Essays on Customization Applications in Marketing

Adiguzel, Feray

Chapter 3
Split Questionnaire Design

3.1 Introduction

Market researchers have traditionally collected consumer information on preferences, attitudes, consumption contexts and lifestyles, by means of often very long questionnaires. In doing so, they need to make tradeoffs between reasonable survey length and the value and quality of additional information. Questionnaire length is a concern since it affects the quality of the data collected in several ways (Berdie, 1989). Long questionnaires lead to higher non-response, item non-response and early break-off rates. They also cause an increase in the use of undesired response styles, increased time to collect the data, and respondent fatigue and boredom. It has been reported that survey respondents become fatigued and irritable when questioned for more than twenty minutes. Many studies indicate that longer questionnaires have lower response rates than shorter ones (Adams and Gale, 1982; Bean and Roszkowski, 1995; Dillman, 1991; Dillman, Sinclair, and Clark, 1993; Heberline and Baumgartner, 1978; Roszkowski and Bean, 1990).

3.1.1 Motivation

We propose a method to design split questionnaire surveys as an effective tool to reduce respondent burden without sacrificing the inferential content of the data. Although Good (1969, 1970) already called for the development of split questionnaire methods to collect survey data more efficiently, in the following thirty-five years, no systematic research on how to best design

split questionnaires seems to have been done. Two decades ago, Herzog and Bachman (1981) advised that a researcher who needs to use a long questionnaire might be well advised to split the material into at least two parts and administer those parts in different orders to different random subsets of the sample. In their split questionnaire survey design, the original questionnaire is divided into sub-components, and subjects only respond to a randomly selected subset of components. A similar idea of designing randomly split questionnaires is applied in what has been called "time sampling". Here, questions are administered in a randomly rotated fashion to different parts of the panel in different episodes (Sikkel and Hoogendoorn, 1995). Incomplete designs in educational testing are based on a similar approach. In test construction, the researcher administers subsets of the total available item pool to the available subjects. The matrix sampling design (Shoemaker, 1973; Thayer, 1983), in which a test instrument is divided in sections, and groups of sections are administered to subjects in a randomized fashion, is used for that purpose.

Each of these previous studies has thus used a randomization approach to design split questionnaires. The important question that remains is how to optimally split the questionnaire such that the least information is lost. Currently, no methods have been published to address that problem, and here lies the contribution of this chapter. Raghunathan and Grizzle (1995) mention that ad-hoc splitting strategies may depend on the purpose and the contents of the survey, contextual placement of certain items, and the partial correlation coefficients of the items. These

correlations may be readily available in tracking or syndicated studies, because here the researcher knows which (groups of) variables are correlated, from their previous measurements. In cross-sectional studies, prior knowledge about inter-relationships between variables can be obtained from a pilot study. However, even when such prior information is available, the construction of a split questionnaire design such that a minimum amount of information is lost is a challenging task. Since the number of possible split questionnaire designs is exponential in the number of questions, it is not feasible to consider all possible splits in designing a questionnaire for real-life applications. Therefore we suggest, in line with previous practice in marketing research, utilizing the natural structure of the questionnaire, in which questions are placed in blocks. Mostly, several questions measuring, for example, one particular attitudinal or lifestyle trait are administered as a group or block. We use this block-structure to generate split questionnaire designs in two different ways: selecting entire blocks of questions, which we call a “between-block design”, or selecting questions in each block, which we call a “within-block design.” In the between-block design, a “split” is comprised of the allocation of selected blocks of questions and respondents answer all questions in these blocks; in the within-block design, a split is comprised of sets of selected questions in each of the blocks, and respondents answer only those questions in each block. For the first method, given the coherent interpretation of the questions in one block, the problem then simplifies to how these blocks should be administered to respondents in an optimal way. On the other hand, for the within-block design, we need to optimally choose questions in each block. The choice between the within-block and the between-block

design should be based on substantive issues, as well as statistical properties of the two types of designs, as will become clear in the following of chapter. We focus on the problem of how to best develop a split questionnaire and propose a method to optimally choose the splits (a set of blocks of questions or questions in each block offered to a respondent) in this chapter.

3.1.2 Outline of the Chapter

The main contribution of this chapter is to propose a method to design split questionnaires. We apply the modified Federov algorithm to find the optimal design from all possible designs because of its speed and reliability. This method has been previously applied in a different context in the design of conjoint experiments (Kuhfeld, Tobias and Garratt, 1994). We propose using Kullback-Leibler (KL) distance between the complete and split questionnaire data as an optimization criterion. The algorithm searches the candidate splits for the split that is optimal in terms of the given criterion. As explained above, we study both between-block and within-block split questionnaire designs. The split questionnaire, once administered, results in data missing by design, which may result in lack of identification of all parameters from the observed data (Little and Rubin, 1997; Rassler, 2002). Specific overlap of the splits of the questionnaire may help to avoid that identification problem. We explain how to construct identified split questionnaire designs, and how to impute the missing data with the Gibbs sampler. Using a small simulated questionnaire, we enumerate all possible designs and compare that with the result of our design generating

algorithm, which reveals that it recovers the optimal split in all cases. We compare the efficiency of split questionnaires generated with our procedure to (random) matrix sampling designs on synthetic data. In practice, market research companies design split questionnaires by randomly choosing blocks, or questions within each block. These methods are similar to the multiple matrix sampling techniques used in testing theory (Shoemaker, 1973), and therefore constitute an appropriate benchmark.

We then apply our approach to data obtained from a questionnaire on web attitudes and perceptions (Novak, Hoffman, and Yung, 2000) to empirically assess the performance of optimal between- and within-block designs, and to compare them to matrix sampling designs and heuristic designs constructed based on a principal components analysis of pilot data. We investigate the sensitivity of the optimal split questionnaire designs to changes in the prior parameters from the pilot study. Finally, we investigate the extent to which the proposed split questionnaire design method may result in better data quality than the complete questionnaire, by studying respondent burden, boredom, and fatigue in a field application of the web-attitude questionnaire. Our conclusion is that optimally splitting questionnaires is worth consideration due to improved questionnaire efficiency and the resulting data quality.

The subsequent sections are organized as follows: Section 2 examines issues in designing a split questionnaire. Since split questionnaire design is one of the methods of collecting data missing by design, we explain other methods of data collection missing by design in Section 3. In Section 4, the design criterion is introduced; the modified Federov algorithm and the

construction of identified split designs are explained. In Section 5, we discuss multiple imputations of the missing data and the estimation of the fraction of missing information. Section 6 provides a simulation study, which investigates the performance of the proposed split questionnaire design method, Section 7 provides the empirical application, and Section 8 summarizes the field study. Finally, in Section 9, the results of this research are discussed and concluding remarks are offered.

3.2 Constructing the Split Questionnaires

Finding an optimal design for a split questionnaire involves finding the configuration of question sets (i.e. those questions given to one respondent, or a "split") such that a minimum amount of information is lost as compared to the complete questionnaire. The design of a split questionnaire, as we propose it, involves two steps. First, one needs to assign questions to blocks with homogeneous content. Second, one needs to allocate either selected blocks to splits, or selected questions within blocks to splits, resulting in between- and within-block designs, respectively. In the first step, one wants to keep thematically closely related questions in the same block[4]. Raghunathan and Grizzle (1995) call this the contextual placement of questions. We start from the assumption that the questionnaire already consists of a number of blocks with questions that

[4] A block structure, if not available a-priori, can be generated using cluster analysis of a pilot with the full questionnaire (Rassler 2002).

need to be kept together, and we will utilize that natural structure of the questionnaire**.** Our approach is thus very suitable for questionnaires comprised of items to measure several multi-item constructs. These are very common in marketing research. Each split questionnaire design is defined by three sets of parameters: the number of splits, the number of blocks/questions per split, and the sampling fraction responding to each split. In this study we investigate the first two parameters and assume throughout that splits are distributed randomly and evenly to respondents. We propose to choose splits from all possible combinations of blocks (between-block designs) or from all possible combinations of questions in each block (within-block designs), using the Kullback-Leibler distance as a measure of information loss, computed from prior parameter estimates. Split questionnaires are one of the methods of collecting data missing by design in surveys with long questionnaires. Now, we explain other methods of data collection that give rise to data missing by design, in order to gain a broader perspective.

3.3 Data Missing by Design

Data collection through surveys requires significant amounts of time, money, and effort. Since time, money and subjects are scarce, in various research areas including marketing, researchers have begun developing more advanced methods to more efficiently collect data. Under time, subject and cost limitations, market research companies sometimes prefer to collect data missing by design, which is also called “planned missingness.” In these studies, companies select sub-parts of the whole

questionnaire to reduce the cost of a study. If planned missignness methods are applied successfully, missingness has little effect on the precision of the parameter estimates of interest. In this section, we talk about these proposed approaches, which are collecting data missing by design. In addition to collecting data missing by design, another currently used procedure in marketing is data fusion, which allows merging data from different sources. Since data fusion and split questionnaires are related, we also discuss data fusion and explain the relationship below. A split questionnaire survey design results in data that is missing by design. Alternative methods are two-stage designs, matrix-sampling designs, subsampling, time-sampling designs, and some experimental design procedures from classical statistics, such as fractional factorial designs or incomplete block designs.

We saw the first applications of data missing by design in experimental psychology and in agricultural experiments (in which plots are used), in which different subsets of questions, plots, or stimuli are administered to different persons, e.g. factorial designs. Since factorial designs take less time (or require fewer resources) and the respondent's task is shorter and less burdening, data collection is more efficient. For instance, Hermkens (1983) uses greco-latin square designs for surveys on equality of income. We also see applications of data missing by design in spatial interpolation problems in environmental science, mining, engineering, geology, soil science and hydrology (Le, Sun, and Zidek, 1997). The most common and widespread usage of data missing by design is in educational testing.

Split Questionnaire Design

Calibration and measurement designs in educational testing are often incomplete designs and used in the framework of item response theory (IRT). The researcher decides to administer only a subset of the total items to the subjects because of the limited testing time (not all available items can be administered to every student). The three commonly used incomplete designs are random incomplete designs, multistage testing designs, and targeted testing designs. In random incomplete designs, the researcher decides which test form is taken by which students without using any a priori knowledge on the ability of a student. In multistage testing designs, the assignment of students to subsets of items from the total item pool in a specific testing stage is based on the observed responses in the previous stage (this is one kind of two-stage design). In targeted testing designs, the structure of the design is determined a priori on the basis of background information. There are two alternative applications of this method. First, the background variables (demographic or income information, etc.) are only used in the assignment of items or tests to students and not in the sampling of the students. In the second application, the background variables are used in the sampling of students as well as in the assignment of tests to students. The efficiency increases if we use a priori knowledge about the difficulty of the items and the ability of students to allocate students to subsets of items (Lord, 1980), which would call for Bayesian methods to develop these kinds of designs. Adaptive or tailored testing in educational testing is another application of “data missing by design.” In adaptive testing, the examinee’s preceding responses are used to select each next item to administer. For example, an examinee answering items correctly would be administered successively more difficult

items, and an examinee answering incorrectly would be administered successively easier items. Although the concept of a "correct" answer may or may not be of use in marketing surveys when designing questionnaires, we believe that the ideas in item response theory models (IRT) from the educational testing literature can be useful in the design of online marketing questionnaires in the future. Some studies on questionnaire designs from item response theory literature are van der Linden, et al. (2004), van der Linden (1999), van der Linden (2004), van der Linden, et al. (1998), and Veldkamp (2002).

The most prominent questionnaire design applications in marketing are conjoint questionnaire designs. Researchers traditionally have constructed designs for (ratings or choice) conjoint experiments using methods from the experimental design literature. Fractional factorial designs are the main method used in experiments. For instance, Lenk et al. (1996) present results that provide shorter questionnaires for metric conjoint analysis. They describe the problems associated with long questionnaires and call for experimental designs and estimation methods to recover parameters with shorter questionnaires. Their paper considers two experimental designs: one in which each subject receives the same set of questions, and one in which subjects receive different blocks of a fractional factorial design. Based on research by Huber and Zwerina (1996), Sandor and Wedel (2001) design conjoint choice experiments based on prior information about the parameters and their associated uncertainty, elicited from managers**.** They use Bayesian design procedures that assume a prior distribution of

likely parameter values and optimize the design over that distribution. Apart from conjoint questionnaire designs, there is to date no research on collecting data missing by design in surveys, and we intend to fill this void with this chapter. Before explaining some alternative tools for collecting data missing by design, we provide some differences and similarities between questionnaire design for conjoint experiments and survey designs.

Conjoint experiments (conjoint questionnaire design) are a specific instance of experimental design, whereas split questionnaire designs toll within the value of survey designs. An experimental design specifies how to allocate resources (attribute levels in conjoint experiments that we want to learn consumer's preferences) in the study. On the other hand, sampling is an economical way to select a small part of the population, so that study of that part permits broad generalizations within reasonable limits of doubt. In this chapter on split questionnaire designs, our purpose is to generate different versions of the questionnaire (which contain fewer questions than the complete questionnaire) with minimum information loss and we would like to know which questions from the whole questionnaire should be chosen to be administered together. The main difference, compared to survey sampling, is that instead of sampling subjects, we select questions and distribute them evenly to subjects. Sufficient subjects should respond to these different versions of the questionnaire. Our approach for designing split questionnaire designs can be modified by selecting questions based on sampled subjects' background information or depending on some selective (classifying) questions. The issue of how many subjects should respond to each version of the questionnaire is also an important issue for future research. After we generate optimal split questionnaires, we collect

data and impute the missing parts. There is no imputation in conjoint designs after data collection. To design conjoint questionnaires, an important assumption that is often made is to assume zero values for the attribute weights. However, in designing split questionnaires, we use the covariance of the questions obtained from pilot studies.

3.3.1 Two-Stage Designs

Two-stage designs are the most common example of procedures that generate data missing by design. The first stage consists of core questions to elicit information which we want to have from all respondents, whereas the second stage, the remaining blocks of questions, are given to a subset of the entire sample, or to a stratified random sample with selection probabilities dependent on the first stage. The correlation between the core measure in the two stages and selection criteria for the second stage sample provide the information needed to make full-sample inferences about the second stage measures (Neff, 1996).

3.3.2 Matrix Sampling Design

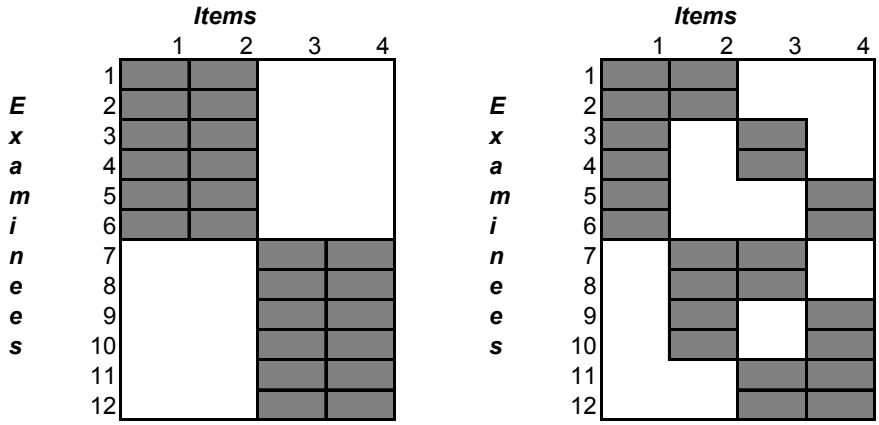
Matrix sampling refers to the random sampling of a rectangular array of row-column entries from a larger matrix from the population. The National Assessment of Educational Progress (NAEP) uses matrix sampling designs in item testing[5]. Item testing is a popular psychometric application of this

[5] NAEP is the only national assessment in the US that measures what American students know and can do over time in various subjects such as reading and mathematics. The analysis of NAEP is IRT based and contains several consecutive steps.

method in which the rows constitute examinees and the columns constitute items. If more than one matrix is sampled, this is referred to in the literature as a multiple matrix sampling. Figure 3.1 depicts a non-overlapping multiple matrix sampling design (NMS) wherein examinees and items are sampled without replacement, and an overlapping multiple matrix sampling design, which in fact is a balanced incomplete block design (BIB). When more than one matrix is sampled, the point estimates for a single matrix are repeated, computed and averaged over all matrices, since the mean of unbiased estimates is also unbiased. When it comes to the computation of standard errors, the situation is more complicated. In matrix sampling designs, different respondents are asked different questions, and the set of questions varies across strata. These designs have common applications in computer-aided interviewing, or as a part of item experiments. In split questionnaire survey design, missing data are imputed to end up with a complete data set, which is not the case in matrix sampling as used in educational testing.

Figure 3.1: An example of NMS and BIB design

1-) Non-overlapping matrix sampling (NMS) 2-) Balanced incomplete block design (BIB)



3.3.3 Time Sampling Design

There are two basic ways to obtain information about the continuous behavior of consumers in time. The first is continuous consumer panels, which help to obtain a continuous record of the behavior of consumers for the entire time period. The second is to sample time, that is, to observe consumers at various points in time and to infer from these observations what behavior took place for those periods for which no measurements were made. Among market researchers, the most commonly used method is to have each sampling unit record its own continuous behavior via a self-administered form, usually referred to as a diary. Although used in a wide variety of contexts, the most frequently used types of consumer panels in

marketing are the purchase panel, the media measurement panel and the product test panel.

Sampling over time enables us to monitor, analyze and understand social processes through the estimation and analysis of changes in variables of interest. In addition to the usual sample design issues considered for a sample used for one time period, the design of a time sampling scheme needs to consider the frequency of sampling and the spread and pattern of inclusion of selected units over time. A key issue is whether to use overlapping or non-overlapping samples over time. For overlapping samples, the precise pattern of overlap must be designed. Factors that affect the design of a sample over time are: the key estimates to be produced, the type and level of analyses to be carried out, cost, data quality, and reporting load. The interaction between the design of the sample in time and the other features of the design, such as stratification and cluster sampling, also needs to be decided. Time series may be produced and analyzed, which may involve seasonal adjustment and trend estimation. Composite estimation is one of the methods of estimation that is used in time sampling that involve using data for the current and previous time periods and give different weights to matching and non-matching sample units.

Repeated, panel, and longitudinal surveys, rotating panel surveys, split panel surveys and rolling samples are important examples of the application of time sampling. A longitudinal survey (or panel survey) is a survey that uses a sample in which the same units are included for several time periods. A repeated survey is a survey conducted at different times

with no attempt to have sample units in common. Rotating panel survey is a panel survey in which a proportion of units are removed from the survey at some time periods and replaced by other units. In this method, a different rotation pattern can be used, i.e. the pattern of inclusion of sample units over time, such as overlapped or nonoverlapped (orthogonal) patterns. One example of time sampling design is illustrated in Figure 3.2.

3.3.4 Subsampling or Multistage Sampling

In subsampling, the aim is to divide the blocks into smaller and preferable subsamples (Figure 3.2). If the blocks represent clusters, subsampling is generally used to divide larger clusters into smaller clusters in sampling design. The advantage of this method is decreasing variance due to a decrease in the degree of clustering, without incurring a proportional increase in cost (Kish, 1965). The difference between estimates from these independent samples may be used to estimate the error variance. Then these error variances are straightforwardly projected to the entire sample formed from the combined subsamples. Subsampling designs require a minimum of two subsamples and homogeneity between samples. Although it is often not practical to include many more than two in the design, this method can be extended to more stages.

3.3.5 Data Fusion

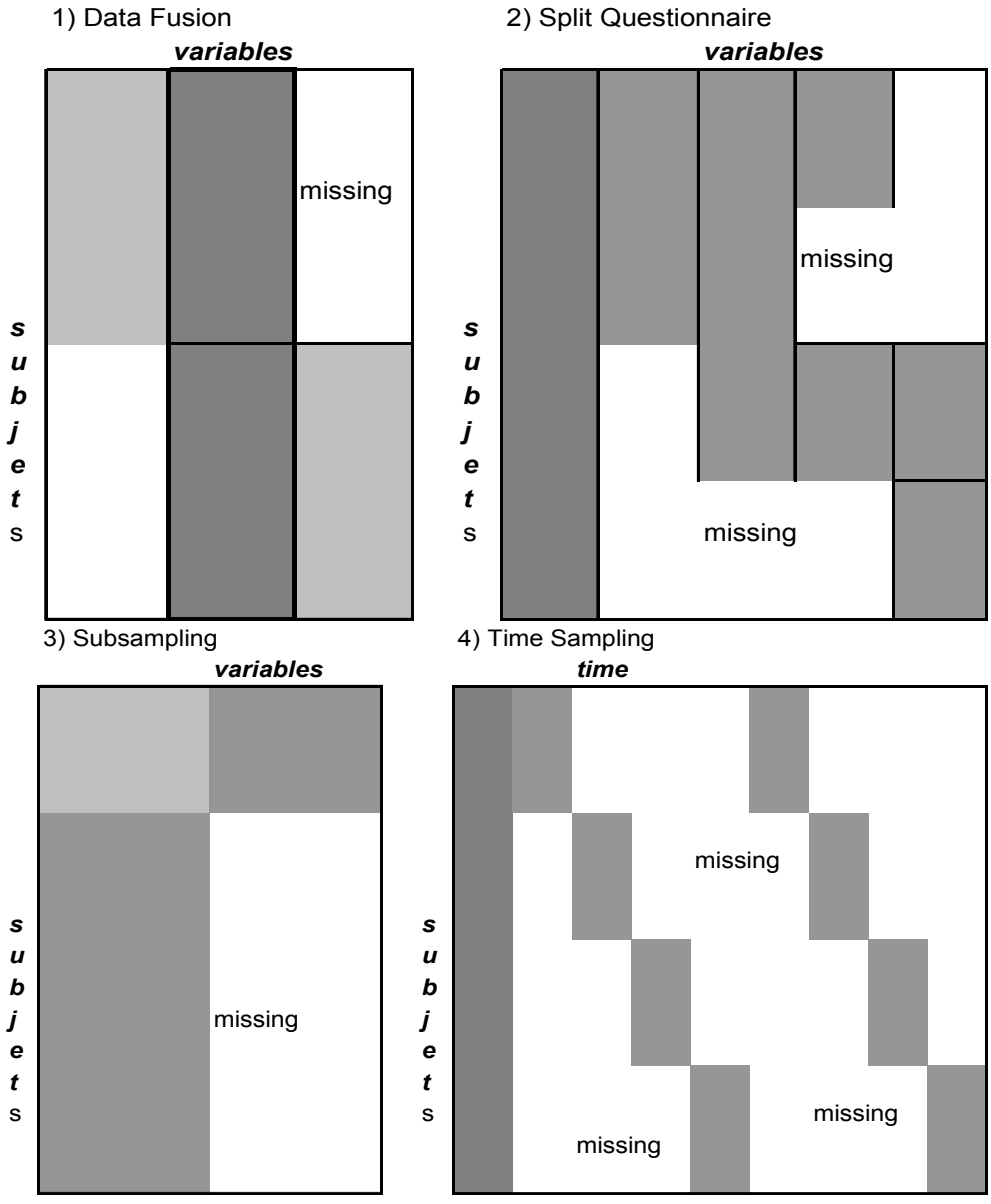
Data fusion is related to split questionnaire designs. Data fusion or statistical file matching techniques merge data sets from different survey

samples to solve the problem that exists when no single file contains all variables of interest (Figure 3.2). Split data arise when data on two different sets of variables are obtained from two independent samples, while a number of variables (usually demographics) are measured in both samples (Kamakura and Wedel, 1997, 2000, and Gilula et al., 2006). Merging data sets is usually done on the basis of variables common to all files, and the methods in question assume conditional independence of the variables never jointly observed given the common variables.

Data fusion can also be used to reduce the required number of respondents or questions in surveys. For example, the Belgium National readership survey on media and products is distributed to two different groups of 10,000 respondents each in Belgium and later merged into a single survey (van der Puttan et al., 2002). In this way, the cost and the time for each respondent to complete the questionnaire was conveniently reduced.

The focus in data fusion studies, however, is more on how to merge two different data sets from different surveys. But in principle, we can design split questionnaires and distribute them to respondents and later merge these different sets using data fusion techniques. Split questionnaire survey design can be applied especially for media and purchase surveys. For instance, data from a television measurement panel and a purchasing behavior panel can be merged together with data fusion.

Figure 3.2 Data missing by design



3.3.6 Incomplete Block Design

The goal in incomplete block design is to construct a design such that any pair of treatments (blocks of questions) occurs equally often within some block (split). A solution can be found for any number of treatments and any size of block, but most of the solutions require too many replications for the usual situations in survey designs. For a given number of treatments and a given size of incomplete block, balanced designs allow little choice in the number of replications. Assignment of blocks of questions to splits can depend on the following constraints: The number of blocks assigned to each split (k), the number of splits to which each block assigned (λ), combinations of blocks are assigned to splits (a minimum number of splits or number of split per pair, etc.).

When an incomplete design is formed so that every pair of treatments occurs together in the same number of blocks, the design is called a balanced incomplete block design (Giesbrecht, 2004). A balanced incomplete block design (BIBD) is expressed with five parameters, (υ, b, r, k, λ), and is a family of b sets, called blocks, each consisting of k (where $k < \upsilon$) elements taken from a set of υ elements, such that each element occurs in exactly r blocks and every pair of elements occurs together in exactly λ blocks. Since b and r can be calculated from υ, k and λ, we use (υ, k, λ) as the parameters for the design. Balanced incomplete block (BIB) designs do not exist for all combinations of blocks sizes (k), numbers of treatments (υ) and number of replications (r). There are four necessary conditions for the existence of BIB design.

1. $\upsilon r = bk$

2. $\lambda(\upsilon - 1) = r(k - 1)$

3. $b \geq \upsilon$

4. if $\upsilon = b$, and

if υ is even then k-λ is a perfect square and if υ is odd then $z^2 = (k - \lambda)x^2 + (-1)^{(\upsilon-1)/2}\lambda y^2$ has a solution in integers with x, y, z not all equal to 0.

Proof: The number of pairs in a block is ${}_kC_2 = \frac{k(k-1)}{2}$. The number of treatment pairs is ${}_\upsilon C_2 = \frac{\upsilon(\upsilon-1)}{2}$. There are b blocks, so the total number of pairs is $b \times \frac{k(k-1)}{2}$. Each pair occurs λ times, so the total number of pairs is $\lambda \times \frac{\upsilon(\upsilon-1)}{2}$. Equating these two expressions gives $r.(k-1) = \lambda.(\upsilon - 1)$

A BIB with 20 blocks would lead to 190 version of the questionnaire for identifiability, and would necessitate unrealistically large sample sizes. If instead a partial BIB is utilized, this leads to many occurrences where questions/blocks do not co-occur, hence, bivariate information is not always available (which leads to identification problems).

We use prior information to design split questionnaires, but incomplete block design is not based on such prior information. Balanced incomplete

block designs depend on the number of blocks assigned to each split (k) and the number of splits to which each block assigned (λ). We use covariance relationships as a prior to generate split questionnaire designs, which allow us to reduce more questions relative to BIB designs. Incomplete balanced block design comes with some certain number of replications (distinct splits) and need more splits (different versions of questionnaires) for identification. Because in split questionnaire designs, we don't have any restrictions on the number of splits in generating identified designs, we need fewer splits.

Figure 3.3: Feasible balanced incomplete block design

Splits	**Blocks**				
	1	**2**	**3**	**4**	**5**
1	1	1	0	0	0
2	1	0	1	0	0
3	1	0	0	1	0
4	1	0	0	0	1
5	0	1	1	0	0
6	0	1	0	1	0
7	0	1	0	0	1
8	0	0	1	1	
9	0	0	1	0	1
10	0	0	0	1	1

One of the applications of incomplete balanced block design in marketing is demonstrated by Rink (1987). He explains and illustrates how these designs can circumvent problems where the respondent must rank many objects. Raghavarao and Federer (1979) present balanced incomplete block design designs as an alternative approach to the randomized

response method for dealing with sensitive questions in a survey context. Their proposed method increases the chances of obtaining honest and unbiased responses by protecting respondents' privacy in a survey, which includes questions that the respondent may not be inclined to answer truthfully. Each respondent is administered a questionnaire containing a subset of the possible questions in these designs. That is, each respondent is assigned a "block" in an incomplete block design[6]. This method applies to questionnaires in which all blocks have at least one quantitative question. The key idea in this approach is that scores for a set of k of the v questions, sensitive and/or non-sensitive, are added, and only a total score for the k questions is reported by the respondent. Different respondents receive different sets of k questions; there are b different sets of questions constructed according to known experimental designs, such as the supplemented block designs and balanced incomplete block design. The block of k questions is randomly assigned a respondent, and all blocks have an equal or nearly equal number of respondents. We can estimate population proportions or means for each question from the block totals; however, we are unable to determine what an individual's response was to a particular question. With the usage of incomplete balanced block designs, one saves interviewing time for questionnaires with several sensitive questions and potentially improves response.

---

[6] In a split questionnaire design, each split (i.e. version of the questionnaire) plays the role of a "block."

3.4 Measuring Information Loss

3.4.1 Optimal Split Questionnaires Using KLD

We use the Kullback-Leibler (KL) measure, the distance between two probability models, to choose the best among all possible designs. The KL-distance was developed by Kullback and Leibler (1951) from "information theory." Here, it is first applied to design construction. The KL-distance defines the distance between the probabilistic models *f* and *g* for as the (usually multi-dimensional) integral:

$$I(f,g) = \int f(y)\log(\frac{f(y)}{g(y \mid \theta)})dy \qquad (3.1)$$

I(f,g) is the "information'' lost when *g* is used to approximate *f*. An equivalent interpretation of minimizing *I(f,g)* is finding an approximating model that is the shortest distance away from "the truth." If *f(y)* and *g(y|θ)* are multivariate normal distributions with a common variance-covariance matrix, then the Kullback-Leibler distance reduces to the Mahalanobis distance (Bar-Hen and Daudin, 1995), which is frequently used as a distance measure in the literature.

We assume that the optimization of the split questionnaire design (SQD) is done under one external constraint fixed by the researcher, which is the total number of splits (*K*) desired. We assume that the researcher knows this number from prior considerations, or that issues related to the implementation of the questionnaire dictate it. The optimization can also accommodate any other practical constraint, such as one that induces

respondents to answer a fixed number of (blocks of) questions, i.e. each candidate split should contain a predetermined number of blocks. These constraints are illustrated below. After generating *K* splits and evenly distributing these splits to respondents, the Kullback-Leibler distance is calculated. In our notation, *K* denotes the total number of splits, *N* is the number of respondents, *B* is the number of blocks, Q_b is the number of questions in block b, *Q* is the total number of questions, $(\sum_{b=1}^{B} Q_b = Q)$, *Y* is the data-matrix containing the answers of the respondents and *D* is the questionnaire design matrix with 0/1 entries (i.e. a fully observed matrix of indicators whose elements are zero or one depending on whether the corresponding elements of Y are missing or observed):

$$d_{ij} = \begin{cases} 1 & \textit{if question } j \textit{ is given to respondent } i \\ 0 & \textit{otherwise} \end{cases}$$

Now f(Y|D) is the likelihood of the incomplete data with respect to the split questionnaire design matrix and f(Y) is likelihood of the data with respect to the complete questionnaire. The Kullback-Leibler distance between the complete data likelihood f(Y) and the split data likelihood f(Y|D) is defined as:

$$KL(D) = \int f(Y) \ln\left[\frac{f(Y)}{f(Y \mid D)}\right] dY, \qquad (3.2)$$

$$= E\ln[f(Y)] - E\ln[f(Y \mid D)],$$

where each expectation is with respect to the true distribution f(Y), where $Y_{N\times Q} = [Y_1, Y_2, \ldots, Y_Q]$. Thus, the KL(D) in this case measures the distance between the distribution of the complete data f(Y) and the incomplete data f(Y|D) given the split questionnaire design D, i.e. it assesses the expected loss of information by deleting data according to the split questionnaire, relative to the complete questionnaire data. The most efficient questionnaire design (D) minimizes KL(D). The first term on the right hand side in the equation for KL(D) is the same for each possible design since it is derived from the complete questionnaire. Consequently, maximization of the second term on the right hand side suffices. Since f(Y) is the same for each possible design, lnf(Y|D) will be maximized in the sequel. Minimizing the KL-distance can be seen as finding the split questionnaire yielding incomplete data, which are closest in expectation to the data that would have been obtained with the complete questionnaire.

We will assume the form of lnf(Y|D) to be a multivariate normal, as a function of the parameters μ and Σ, as shown below. In Appendix I we provide an extension of the KL-distance for mixed data consisting of continuous and discrete variables using a general location model. However, multivariate normality is often assumed for responses of scales in many marketing surveys, including those measuring attitudes, satisfaction, lifestyles etc. (Huber et al. 1993). In addition, the normal distribution has minimal KL-distance to any unknown distribution function (O' Hagan 1994), and in this case minimizing the KL-distance is equivalent to minimizing the Mahalanobis distance.

We have Q-variate normal data $N_Q(\mu, \Sigma)$ with $\mu = (\mu_1, \ldots.., \mu_Q)$ and $\Sigma_{Q\times Q}$. For now, $\mu_{Q\times 1}$ and $\Sigma_{Q\times Q}$ are assumed known. These are considered prior information that can be obtained from past data or through a pilot experiment. The aim is to construct the design using $\mu_{Q\times 1}$ and $\Sigma_{Q\times Q}$ as prior information. Thus, we have the following optimal design criterion:

$$L = \ln L(Y \mid D, \mu, \Sigma)$$

$$= \prod_{i=1}^{n}\left(\frac{-p_D}{2}\right)\ln(2\pi) - \frac{\ln|\Sigma(D)|}{2} - \frac{1}{2}\left[(Y_{obs} - \mu(D))'\Sigma(D)^{-1}(Y_{obs} - \mu(D))\right] \qquad (3.3)$$

where p_D is the number of parameters under design D, *n* the total number of respondents, $Y_{obs}=Y_{ij}d_{ij}$ the data observed under the split questionnaire D, and $\mu(D)$ and $\Sigma(D)$ denote the subvector of the mean vector μ and the square submatrix of the covariance matrix Σ which are obtained from complete data estimates from a pilot study, respectively, that pertain to the variables that are observed in design D.

3.5 Identification Issues in Constructing SQD

When we construct a split questionnaire design, we should be able to estimate all parameters from the observed incomplete data. We call a design that enables the estimation of all parameters (of the multivariate normal distribution) a fully identified design. Clearly, not all designs are fully identified. We illustrate the identification problem briefly through the following example. Assume we want to estimate the parameters of a

multivariate Normal distribution for three blocks, X, Y and Z in a between-block design. However, we have a split A- with only X and Y and a split B- with only X and Z observed together. The covariance matrix of Y and Z is written $V(Y,Z) = V(Y,Z \mid X) + V'(X,Y)V(X)^{-1}V(X,Z)$, where V(X) the covariance matrix of X, and V(Y,Z|X) the covariance matrix of Y and Z conditional on X. We can estimate V(X,Y) from split A, V(X,Z) from split B, and V(X) from both splits, but we cannot only directly estimate V(Y,Z|X) from the available incomplete data. However, if we assume conditional independence of the Y and Z variables given X, we can estimate V(Y,Z) from $V(Y,Z) = V'(X,Y)V(X)^{-1}V(X,Z)$, since all terms on the right hand side are estimable (see Gilula, McCulloch and Rossi, 2006; Rassler, 2002; Rodgers, 1984). However, if we use this conditional independence assumption in a model for imputing the missing data, this implies that for all parameter estimates or statistics subsequently computed from the imputed data this conditional independence assumption should also hold. That assumption is a strong one, which may limit the usefulness of such split questionnaire designs in practice.

Rassler (2002) and Gilula, McCulloch and Rossi (2006) suggest (in the context of data-fusion) to use informative priors in the imputation to overcome the identification problem. The use of priors adds information that enables estimation of the parameters that are not identified by the split questionnaire design. The fact that V(Y,Z|X) is inestimable results in non-positive definite variance-covariance matrix V(X,Y,Z), which we can avoid using prior information. If one uses the Gibbs sampler for imputation, as we will below, such prior information also overcomes lack of convergence.

Using informative priors for the means and covariance matrix of the normal distribution results in an imputed dataset devoid of conditional independence properties induced by the design, which is highly desirable. Since the design itself is constructed based on such prior information, it is natural to also include that same prior information in imputing the missing data. However, it is even more desirable to address the identification problem by constructing designs that do not suffer from it, which we do below.

If all possible pairs of questions occur in an optimal split questionnaire design, this ensures that all parameters of a multivariate normal distribution are identified and estimable from the observed data. Let us consider the between-block design: if we have a questionnaire with n_B blocks and we impose the constraint of n_S blocks per split, then the number of splits K for a fully identified design needs to satisfy $\binom{n_B}{n_S} \le K \le \frac{n_B(n_B-1)}{n_S(n_S-1)}$, where $\binom{n_B}{n_S}$ is the size of the candidate split-set. Note that this is a necessary, but not sufficient condition. In practice one can easily check the identification of any design by looking at the (D'D) matrix: only designs with all off-diagonal elements greater than 0 are fully identified designs. In generating constrained split questionnaire designs, we recommend that one only considers fully identified designs by imposing the identification constraint $(d_i' d_j) \neq 0, \quad \forall_{i \neq j}$, and employ the prior information used to construct the design also in imputing the missing values. This is what we will do

throughout the remainder of this chapter, and we recommend it in general as a procedure for constructing split questionnaires.

3.6 Design Generating Algorithm

We assume that the split questionnaire design (SQD) is constructed under the external constraint that the total number of splits (*K*) is fixed. The optimization can also accommodate other practical constraints, such as that one or more blocks are included in every split, or that each candidate split contains a predetermined number of blocks. Note that these constraints are possible, but not needed (such constraints are illustrated in the applications below). In order to find the most efficient *K* splits out of all possible candidate splits ($N_S = 2^Q$, with Q the number of questions), one could generate all $N_D = \binom{2^Q}{K}$ possible designs and retain the one with the smallest value of the KL-measure. In most practical situations, it is not feasible to do this, since it is usually not computationally feasible to list all N_D possible subsets out of 2^Q designs. Therefore, we need to use an efficient algorithm to search the design-space. Such an algorithm would conduct a search among all possible candidate splits for one that improves the KL criterion. We apply the modified Federov algorithm for that purpose. The modified Federov algorithm is a popular algorithm for experimental design construction, since it is robust and fast. Kuhfeld, Tobias and Garratt (1994) applied it to generate conjoint choice designs.

We start describing the procedure that is used to generate the between-block designs. We assume that if there are N individuals, then N/K

individuals will be assigned randomly to each of the K splits. Each alternative split questionnaire design then consists of an *N* x *Q* matrix *D* with *K* different split patterns. Each entry in the matrix D is a 0 or 1, indicating whether a question is included or excluded in that particular split. In constructing between-block designs, we constrain all questions in one block to be assigned to the same respondent. That is, if we have five blocks with four questions and one particular split at the block-level is [11010], we will use d_{ij}=[1111 1111 0000 1111 0000] as a row in the design matrix D. The proposed procedure to construct split questionnaire designs operates as follows:

Step 1. Build a candidate split-set (C, a $N_S \times Q$ matrix), which is a list of all potential splits contained in its rows. Inadmissible designs are removed from C.

Step 2. Choose a starting design at random, say D_0. Using the pilot data, obtain estimates for the parameters of the model for each of the questions in the questionnaire. Compute the KL-measure for the starting design $KL(D_0)$ based on these estimates, using (3.3).

Step 3. Take the first split (first N/K rows) in the starting design D_0. Exchange that with the candidates, $\ell = 1,...,N_S$, i.e. each of the rows in C, in turn. For every exchange, compute the KL-distance in (3.3), i.e. D_0^{ℓ}. Keep that split that minimizes the KL-distance, i.e. $D_1 = \min_{\ell} KL\left(D_0^{\ell}\right)$, and replace D_0 by D_1.

Step 4. Find the best exchange (if one exists) for the next split in the target design D_1 (i.e. the second set of N/K rows), by sequentially processing the candidates $\ell = 1,...,N_S$ in C, and replacing the design matrix D_1 by $D_2 = \min_{\ell} KL(D_1^{\ell})$.

Step 5. Ensure that the design is fully identified by checking off-diagonals of the (D′D) matrix at every step, and reject splits that cause zero off-diagonal values.

Step 6. The first iteration is completed once the algorithm has found the best exchanges for all of the splits in the target design matrix. Then, the algorithm moves back to the first split in the target design matrix and replaces it again with each candidate in C, cycling through steps 3 and 5, until no improvement is possible.

Step 7. To avoid local optima, the whole process is restarted with different (random) starting designs and the best design is selected, i.e. the one that yields the lowest KL-distance.

3.6.1 Generating Within-Block Designs

Whereas the construction of between-block designs is feasible with the modified Fedorov algorithm described above, that of the within-block design is not in most practical situations, because of the enormous size of the design space. Therefore, we choose questions within each block using a "greedy" approach, as follows. Instead of optimizing the full within-block split design, we generate splits for each block sequentially. For block B there are 2^{Q_B} possible splits, with Q_B the number of questions. We have a

candidate split-set for each block, denoted as C_b, for b=1,…,B. The procedure then operates as follows.

Step 1. Build a candidate split-set (C_b), for each block. Inadmissible designs are removed from C. Choose a starting design at random for every block, say $D_{0,b}$.

Step 2. Find the optimal K splits in the first block from C_1 using the modified Federov algorithm as described in the Steps 3-6 above, assuming the other blocks are complete, to obtain $D_{1,1}$.

Step 3. Then, find the optimal splits in the second block searching across the candidate splits in C_2, as described in steps 3-6 above, given the optimal splits of the first block and assuming the remaining blocks are complete, to obtain $D_{2,1}|D_{1,1}$.

Step 4. Continue this procedure by sequentially passing through the remaining blocks, finding the optimal splits for each block using steps 3-6 above, given the optimal designs of the previous blocks, and assuming the remaining blocks complete, thus obtaining $D_{b,1}| D_{b-1,1},\ldots,D_{1,1.}$

Unfortunately, it proves difficult to produce fully identified within-block designs using the “greedy” approach described. We therefore choose to generate only locally identified designs by checking the $D_b'D_b$ matrix of each block b separately. This does not guarantee the appearance of all question-pairs in the complete design, which is needed for the design to be fully identified. Thus, the constructed within-block split questionnaire designs are neither fully identified nor globally optimal, but are still more

efficient than designs constructed by choosing questions within each block at random or with heuristic procedures.

For within-block designs, constraints can be imposed by only considering admissible designs in the candidate split set C_b. One important class of constraints is imposed by forced within-block skip patterns in the questionnaire (see Sudman and Bradburn, 1989, p.224). The within-block branching structure of the questionnaire can be accommodated in the split questionnaire design, by forcing a higher node question into any split that also contains the lower node question.

3.7 Multiple Imputations with Gibbs Sampling

The within- and between-block split questionnaire designs produce datasets with intentionally missing data. To obtain complete data, instead of using a single imputation, which ignores uncertainty due to imputation and therefore underestimates the variability of the resulting estimates (Rubin, 1987), we use Bayesian proper multiple imputations by drawing values of missing data (Y_{mis}), and μ and Σ from their full conditional posterior distributions using Gibbs sampling (Gelfand and Smith, 1990). We use informative priors, μ_{pr} and Σ_{pr}, obtained from the full questionnaire in a pilot study, with n_0 and ρ the prior number of observations and degrees of freedom on which the μ_{pr} and Σ_{pr} are based, respectively. Let $\Sigma_{obs,obs}$, $\Sigma_{mis,mis}$, and $\Sigma_{mis,obs}$ denote the sub-matrices of Σ formed by the indices corresponding to the observed and missing Y values; μ_{obs}, μ_{mis} denote the corresponding sub-vectors of μ. The conditional distribution of Y_{mis}, given

Y_{obs}, μ_m, and Σ, is normal distribution with mean $\mu_{mis} + \Sigma_{obs,mis}\Sigma^{-1}_{obs,obs}(Y_{obs} - \mu_{obs})$ and variance $\Sigma_{mis,mis} - \Sigma_{obs,mis}\Sigma^{-1}_{obs,obs}\Sigma_{mis,obs}$. The Gibbs sampler iterates between:

Step 1. Draw $Y^{(t+1)}_{mis}$ given μ_0, Σ_0, and Y_{obs}:

$$Y^{(t+1)}_{mis} \mid Y_{obs} \sim MVN\left(\mu_{mis} + \Sigma_{obs,mis}\Sigma^{-1}_{obs,obs}(Y_{obs} - \mu_{obs}); \Sigma_{mis,mis} - \Sigma_{obs,mis}\Sigma^{-1}_{obs,obs}\Sigma_{mis,obs}\right) \tag{3.4}$$

Step 2. Draw $\Sigma^{(t+1)}$ given $\mu^{(t)}$ and $Y^{(t+1)} = (Y_{obs}, Y^{(t+1)}_{mis})$ from[7]:

$$\Sigma^{(t+1)} \mid Y \sim IW(n_{obs} + \rho, (n_{obs} - 1)S + \rho \times \Sigma_{pr} + S_m) \tag{3.5}$$

where S is the sample covariance matrix and

$$S_m = \frac{n_{obs} \times n_0}{(n_{obs} + n_0)}(\bar{y} - \mu_{pr})(\bar{y} - \mu_{pr})',$$

Step 3. Draw $\mu^{(t+1)}$ given $\Sigma^{(t+1)}$ and $Y^{(t+1)}=(Y_{obs}, Y^{(t+1)}_{mis})$ from[8]:

$$\mu^{(t+1)} \mid \left(\Sigma^{(t+1)}, Y\right) \sim N\left(\frac{1}{n_{obs} + n_0}(n_{obs}\bar{y} + n_0\mu_{pr}), \frac{1}{n_{obs} + n_0}\Sigma^{(t+1)}\right) \tag{3.6}$$

---

[7] With noninformative priors: $\Sigma^{(t+1)} \mid Y \sim IW(n_{obs} - 1, (n_{obs} - 1)S)$

[8] With noninformative priors: $\mu^{(t+1)} \mid \left(\Sigma^{(t+1)}, Y\right) \sim N\left(\bar{y}, \frac{1}{n_{obs}}\Sigma^{(t+1)}\right)$

The Gibbs sampler is easy to implement and enables quick imputation of the missing values. In addition, it can be used simultaneously and in the same manner to impute missing values arising to item non-response (Schaffer, 1997).

3.8 Estimation of the Fraction of Missing Information

The incomplete data generated through the split questionnaire design contain less information on the parameters than the complete data. We estimate the fraction of missing information of the parameters using the missing information principle (Orchard and Woodbury, 1972, see appendix B). Since the complete data information is the sum of the observed data information and the missing data information, we can write:

$$\frac{1}{V(\hat{\theta})} = \frac{1}{V(\hat{\theta}_{obs})} + \left(\frac{1}{V(\hat{\theta})} - \frac{1}{V(\hat{\theta}_{obs})} \right) \tag{3.7}$$

Here $V(\hat{\theta})$ is the complete information on θ estimated from the Fisher information matrix. $V(\hat{\theta}_{obs})$ is the expected observed data information, which we estimate after the multiple imputation of the missing data with the Gibbs sampler. If we divide both sides by the missing information and take the fraction of missing information (γ) to be equal to the missing information divided by the complete information, we obtain:

$$\gamma = \frac{\left(\frac{1}{V(\hat{\theta})} - \frac{1}{V(\hat{\theta}_{obs})} \right)}{\frac{1}{V(\hat{\theta})}} \tag{3.8}$$

This quantity shows how much information there is in the data on the parameters in question, and can be used as a statistic to evaluate the efficiency of split questionnaire designs.

3.9 Simulation Studies

Before we extensively investigate the performance of split questionnaire designs on empirical data below, we first illustrate them with simulated data. We conduct two simulation studies, focusing on between-block designs. First, we investigate the performance of the modified Fedorov algorithm in identifying the optimal design. Second, we compare optimal split questionnaire designs to matrix sampling designs.

We construct a split questionnaire design that is small enough to enumerate all possible designs, which makes it possible to investigate the performance of the modified Federov algorithm in finding the optimal design. Let Y_{ij} denote the answer of respondent $i \in \{1, \ldots, N\}$ to question $j \in \{1, \ldots\ldots, Q\}$, which forms the complete data matrix Y. We assume a between-block design, with $B = 5$ blocks and each block containing $Q_b = 4$ questions, so that in total we have twenty questions. We generate Y from a multivariate normal distribution with given $\mu_{Q\times 1}$ and $\Sigma_{Q\times Q}$. The matrix X is an

$N_S \times B$ matrix containing N_S possible or candidate splits, 1 denoting an included block and 0 denoting an excluded block. There are 32 candidate split points contained in the matrix *X*, but unrealistic or undesirable combinations such as one where none of the questions is asked (a row with only zeros in the design matrix *X*) or where just one block of questions is asked, are excluded, as indicated in the candidate split set shown in Table 3.1. Even under the external constraint that fixes the number of desired splits (*K*), there are many possible designs. For example, there are in total 5311735 (= 26!/(16!10!)) different designs for *K* = 10 splits. We choose *K* splits from the candidate split matrix in Table 3.1, and distribute these splits evenly to one hundred subjects. We do this both with the modified Fedorov algorithm and through complete enumeration. The matrix *D* contains the design with the *K* splits. We eliminate the responses of the subjects from the complete data matrix (Y) according to the split design (D) and compute the KL distance. We choose the SQD design with the maximum Inf(Y|D) among all possible designs as the optimal design. We investigate three different numbers of desired splits: *K* = 5, *K* = 10 and *K* = 15.

The time that the modified Federov algorithm needed to find the optimal questionnaire design with *K*=5, 10 or 15 splits is compared to that for complete enumeration in Table 3.2. All calculations are done with a Pentium 3 computer, using GAUSS software. For the Federov algorithm, we used 10 iterations, and 1000 different random starts. All 1000 random starts produced the same optimal design in all three cases in $1/10^{th}$ or less of the computation time of complete enumeration, as shown in Table 3.2. This indicates that the performance of the Federov Algorithm as applied to the problem of split questionnaire design is highly satisfactory.

We now illustrate the performance of optimal between-block split questionnaire designs (SQD) relative to matrix sampling designs (MSD) in a second simulation study (within-block designs are investigated more extensively in the empirical application below). We have six blocks and five questions per block. We optimally design the questionnaire and impute the resulting missing data with the Gibbs sampler. We investigate constrained and unconstrained between-block designs, with 5 or 10 splits. To assess the performance of the proposed method, next to the fraction of missing information, we compute the KL-distance and the Bayes information criterion (BIC), where BIC=-2×lnf($Y|D$)+ln(N)×2. Further, we calculate the mean absolute deviation (MAD) and the root mean square error (RMSE) of the estimates of variance and covariance parameters for the SQD and the MSD relative to the complete data (the optimal design procedure improves efficiency and thus affects only variance and covariance estimates). The results are shown in Table 3.3. We obtain better values for the BIC- and KL- statistics and less missing information for the SQD as compared to the MSD. Parameter estimates are also closer to the true values for the SQD: the MAD is equal to 3.143 for 10 splits and 2.817 for 5 splits while these values are equal to 3.730 and 3.210 for the matrix sampling design. The missing information for the unconstrained split designs is 24% (ten splits) and 27% (five splits), and 22% and 29%, for constrained split designs, respectively, when we eliminate 50-60% of the questions. In contrast, the fraction of missing information for the MSD is consistently higher. Since

these results support the performance of the SQD, we investigate its performance in an empirical setting in the next section.

Table 3.1: Candidate split set for a five block between-block design

N_S	Block 1 Q1-4	Block 2 Q5-8	Block 3 Q9-12	Block 4 Q13-16	Block 5 Q17-20
1	0	0	0	0	0
2	1	0	0	0	0
3	0	1	0	0	0
4	0	0	1	0	0
5	0	0	0	1	0
6	0	0	0	0	1
7	1	1	0	0	0
8	1	0	1	0	0
9	0	1	1	0	0
10	1	1	1	0	0
11	1	0	0	1	0
12	0	1	0	1	0
13	1	1	0	1	0
14	0	0	1	1	0
15	1	0	1	1	0
16	0	1	1	1	0
17	1	1	1	1	0
18	1	0	0	0	1
19	0	1	0	0	1
20	1	1	0	0	1
21	0	0	1	0	1
22	1	0	1	0	1
23	0	1	1	0	1
24	1	1	1	0	1
25	0	0	0	1	1
26	1	0	0	1	1
27	0	1	0	1	1
28	1	1	0	1	1
29	0	0	1	1	1
30	1	0	1	1	1
31	0	1	1	1	1
32	1	1	1	1	1

Note: The size of the restricted split is 26 by excluding the splits with indices 1 to 6.

Table 3.2**:** Performance of the modified Federov algorithm

K	# of Possible designs (N_D)	Complete Enumeration (sec.)	Modified Federov Algorithm (sec.)[1]
5 splits	65780	260	20
10 splits	5311735	10456	50
15 splits	7726160	13343	78

[1] The modified Federov Algorithm results are based on 1000 random starts.

Table 3.3: Simulation results for between-block designs

Design	Unconstrained 10 Splits		Constrained 10 Splits		Unconstrained 5 Splits		Constrained 5 Splits	
Measure	SQD[a]	MSD	SQD	MSD	SQD	MSD	SQD	MSD
MAD	3.143	3.73	2.471	2.773	2.817	3.21	3.001	3.102
RMSE	3.454	4.283	2.753	3.252	3.288	3.764	3.514	3.701
γ[b]	0.243	0.317	0.217	0.284	0.269	0.306	0.294	0.304
% Missing	0.600	0.600	0.500	0.500	0.533	0.533	0.500	0.500
BIC	5232	7193	8777	8989	4570	8170	8764	8796
logL(D)	-2608	-3589	-4380	-4486	-2277	-4077	-4374	-4390

[a] SQD = Optimal Split Questionnaire Design, MSD= Matrix Sampling Design

[b] γ is the fraction of missing information

3.10 Empirical Data Application

We apply our procedure to a previously published empirical dataset obtained with the "Project 2000 Ninth GVU Survey Web Attitude and Perceptions Questionnaire"[9], which assesses how people use the Web and

[9] http://elab.vanderbilt.edu/research/topics/flow/project2000.gvu9.htm

their attitudes towards using it (Novak, Hoffman, and Yung, 2000). This type of survey, applied repeatedly to the same panel for the purpose of tracking consumer attitudes and behavior, may benefit from application of split questionnaire designs since it is conducted on a regular basis with an almost identical structure. Although this particular application is less than ideal to illustrate the performance of SQD, since the questionnaire is relatively short, we consider the use of a published questionnaire and publicly available data attractive. There are sixty-five questions, grouped into nine blocks according to content. The first block contains five questions about the role of the Web in life, the second block consists of eight questions on feelings while using the Web, the third block is composed of five questions related to Web activities, there are seven questions in the fourth block about perceptions on using the Web, the fifth block consists of seven questions about attitudes about using the Web, the sixth block contains eight questions about people feelings towards using the Web, the seventh and eighth block are comprised of ten and nine questions, respectively, about attitudes and perceptions, and the last block contains questions on flow and usage of Web information. The questions are assessed on 9-point Likert scales and are considered to be continuous and normally distributed for the purposes of the present study.

Data are available for two waves of the study conducted in two consecutive years. We use these as initialization and validation data, containing 500 and 1150 respondents, respectively. All data are complete. The advantage of having access to complete data is that it allows us to assess the performance of the SQD. A disadvantage of using such complete data is that we may underestimate the effect of the split

questionnaire design, since we do not benefit from the advantages of improved quality of the responses due to reduced respondent burden. Therefore, we also construct a field study with this questionnaire on which we report in the final part of this chapter. The initialization data are derived from the first wave of the survey, which we use for creating the split questionnaire. From the initialization data, we calculate the complete data parameter estimates. This enables us to obtain the design using the Federov algorithm to minimize the Kullback-Leibler distance. We investigate the following designs, where all designs in this study are constructed to be fully identified:

a) Optimal split questionnaire (SQD) and matrix sampling designs (MSD),

b) Designs with five or ten splits,

c) Between-block and within-block designs,

d) Unconstrained or constrained designs.

We consider the MSD (matrix sampling design) as a benchmark for the between-block design. For the within-block SQD, we use as benchmarks a random questionnaire design (RQD, in which questions within blocks are randomly assigned) as well as an ad-hoc procedure based on a principal components analysis of the items, as explained in more detail below. We use about the same total number of questions in all designs. We generate the MSD by randomly choosing five or ten splits from the candidate split matrix and evenly distributing them among respondents, eliminating responses from the complete data matrix Y according to the design in

question. For the RQD we apply the same procedure for each block separately, each time randomly selecting splits from the candidate split set. Since we have access to the complete data, we apply the constructed designs to those data to generate the missing data pattern. To compare the designs, we compute the KL distance and BIC statistics, the fraction of missing information, and MAD and RMSE, after imputing the missing data with the Gibbs sampler. We use informative priors obtained from the initialization data, for all designs. We run the Gibbs for 3000 iterations and save the last 600 draws from the predictive distribution for Y_{mis} as imputations; iteration plots show that the chains converge well before the end of the burn-in period.

3.10.1 Between-Block Designs

The MAD and RMSE measures shown in Table 3.4 reveal that the estimated parameters for the optimal SQD design are close to the complete data parameters. For both the five- and ten-split cases, the SQD improves significantly over the MSD, the MAD being 35% and 45% smaller respectively, and RMSE 34% and 45%. The improvement of the optimal designs over the currently used matrix sampling designs is substantial. The reason for the better performance of the five-split design, which results in 32% lower MAD and 31% lower RMSE than the ten-split design, is that the lower number of splits is associated with a smaller percentage of missing questions. For this particular application, the five-split optimal SQD results in a reduction of around 66% of the questions, with only a 14% information loss. With ten splits we obtain a greater reduction in the number of questions as compared to five splits. Here, while the SQD results in a 14%

loss of information, for the MSD the fraction of missing information is larger, 18%. The split questionnaires with five and ten splits are provided in Figure 3.3.

Table 3.4: Comparison of designs on empirical data

BETWEEN-BLOCK DESIGNS								
	Unconst. 10 Splits		Const. 10 Splits-5Blocks/Split		Unconst. 5 Splits		Const. 5 Splits-5 Blocks/Split	
	SQD	MSD	SQD	MSD	SQD	MSD	SQD	MSD
MAD	0.265	0.483	0.169	0.197	0.18	0.277	0.148	0.159
RMSE	0.378	0.682	0.24	0.319	0.262	0.399	0.203	0.215
γ	0.143	0.182	0.074	0.134	0.140	0.170	0.089	0.109
%Missing	0.735	0.735	0.492	0.492	0.662	0.662	0.440	0.440
BIC	18410	30298	57284	57655	15070	38740	64489	64675
logL(D)	-9195	-15139	-28631	-28817	-7525	-19360	-32234	-32327

WITHIN-BLOCK DESIGNS						
	10 splits			5 splits		
	SQD	RQD	PCA	SQD	RQD	PCA
MAD	0.156	0.163	0.164	0.125	0.125	0.129
RMSE	0.227	0.243	0.251	0.201	0.211	0.216
γ	0.078	0.087	0.084	0.056	0.060	0.058
%Missing	0.515	0.515	0.515	0.406	0.406	0.406
BIC	44134	45186	45085	54890	55126	54979
logL(D)	-22056	-22582	-22532	-27434	-27552	-27479

[a] SQD = optimal Split Questionnaire Design, MSD= Matrix Sampling Design, RQD = Random Questionnaire Design, PCA = Principal Components Design

In addition, we investigate the case where constraints are imposed on the SQD. In particular, we construct designs in which each split consists of exactly five blocks. We choose this number, since we need at least five splits to generate fully identified designs under the constraint of five blocks per split. We repeat the design construction and imputation procedure on the empirical data, using five and ten splits, fixing each split to contain five

blocks. The results are given in Table 3.4. We focus first on the five-split design. In this case we reduce the number of questions with about 44%, while it was 66% for the unconstrained SQD. As a result, the constrained SQD yields 9% of missing information, while the unconstrained SQD yields 14% of missing information (these numbers are 7% and 14%, respectively, for the ten-split SQD). The fraction of missing information is also less for the constrained SQD than for the constrained MSD, as expected, but the logL(D) and BIC for the constrained designs are worse than for the unconstrained designs. The RMSE and MAD measures reveal that the SQD estimates are close to those of the complete data, these measures are even smaller that for the unconstrained design. They are better than for the comparable MSD's, although the differences are smaller than for the unconstrained designs. The reason is that the constraints strongly limit the degrees of freedom for improvement over the MSD, since they reduce the size of the candidate split set. The optimal constrained five and ten-split designs are shown in Figure 3.4.

3.10.2 Within-block Designs

Using the prior estimates from the initialization data, we also construct optimal within-block designs by selecting questions within blocks, as described above. We compare the optimal SQD with designs in which the questions within blocks are selected randomly (RQD). To also compare to a stronger benchmark, we construct split designs using principal component analysis (PCA)[10]. We extract five and ten Varimax rotated components to

[10] We acknowledge an anonymous reviewer for this suggestion.

construct the splits. Questions in a block are discarded for a split if they contribute the least variance for that component. Every question was included at least once, and the design has the same number of questions as the SQD and RQD designs.

The results are shown in Table 3.4. We reduce 41% and 52% of the questions with the five- and ten- split within-block designs. The BIC and KL-distance of the optimal within-block designs are lower than the random design and the principal components design. The optimal within-block designs are also somewhat better in terms of RMSE and MAD of the parameter values, but the differences are not as large as for the between-block designs. The PCA designs are in between the RQD and optimal SQD on these measures. The average percentage of missing information is around 7.8% and 5.6% respectively for the optimal five- and ten-split designs. These numbers are better than for the corresponding random designs, with 8.7% and 6.0% respectively, and for the PCA designs, with 8.4% and 5.8%, respectively. The fraction of missing information for within-block designs, however, is substantially lower than for the between-block designs. MAD and RMSE of the five-split within-block designs are 31% and 23% lower than those of the between-block designs. For the ten-split designs they are 41% and 40% lower than those of the between-block design. However, the MAD and RMSE of the within-block designs are comparable to those of the constrained between-block designs. The optimal within-block designs are shown in Figure 3.5.

The estimates of the variances of the responses to the questions for the prior data, full and split questionnaires (after imputation) are shown in Table 3.5. As can be seen from the table, the prior estimates are close to complete questionnaire estimates of the current study. This illustrates the value of such prior estimates for the construction of split designs, but we further investigate the sensitivity of the optimal between- and within-block designs to these prior parameter values. For this purpose, randomly draw 50 sets of values from the sampling distribution of the parameters obtained from the initialization data and obtain optimal ten-split unconstrained and constrained between-block designs and within-block designs based on each of these sets. On average, we found 9.7 splits to be the same across these replications for the unconstrained between-block design[11]. For the constrained ten-split between-block design we find a lower average number of corresponding splits, 5.5. For the within-block design, on average only 2.2 splits were the same. Clearly, the within-block design is much more sensitive to the choice of the prior than the between-block designs. The size of the full candidate split set, as well as the use of the greedy design generating algorithm contribute to the high prior sensitivity of the within-block design. In particular, we find the sensitivity of the between-block design to the prior specification highly satisfactory.

[11] The maximum is 10, if all prior values yield exactly the same design, since there are ten splits in the design.

Table 3.5: Variance estimates after imputation[1]

	Full Wave 1	Full Wave 2	Between Con. SQD	Between Unc. SQD	Within SQD		Full Wave 1	Full Wave 2	Between Con. SQD	Between Unc. SQD	Within SQD
1	2.29	2.27	2.27	2.16	2.36	34	3.19	3.09	3.45	3.47	3.64
2	2.56	2.38	2.38	2.34	2.54	35	3.41	3.51	3.96	3.95	4.29
3	1.92	2.22	2.22	2.16	2.20	36	2.17	1.78	1.88	1.88	1.84
4	1.96	2.13	2.13	2.14	2.08	37	1.96	1.89	2.14	2.07	1.98
5	2.33	2.15	2.15	2.19	2.41	38	2.49	2.47	2.76	3.03	2.58
6	4.35	4.14	4.66	5.02	4.33	39	1.87	1.84	2.06	1.92	1.88
7	2.63	2.36	2.50	2.74	2.34	40	2.04	2.29	2.65	2.07	2.49
8	2.82	2.81	3.02	2.97	3.17	41	2.80	2.79	2.93	4.54	3.16
9	2.29	2.52	2.79	2.49	2.76	42	3.85	4.00	4.45	5.19	3.95
10	1.69	1.89	1.98	1.98	2.19	43	2.90	2.89	3.12	3.97	3.08
11	3.08	3.11	3.30	4.01	3.13	44	4.59	4.34	4.85	7.28	4.41
12	2.42	2.51	2.74	2.72	2.88	45	4.13	4.04	4.66	4.97	3.96
13	2.69	2.28	2.71	2.71	2.32	46	2.95	2.92	3.37	4.76	3.02
14	1.86	2.18	2.18	2.31	2.24	47	3.04	3.20	3.75	4.52	3.52
15	3.77	3.62	3.72	3.87	4.03	48	4.87	4.60	5.64	6.23	4.66
16	4.31	3.87	4.05	3.91	4.08	49	4.86	4.66	4.77	6.09	4.70
17	5.48	4.62	4.74	4.66	5.21	50	3.77	3.96	4.66	5.84	4.51
18	3.25	3.54	3.60	3.67	3.59	51	3.05	3.05	3.11	3.49	3.19
19	4.97	4.62	5.25	5.23	5.03	52	2.13	2.22	2.52	2.96	2.25
20	4.79	4.86	5.29	5.63	6.46	53	5.38	5.48	5.85	7.17	5.50
21	3.08	2.91	3.08	3.55	3.06	54	4.89	4.59	5.19	6.51	5.00
22	2.87	2.90	3.07	2.99	3.24	55	3.19	3.47	4.25	5.62	3.44
23	3.06	3.43	3.59	4.09	3.61	56	5.03	4.67	5.19	7.20	4.79
24	5.22	5.36	6.07	6.03	5.75	57	2.94	3.12	3.44	4.03	3.29
25	2.27	2.04	2.28	2.53	2.28	58	2.93	2.77	2.99	3.78	2.94
26	4.40	4.08	4.30	4.53	4.40	59	3.52	3.60	3.81	5.00	3.58
27	5.33	5.49	5.97	6.34	5.49	60	6.78	6.74	7.07	7.81	6.91
28	3.66	4.20	4.59	5.24	4.35	61	4.54	4.68	4.87	5.26	4.75
29	2.76	2.96	3.08	3.20	3.02	62	5.21	5.23	5.37	5.97	5.56
30	4.83	4.87	5.42	6.04	5.10	63	1.07	1.12	1.19	1.23	1.27
31	3.45	3.88	4.59	5.37	3.97	64	1.66	1.84	1.85	1.95	1.93
32	4.12	4.25	4.65	5.51	4.64	65	0.56	0.53	0.55	0.60	0.52
33	1.43	1.41	1.71	1.61	1.60						

[1] From the full questionnaire from the first and second wave survey, the constrained and unconstrained between- and within ten-split optimal designs

3.11 Field Study

The above analysis illustrates that optimally designed split questionnaires can be beneficial, but only address that issue from a statistical perspective. In this section, we look into the behavioral issues of providing subjects split questionnaires. We conducted a field experiment to investigate whether with split questionnaires one may reduce boredom, fatigue, and completion time, which ultimately should increase the quality of data. We will also investigate respondents' attitudes towards the questionnaires, and assess whether using split questionnaires improves the reliability of constructs, compared to the full questionnaire.

For the field study, we use the exact questionnaire that was used in the empirical study above. We asked additional questions about boredom, which is scaled 1 (not at all bored) to 9 (extremely bored), fatigue which also is scaled 1 (not at all tired) to 9 (extremely tired). In addition, we assessed attitudes towards the questionnaire (three questions, five-point scale: repetitive-varied, very long-very short, boring-stimulating). We tested the full questionnaire, a ten-split between-block design, and a ten-split within-block design (see above) each on 63 subjects recruited from the subject pool from [withheld for confidentiality]. In total, 189 subjects responded to 21 versions of the questionnaire that were displayed on computer screens in the experimental lab. Computer aided questionnaires allowed us to record the exact time it took respondents to complete them. These average times to complete the full and split questionnaires differed significantly: 8 minutes for the complete questionnaire, and about 6 minutes

for each of the split questionnaires. This is a significant reduction of about 25% in completion time, with a 50% reduction in the number of questions. Note that even the full questionnaire with 65 questions can be completed relatively quickly --the longest it took any respondent was 10 minutes--, which makes it more difficult to identify the behavioral effects of the split questionnaires.

The mean scores for the scales are shown in Table 3.6. A MANOVA across all measures reveals a significant difference between the complete and between-block design ($p<0.01$) and the complete and within-block design ($p<0.01$), but not between the latter two. The mean boredom score for the full questionnaire is 5.44, which is significantly higher than that for the within-block questionnaire, which is 4.98. The differences with the between-block design, which has an intermediate boredom score of 5.23, are not significant. This may indicate that feelings of boredom are primarily caused by repetition of the relatively similar questions within blocks, which occurs less in the within-block design. The respondents that completed the full questionnaire report feeling more tired than those receiving the between-block design, the mean scores being 4.32 and 3.57. The within-block tiredness score is intermediate, 3.73, and not significantly different from the other two. This may point to feelings of tiredness being more strongly associated with switching between different topics, which occurs less often in the between-block design due to a reduction of the number blocks. The split questionnaire designs are evaluated more favorably than the complete questionnaire, the between- and within-block designs being

seen as less repetitive (5.32 and 4.20 versus 5.68) and less boring (4.77 and 4.42 versus 4.94) than the complete questionnaire. The scores for the within-block design are significantly better than those for the between-block design. The within-block design is also considered to be significantly less long than the complete questionnaire design (3.13 versus 3.68; and 3.54 for the between-block design, which is not significantly different from the former two). The shorter perceived duration of the within-block design may be associated with its lower perceived boredom discussed above, since its actual duration is about 20 seconds longer than that of the between-block design (the longer duration may have to do with the reading and processing of the separate instructions for each block).

Table 3.6: Item means and SDs from the field experiment

	FULL QUESTIONNAIRE	BETWEEN-BLOCK SQD	WITHIN-BLOCK SQD
Duration	476.92	344.48[a1]	364.02[b1]
	(95.01)	(146.552)	(93.57)
Boredom	5.44	5.23	4.98[b1]
	(2.09)	(1.95)	(2.00)
Fatigue	4.32	3.57[a2]	3.73
	(2.55)	(2.27)	(2.02)
Repetitive	5.68	5.32[c1]	4.70[b1c1]
	(1.37)	(1.22)	(1.78)
Long	3.68	3.54	3.13[b1]
	(1.54)	(1.56)	(1.25)
Boring	4.94	4.77[c1]	4.42[b1c1]
	(1.28)	(1.11)	(1.25)
Cronbach's alpha	0.66	0.66	0.67
Item Variance	3.34	2.36[a1]	2.30[b1]

Notes: The values in parenthesis are standard deviations. N=189. Duration mean values are in seconds. Superscripts indicate the significance of the differences between means of the full & between- ([a]), full & within- ([b]) and between- & within- ([c]) block questionnaires; [1] p=0.05 , [2] p=0.10

In short, split questionnaire designs decrease completion time, fatigue, boredom and non-response and are evaluated more positively by respondents, where it seems that the within-block design has a somewhat more favorable behavioral effect than the between-block design. These effects may impact the quality of the data. For each of the three questionnaires, respondents could skip every question displayed on the screen. There were 33 skip-responses for the full questionnaire, 7 for the

between- and 5 for the within-block design. These responses start only after the first twelve questions and mostly occur in the last half of the questionnaires. This indicates that the use of split questionnaires may substantially reduce item non-response. Second, the effect of the questionnaire design on the average item variances and Cronbach's alpha were investigated. The questionnaire consists of 13 constructs that are each measured with several items. There were no statistically significant differences in the average Cronbach's alpha, estimated after multiple imputation of the missing data of the between- and within-block split questionnaire designs. However, we did find significant differences in item variances between the full- and split questionnaire designs. The differences between between-block and within-block designs are not significant. The average item variance for the full questionnaire is 3.34, which is significantly higher than for the between-block design, with 2.36, and the within-block design, 2.30. This means that subjects who answered the questions in the within-block or between-block designs responded to the items that measure the same construct more consistently. Thus, the quality of the data we obtained from the between-and within-block split questionnaire designs tends to be better than that of the full questionnaire. Again, we note that with a maximum average completion time of eight minutes, the complete questionnaire is relatively short. For longer questionnaires, the effects may be even larger.

3.12 Conclusion

Split questionnaires present opportunities for application in consumer panels, offering the potential to obtain higher quality information from respondents faster and at a substantially lower cost. In this chapter, we first propose a methodology to split questionnaires optimally into sub-components with minimal information loss by applying optimal experimental design methods. We proposed the Kullback-Leibler distance as a design criterion, applied the modified Federov algorithm to search over the design space, and illustrated that good designs can be constructed rapidly in spite of the demanding task. Split questionnaire designs were shown to have desirable statistical and behavioral properties, relative to complete questionnaires or questionnaires constructed with ad-hoc methods.

We have investigated two different types of split questionnaire designs based on the contextual placement of questions in blocks. The first method, producing between-block designs, places blocks as a whole into different split versions of the questionnaire. Optimizing the allocation of the blocks across the splits is a much more feasible task than allocating individual questions to splits. Additional constraints, such as on the number of blocks per split, can easily be accommodated and may further reduce the number of questions asked from each respondent. Between-block designs result in estimates close to those obtained from the complete data, reducing completion time and respondent fatigue. The second method, producing within-block designs, is based on choosing questions in each block. For

these designs, the optimization task is very demanding, so that we needed to use a greedy algorithm to find the optimal design. As a consequence, the within-block designs are not strictly optimal, nor can they easily be constructed to be fully identified. However, they do provide improved efficiency, yielding parameter estimates that are closer to the complete data estimates and less missing information than designs constructed with heuristic procedures. Their performance in terms of parameter estimates and missing information tends to be better than that of the between-block designs, but they are substantially more sensitive to the values of the prior estimates.

Our field study shows that the behavioral reaction of respondents to split questionnaires is more favorable than to the complete questionnaire, in terms of duration, boredom, and fatigue, amongst others. The response to within-block designs tends to be more positive than that to a between-block design, since respondents feel less bored, and perceive the questionnaire as less long, boring and repetitive. The between-block design, however, results in less respondent fatigue. The choice between the within-block and between-block designs may therefore be based on either statistical or behavioral criteria. From our investigation, it appears that the between-block design has better statistical properties, since it is feasible to construct fully identified designs with little sensitivity to the prior estimates. However, the within-block design still performs quite satisfactorily, yields parameter estimates comparable to constrained between-block designs, and elicits a more positive reaction from respondents. However, the high sensitivity of these designs to prior estimates warrants further study.

The validity of the prior knowledge when constructing the split questionnaire design is an important issue. Whereas prior knowledge can be easily obtained in panel or tracking surveys conducted on a regular basis with almost identical questions and blocks, it may be less easy to obtain in other settings. In those cases, subjective prior distributions for the model parameters can be assessed, which in many cases would involve the elicitation of priors from consumers, decision makers or other subject-matter experts. Chaloner (1996) provides an overview of the various approaches to elicitation based on the ways people think about and update probabilistic statements. It is of interest to consider prior uncertainty on the parameters in constructing the designs, and to construct designs integrating the design criterion over the prior distribution of the parameters (Sandor and Wedel, 2001). This may in particular be worthwhile for within-block designs, which were revealed to have high sensitivity to the prior specification. For between-block designs, in particular in panel data applications such as the one presented above, this may not be needed, since the prior parameter values can be fairly precisely estimated from the available pilot data, and the designs themselves were shown to be quite insensitive to the prior parameter values. We leave these issues for future research.

3.13 Appendix

3.13.1 KL-Distance for Mixed Data

The incomplete data log-likelihood of mixed data is derived below using the general location model (Olkin and Tate, 1961; Krzanowski, 1983). We have the data matrix $Y_{N\times(p+q)}=(X,Z)$, where $X=(X_1,..,\ X_p)'$ and $Z=(Z_1,\ldots,Z_q)$ represent the continuous and categorical variables, respectively. Each column variable in Z, z_j has c_j levels, and these categorical variables form a q-dimensional contingency table with a total number of cells $C=\prod_{j}^{q} c_j$. The frequencies in this table are contained in $W = (w_{c_1}, w_{c_2},.....w_{c_q})$. The marginal distribution of the categorical variable Z is multinomial distribution $(w \mid \pi=(\pi_1,\pi_2,.....\pi_c)' \sim$multinomial (π) with $\sum_{i=1}^{C}\pi_i=1)$ and the conditional distribution of the continuous variables X given categorical variables Z (i.e. given a particular cell) is multivariate normal with different means across the cells defined by the categorical variables, but with a common covariance matrix $(X \mid Z=w,\mu_i,\Sigma \sim N(\mu_i,\Sigma)$, where μ_i is the mean of X in the cell specified by z, and Σ is the common conditional covariance of X across cells of the contingency table). The KL-distance in this case reduces to the

incomplete-data log-likelihood:

$$L\ (\mu_z, \Sigma, \pi \mid D) = \sum_{i=1}^{N} \log f(x_{i,obs} \mid C_i, \mu_z, \Sigma) + \sum_{c \in C_i}^{N} \log f(\pi_c)$$

$$= -\frac{1}{2}\sum_{i=1}^{N} p_i \log(2\pi) - \frac{1}{2}\sum_{i=1}^{N} \log\left|\Sigma_{i,obs}\right|$$

$$+ \sum_{i=1}^{N} \log\left[\sum_{c \in C_i} \pi_i \exp\left\{-\frac{1}{2}(X_{i,obs} - \mu_{i,obs,c})'\Sigma_{i,obs}^{-1}(X_{i,obs} - \mu_{i,obs,c})\right\}\right] \quad \textit{(A1)}$$

where $X_{i,obs} = X_{ij}d_{ij}$, where d_{ij} is the element of design matrix D.

3.13.2 Missing Information Principle

Assume that f(Y|θ) is the probability distribution of $Y = (Y_{obs}, Y_{mis})$ and parameter θ. The distribution of the complete data Y can be factored with f(Y_{obs}|θ), the density of the observed data Y_{obs}, and f(Y_{mis}|Y_{obs},θ), the density of the missing data given the observed data, is represented as

$$f(Y \mid \theta) = f(Y_{obs}, Y_{mis} \mid \theta) = f(Y_{obs} \mid \theta) f(Y_{mis} \mid Y_{obs}, \theta) \quad \text{(B.1)}$$

The decomposition of the loglikelihood that corresponds to (B.1) is

$$\ell(\theta \mid Y) = \ell(\theta \mid Y_{obs}, Y_{mis}) = \ell(\theta \mid Y_{obs}) + \ell(Y_{mis} \mid Y_{obs}, \theta) \quad \text{(B.2)}$$

Since directly maximizing the incomplete-data likelihood *l*(θ|Y_{obs}) with respect to θ for fixed Y_{obs} to estimate θ can be difficult, we can write B.3 with the observed loglikelihood *l*(θ|Y_{obs}), the complete-data loglikelihood *l*(θ|Y), and the missing part of the complete-data loglikelihood *l*f(Y_{mis}|Y_{obs},θ)

$$\ell(\theta \mid Y_{obs}) = \ell(\theta \mid Y) - \ell f(Y_{mis} \mid Y_{obs}, \theta) \quad \text{(B.3)}$$

The observed information matrix *I*(θ|Y_{obs}) can be found directly by differentiating the loglikelihood *l*(θ|Y_{obs}) twice with respect to θ. Alternatively, differentiating *l*(θ|Y_{obs}) twice with respect to θ yields for any Y_{mis}

$$\ell(\theta \mid Y_{obs}) = \ell(\theta \mid Y_{obs}, Y_{mis}) + \frac{\partial^2 \ln f(Y_{mis} \mid Y_{obs}, \theta)}{\partial\theta\partial\theta}, \quad \text{(B.4)}$$

where $I(\theta|Y_{obs},Y_{mis})$ is the observed information based on $Y=(Y_{obs},Y_{mis})$ and the negative of the last term is the missing information from Y_{mis}. Taking expectations over the distribution of Y_{mis} given Y_{obs} and θ yields

Observed Information=Complete Information-Missing Information (B.5)

The observed information equals the (conditional expected) complete information minus the missing information. This has been called the missing information principle by Orchard and Woodbury (1972).

The decomposition of the observed information is particularly simple in case the complete data Y have a distribution from the regular exponential family defined by

$$f(Y \mid \theta) = b(Y)\exp(s(Y)\theta)/a(\theta) \tag{B.6}$$

where θ denotes a ($d\times1$) parameter vector, s(Y) denotes a ($1\times d$) vector of complete-data sufficient statistics, and a and b are functions of θ and Y, respectively. The complete information is $Var(s(Y)|\theta)$, and the missing information is $Var(s(Y)|Y_{obs},\theta)$. Thus the observed information is the difference between the unconditional and conditional variance of the complete-data sufficient statistic.

$$\ell(\theta \mid Y_{obs}) = Var(s(Y) \mid \theta) - Var(s(Y) \mid Y_{obs},\theta), \tag{B.7}$$

In sum, according to the missing data information principle, the missing information is equal to the variance difference between the complete data and the incomplete data. In questionnaire design, the missing information measures the increase in variance of estimation due to nonresponse, and

is determined by response rates and the ability of observed values to predict missing values successfully.

3.13.3 Figures

Description of Blocks:

Block 1: Five questions about the role of the Web in life.

Block 2: Eight questions about the feeling while using the Web

Block 3: Five questions related to the Web activities feeling while using the Web

Block 4: Seven questions about and perceptions on using the Web

Block 5: Seven questions about attitudes and perceptions on using the Web

Block 6: Eight questions about peoples' feelings towards using the Web

Block 7: Ten questions on attitudes and perceptions

Block 8: Nine questions about attitudes and perceptions on using the Web

Block 9: Six questions about flow and usage of the web.

Figure 3.3: Optimal Unconstrained Between-Block Designs for the Empirical Data

THE OPTIMAL 10-SPLIT UNCONSTRAINED BETWEEN-BLOCK SQD

Resp.No.	Block 1 Q1-5	Block 2 Q6-13	Block 3 Q14-18	Block 4 Q19-25	Block 5 Q26-31	Block 6 Q32-40	Block 7 Q41-50	Block 8 Q51-59	Block 9 Q60-65
1-115	■	■	■	■	■	■	■	■	■
116-230	■	■							
231-345	■		■						
346-460	■			■					
461-575	■				■				
576-690	■					■			
691-805	■								■
806-920			■	■					
921-1035			■			■			
1036-1150			■						■

THE OPTIMAL 5-SPLIT UNCONSTRAINED BETWEEN-BLOCK SQD

Resp.No.	Block 1 Q1-5	Block 2 Q6-13	Block 3 Q14-18	Block 4 Q19-25	Block 5 Q26-31	Block 6 Q32-40	Block 7 Q41-50	Block 8 Q51-59	Block 9 Q60-65
1-230	■	■	■	■	■	■	■	■	■
231-460	■		■						
461-690	■					■			
691-920	■								■
921-1150			■						■

Note: shaded are observed, blank are missing blocks.

Figure 3.4: Optimal Constrained Between-Block Designs for the Empirical Data

THE OPTIMAL 10-SPLIT 5-BLOCK BETWEEN-BLOCK SQD

Resp.No.	Block 1 Q1-5	Block 2 Q6-13	Block 3 Q14-18	Block 4 Q19-25	Block 5 Q26-31	Block 6 Q32-40	Block 7 Q41-50	Block 8 Q51-59	Block 9 Q60-65
1-115	■	■	■	■					■
116-230	■	■	■			■			■
231-345	■	■	■		■				■
346-460	■	■			■	■	■		
461-575	■	■			■	■		■	
576-690	■		■	■	■				■
691-805	■		■	■				■	■
806-920	■		■	■		■			■
921-1035	■		■	■			■		■
1036-1150	■		■				■	■	■

THE OPTIMAL 10-SPLIT 5-BLOCK BETWEEN-BLOCK SQD

Resp.No.	Block 1 Q1-5	Block 2 Q6-13	Block 3 Q14-18	Block 4 Q19-25	Block 5 Q26-31	Block 6 Q32-40	Block 7 Q41-50	Block 8 Q51-59	Block 9 Q60-65
1-230	■	■	■	■					■
231-460	■	■	■		■	■			
461-690	■	■	■				■	■	
691-920				■	■	■	■	■	
921-1150					■	■	■	■	■

Note: shaded are observed, blank are missing blocks.

Figure 3.5: Optimal Within-Block Designs for the Empirical Data

THE OPTIMAL 10-SPLIT WITHIN-BLOCK SQD

Bl. 1 Q1-5	**Bl. 2 Q6-13**	**Bl. 3 Q14-18**	**Bl. 4 Q19-25**	**Bl. 5 Q26-31**	**Bl. 6 Q32-40**	**Bl. 7 Q41-50**	**Bl. 8 Q51-59**	**Bl. 9 Q60-65**
00110	00000101	00101	1100000	0011101	00011111	0110001100	01111110	111010
11111	11111111	11111	1000100	1101010	01000010	0011100100	11100011	011111
00011	10000001	10100	1010000	0101010	00010001	1100110100	01101111	100010
10010	01000100	00101	0010001	0111110	11011110	0111110010	10011100	110010
10100	01010000	00101	1000010	1001001	10101100	1110111010	11001101	010011
01100	10010000	00011	0010010	1101110	10010010	0100011110	01011101	001101
00101	00101000	11000	1000010	1110101	00011100	1001111101	00101100	110101
11000	10010000	10010	0010001	1100110	01011110	1011011011	11001110	010001
01001	01000100	00011	0011000	0010011	11100011	1110010111	01100000	110111
10001	00110000	10100	0001001	0111100	10110111	0000000011	10101101	110011

THE OPTIMAL 5-SPLIT WITHIN-BLOCK SQD

Block 1 Q1-5	**Block 2 Q6-13**	**Block 3 Q14-18**	**Block 4 Q19-25**	**Block 5 Q26-31**	**Block 6 Q32-40**	**Block 7 Q41-50**	**Block 8 Q51-59**	**Block 9 Q60-65**
00110	01101101	00110	1010110	1111111	01100100	1110011101	00010101	101011
10010	10000101	11111	0101100	1101110	11110111	0000110011	01111110	011101
11111	00100011	01110	1011110	1101100	11011000	0111101011	11111111	100111
10100	01001111	00011	1101001	1001101	11101111	1101101110	10000110	110010
00101	11111111	10001	0110111	0101010	11001010	0001011001	00011100	001011