A Split Questionnaire Survey Design applied to German Media and Consumer Surveys

Susanne Rässler, Florian Koller, Christine Mäenpää
Lehrstuhl für Statistik und Ökonometrie
Universität Erlangen-Nürnberg
Lange Gasse 20
D-90403 Nuremberg
Germany
Email: susanne.raessler@wiso.uni-erlangen.de
florian.koller@gfk.de

Diskussionpapier 42 a / 2002

Abstract

On the basis of real data sets it is shown that splitting a questionnaire survey according to technical rather than qualitative criteria can reduce costs and respondent burden remarkably. Household interview surveys about media and consuming behavior are analyzed and splitted into components. Following the matrix sampling approach, respondents are asked only the varying subsets of the components inducing missing data by design. These missing data are imputed afterwards to create a complete data set. In an iterative algorithm every variable with missing values is regressed on all other variables which either are originally complete or contain actual imputations. The imputation procedure itself is based on the so-called predictive mean matching. In this contribution the validity of split and imputation is discussed based on the preservation of empirical distributions, bivariate associations, conditional associations and on regression inference. Finally, we find that many empirical distributions of the complete data are reproduced well in the imputed data sets. Concerning these long media and consumer questionnaires we like to conclude that nearly the same inference can be achieved by means of such a split design with reduced costs and minor respondent burden.

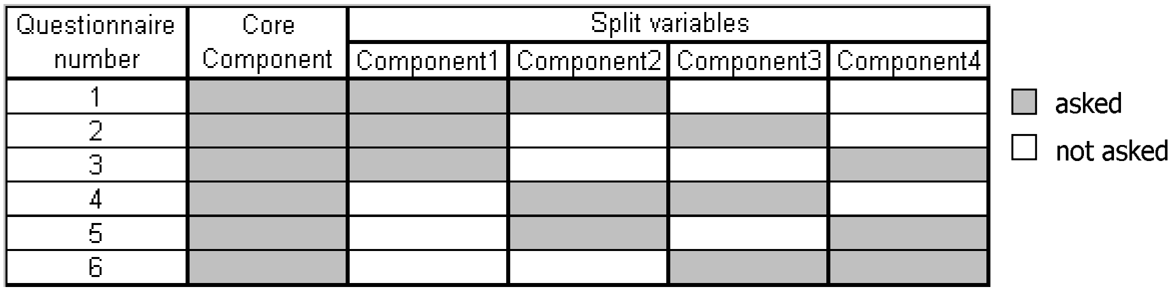
1 Introduction

Owing to price competitiveness market research institutes face the problem of deriving an increasing amount of information from data sources at steadily decreasing costs. Moreover, in view of the growing importance of multi-media market investigations which include many magazines, tv-viewing behavior and purchasing behavior, the problem of excessively long questionnaires has been arising with increasing frequency, along with the related problem of declining response rates and an increased investment of time for the respondent and the interviewers, for an early discussion see Tennstädt (1987). Since questionnaire based surveys are widespread means to gather all this information, it is hardly surprising that a variety of methods like split questionnaires with rotational elements have been developed to solve the above problem. However, these approaches usually lead to reduced sub samples of the original data set such that in some cases the sample size for multi-dimensional analyses gets very small.

Raghunathan and Grizzle (1995) introduced a split questionnaire survey design where the original questionnaire is divided into several components with each component containing a roughly equal number of questions. The split approach is based on the multiple matrix sampling design which has long been used in US educational testing, achievement testing, and program evaluation, see Shoemaker (1973), Holland and Rubin (1982), or Munger and Lloyd (1988). With multiple matrix sampling basically there are subgroups of variables created randomly and these subgroups are randomly assign to subgroups of units; these random assignments can lead to estimation problems due to non-identification and highly reduced data sets for multivariate analysis. In the split questionnaire survey design, apart from a core component with questions that are considered to be vitally important (e.g., sociodemographic questions), also only a selection of the other components is administered to every interviewee. This clearly reduces interview time, yielding lower survey costs as well as reducing the respondent burden. But unlike the matrix sampling approaches, the missing data now are imputed to finally end up again with a complete(d) data set. This approach merely requires that combinations of variables which are to be evaluated must be jointly observed in a small sub sample (to avoid estimation problems due to non-identification). Thus, depending on the split design any desired analysis can be carried out while retaining the original sample size.

To illustrate this Figure 1.1 shows a split questionnaire survey (SQS) design where interactions of second order among the split variables are assumed to be analyzed. In our exemplary design the questionnaire is divided into four components (plus the core component with questions administered to all sample individuals).

Figure 1.1: Split questionnaire design with four components



Questionnaire number	Core Component	Split variables			
		Component1	Component2	Component3	Component4
1	■	■	■	□	□
2	■	■	□	■	□
3	■	■	□	□	■
4	■	□	■	■	□
5	■	□	■	□	■
6	■	□	□	■	■

■ asked

□ not asked

In the case of Figure 1.1 a split design with two selected and four total components would generate six (4 choose 2) different questionnaires.

The general idea behind this approach is so appealing that Germany’s largest market research company, GfK AG, decided to apply split questionnaire designs to market-media data bases. In a first test run two surveys are examined. The first one is a marketing tracking survey with about 300 individuals between 14 and 49 years being interviewed about recollection of commercials and advertisement in general for TV channels and programmes. The survey is conducted on two days a week using CATI. The second one is a combination of face-to-face interviews and household books and is conducted once a year among 4000 individuals with focus on questions about varying possibilities of advertising and interest in specific product categories. The goal of this article is to show that the split questionnaire survey design can offer solutions to a wide range of questionnaire based surveys which suffer either or both from high costs or a high respondent burden. Although based on the work of Raghunathan and Grizzle (with some slight modifications) our article will focus even more on the practitioner’s perspective. For this reason the article structure reflects loosely the chronological order of a conducted SQS project. More details of the project are described in Koller (2001) and Mäenpää (2001).

In the next section we will cover the component structure of the questionnaire and also discuss the rules for the assignment of questions to components. Raghunathan and Grizzle (1995) state that variables with high partial correlation coefficients should go into different components. This sounds reasonable, because following this condition you avoid that variables which explain each other very well are always jointly missing for any observation. However, in some cases the purely data generated solution of the questionnaire design must be modified. We try to show how robust the imputation of the missing answers is to a violation of this condition. The third section describes the imputation algorithm we used for both surveys which is, unlike to Raghunathan and Grizzle (1995), not based on Bayesianly proper multiple imputations (see Schafer, 1997). Due to the needs of the marketing research company, one imputed data set had to be produced for being passed to their customers. German agencies are well equipped with computerised media planning tools but for the time being there is no possibility to include correct multiple imputation inference in this planning process.

The missing data are imputed using regression imputation and predictive mean matching (PPM) as introduced by Rubin (1986) and Little (1988). Section four is divided into two parts: After a description of the application and implementation of both SQS designs, some results of the imputed and analyzed data sets will be compared with the available case equivalents (without imputation), and, this being a simulation study where initially complete data sets have been “punched”, with results based on the complete data sets. Finally, the last section will resume some of the problems and pitfalls encountered and discuss possible solutions.

2 Component Structure

In the first section we have learned that all split variables should be administered to components such that you get high partial correlations for variables in different components. To fulfil this requirement we need a complete data set to generate the component structure. Surveys which are conducted on a regular basis with at least almost identical questions like panels or the considered tracking surveys are therefore especially suitable for SQS designs. Otherwise you have to draw a small sub sample (with the complete questionnaire answered) in advance in order to calculate the associations/correlations. The calculation of a suitable measurement of association can occure to be difficult, because the survey variables are often measured on several scales. Problems occur when the considered variables are of different scale, especially if one of the variables is ordinal. Hence, to simplify matters ordinal variables will be treated as metric and the Bravais-Pearson coefficient is used to calculate the correlation. The formula for the Bravais-Pearson coefficient is given by

$$-1 \leq r = \frac{\sum_{i=1}^{n}(x_i - \overline{x}) \cdot (y_i - \overline{y})}{n \cdot s_x \cdot s_y} \leq 1 \quad , \tag{2.1}$$

where x_i and y_i are values for the *i*th observation of the variables x and y, $\overline{x}$ and $\overline{y}$ are the corresponding means and s_x, s_y the corresponding standard deviations, n is the total sample number.

For two binary variables the Phi coefficient can be used; its formula is given by

$$-1 \leq \Phi = \frac{a \cdot d - b \cdot c}{\sqrt{(a+c) \cdot (b+d) \cdot (a+b) \cdot (c+d)}} \leq 1 \quad , \tag{2.2}$$

where a, b, c, and d each represent one quarter of a 2×2-table with a being in the top left and b being in the top right cell.

After obtaining the correlation matrix a cluster analysis can be used to generate the component structure. Instead of cases we cluster variables, and by using the correlation matrix rather than a distance matrix, those variables with low correlations will be put into the same cluster. This means that variables with high correlations end up in different clusters. These clusters represent the questionnaire components and statistical standard software can be used to calculate the required number of clusters/components. “Average linkage within groups” minimises the average Euclidean distances for variables within each cluster while maximising the distances for variables in different clusters. If the cluster sizes tend to vary heavily (and hence the length of the different questionnaires), “Ward” can be used as an alternative distance measure. Thus, a “reversed cluster analysis” is an easy way to generate the required component structure.

However, some questions are required to be in the same block or even in a specific order as the following example might demonstrate: Suppose you ask interviewees who own at least one computer (or claimed to do so), what kind of computer they have (“Do you own a PC?/…home computer?/…laptop?”) asking a singular question for each type of computer. The empirical distributions would probably change drastically if you leave out one of the categories or maybe even if you asked about home computers before you asked about PC’s, because many people would not distinguish between the different types, especially if one category doesn’t even occur at all.

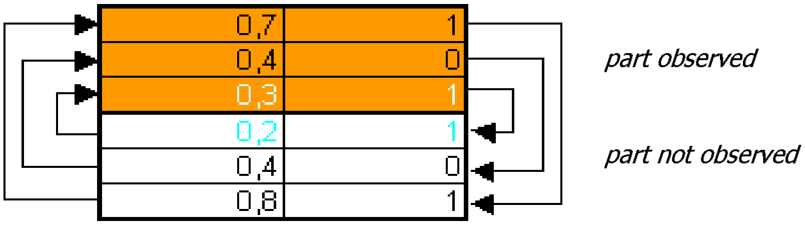
To test the robustness of the imputation to modifications in the data generated component structure we will use three different designs: the standard (“correct”) solution where the variables are clustered in the way described above, a solution where the assignment of questions to components is randomised and a “regular” cluster solution based on a distance matrix (worst case scenario) where variables with high correlations will be assigned to the same component. The results of the three designs are given in section four.

3 Imputation algorithm

Under the assumption that the missing data are missing at random (MAR) or even missing completely at random (MCAR), as defined by Rubin (1976, 1987) and Little and Rubin (1987), and that the parameters are distinct (see Schafer, 1997) the missing data mechanism is said to be ignorable. If these assumptions hold, it is possible to suitably impute the missing data by a standard multiple imputation technique. As the assignment of components to individuals is randomised the missing data mechanism can be treated to be MCAR, or, at least, MAR. However, as previously discussed, due to the customer’s demands we are not yet asked to multiply impute data sets. The production of one carefully imputed data set is the actual task.

The imputation of missing values is carried out by predictive mean matching which is basically both regression and nearest neighbour approach. Regression imputation (with rounding to the nearest observed value) tends to overestimate the explained sum of squares yielding a higher R^2 than the original data would do. In order to avoid this, predictive mean matching (PMM) combines regression imputation with nearest neighbour approaches. Instead of rounding the regression estimate to the nearest observed value, each case with an initially missing value scans the observed values to “find” the case with the closest regression estimate to its own regression estimate and adopts the corresponding observed value (see Figure 3.1).

Figure 3.1: Imputation algorithm (predictive mean matching).

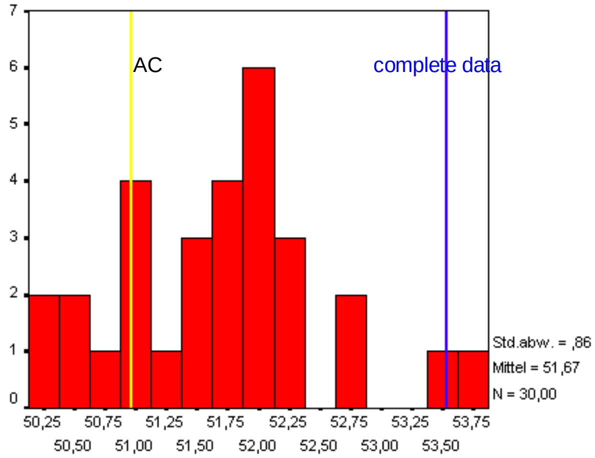


Predictive mean matching is more likely to preserve original sample distributions than rounding to observed values, because outliers like the high-lighted values in Fig. 3.1 do not necessarily change the structure of the sample. One minor drawback of PPM in this situation is that only “observed” rather than “possible” values can be imputed. However, in our surveys items usually have a very limited number of possible categories and therefore this problem can be safely ignored.

For a first run (starting solution) the computer algorithm includes all variables of the core component to generate initial estimates for all partly missing variables. Then, all variables (with exception of variables within the same component) are included in the regression, thus transporting any combination of variables implemented in the split design.[1] This second step will be repeated until the results for the imputed values converge to a certain level.

Although the algorithms does not contain a stochastic component it did not converge in any of the conducted test runs at a maximum number of 2000 iterations. One explanation might be the imputation of missing values based on outliers which might bias a converging process away rapidly within an iteration step (especially for binary variables). In order to generate one single reliable solution every tenth of a total of 300 iterations was saved as imputation yielding M = 30 imputations.[2] Figure 3.2 shows exemplary mean estimates of a split variable based on imputations, available cases and complete data set.

Figure 3.2: Histogramme for mean estimates of all 30 imputations.



Most of the mean estimates partly based on imputed values were closer to the complete data set estimator than the available case (AC) estimator. In order to identify the "best" imputation we would compare the deviations between imputation and complete data estimator for all considered estimators selecting the one imputation that yields a total minimum deviation sum of squares. However, the complete data estimator is, apart from the test situation, unknown, so our idea was to derive a chi^2-based measure which yields the deviation sum of squares to the mean over all imputation means. Let n be the number of total sample observations and q denote the number of considered variables. Then x_{ij} denotes the i-th observation of a split variable X_j with i = 1,2,…,n and j = 1,2,…,q. The estimated mean (or proportion for binary variables) of the w-th imputation is given by $\bar{x}_j^{(w)} = \frac{1}{n}\sum_{i=1}^{n}\hat{x}_{ij}$, with w = 1,2,…M. $\hat{x}_{ij}$ denotes either an observed or imputed value for X_j . The mean of all imputation means is given by

$$\hat{T}_j^{MI} = \frac{1}{M}\sum_{w=1}^{M}\bar{x}_j^{(w)} \quad , \qquad \text{for } j=1,2\ldots q, \qquad (3.1)$$

which shall be referred to as MI estimator, although the imputations themselves are not based on proper or Bayesianly proper multiple imputations (see Schafer, 1997).

[1] A special linux based tool (programming by Rässler, 2001) is used to generate the imputed data sets.

[2] Notice that the whole imputation procedure as presented herein is deterministic, no random values are drawn at any time.

In order to avoid that deviations of parameter estimates based on small sample sizes have a strong influence on the value of the measure a limit L is used to weigh down those deviations. Thus, we obtain a chi^2-like measure

$$G^{(w)} = \sum_{j=1}^{q} \frac{\left[n \cdot \left(\overline{x}_j^{(w)} - \hat{T}_j^{MI}\right)\right]^2}{n \cdot \hat{T}_j^{MI}} \quad \text{, for } n \cdot \hat{T}_j^{MI} \geq \text{L} \qquad (3.2a)$$

and

$$G^{(w)} = \sum_{j=1}^{q} \frac{\left[n \cdot \left(\overline{x}_j^{(w)} - \hat{T}_j^{MI}\right)\right]^2}{L} \quad \text{, for } n \cdot \hat{T}_j^{MI} < \text{L} \qquad . \qquad (3.2b)$$

yielding weighted deviations from the MI estimators of all considered parameter estimates.

The results presented in section four are based on the imputation that yielded the minimum value for G$^{(w)}$ with L = 30. The imputed data set so chosen was always either the “best” or, at least, among the “best” imputations when compared to the actual complete data estimates rather than the MI estimators.

4 Results

For the first survey we examined data from ten successive months with the earliest data being used to calculate the component structure. As mentioned earlier we wanted to examine possible effects of modifications within the component structure. The results for the three different designs described in section 2 for the average correlations yield

“correctly” calculated component structure:	0.22 (within)	0.27 (between)
random assignment to component :	0.24	0.24
“worst case scenario”:	0.34	0.22

Disregarding the design of the component structure, the average between- and within-correlations show rather small deviations. However, the “correct” component structure yielded a maximum average correlation within one component which was still smaller than the minimum average correlation between components. The rather small range for the between correlations might be put down to a small “variance” of the correlations within this data set.

The results based on the different designs were compared to results based on the complete data set. Due to the similar between-correlations it is not surprising that nearly no differences could be found between correct component structure results and those based on the random structure. However, the “worst case scenario” yielded significantly worse results than the correct component structure data set. Throughout this section all analyses of imputed data sets are based on a correctly calculated component structure with a 4-choose-2-design. The analyzed imputed data sets were selected using G$^{(w)}$ as introduced in the previous section. To check for effects due to the number of observations the data were combined to three data sets of different size (n = 300, n = 1000 and n = 2000) Table 4.1 provides an overview of the results for these data sets.

Table 4.1: Deviations between complete and imputed data sets, t-statistics (for testing mean differences) and Chi^{2-}tests (for testing distributional differences) ($\alpha = 0.05$) for bivariate distributions between variables of the core component and split variables and for bivariate distributions among split variables.

	Number of observations		
	300	1000	2000
Marginal distributions			
Absolute average deviation (in perc. points)	2.88	1.38	0.57
Proportion dev. (share of sig. t statistics in %)	0.00	0.00	0.00
Bivariate distributions (split vs. core component)			
Absolute average deviation (in perc. points)	6.16	3.36	2.14
Proportion dev. (share of sig. t statistics in %)	4.72	3.45	1.59
Chi2 tests (share of sig. Chi2 statistics in %)	22.81	2.38	2.88
Bivariate distributions (split vs. split)			
Chi2 tests (share of sig. Chi2 statistics in %)	65.50	40.90	20.50

While marginal and bivariate distributions between split and completely observed variables can be reproduced for data sets with a sufficient number of observations, the results indicate that chi^{2-}tests for bivariate distributions between two split variables yield a much higher share of significant deviations than the expected 5 percent (for $\alpha = 0.05$). One possible reason for this specific data set can be found in the very low correlations among the split variables variables but, of course, also in the lower basis of actual observations. Moreover, all regressions were carried out using a linear regression model disregarding the true scale of the dependent variable. An update version of the computer algorithm is planned that includes an extension to the general linear regression model which might also improve the imputation of bivariate distributions between split variables.

The next table provides a comparison between estimates based on the imputed data set and the available case data set with a reduced number of observations. The intention in this case is to control for any misspecification in the underlying linear regression model that might lead to a systematic bias in the imputed data set estimates. Again, deviations to the complete data set are stated for both data sets.

Table 4.2: Deviations, t-statistics and Chi^2-tests ($\alpha = 0.05$) for bivariate distributions between variables of the core component and split variables and for bivariate distributions among split variables for the available case and the imputed data set originally based on 2000 obserations.

	2000 observations	
	AC	Imputed data set
Marginal distributions		
Absolute average deviation (in perc. points)	0.91	0.57
Proportion dev. (share of sig. t statistics in %)	0.00	0.00
Bivariate distributions (split vs. core component)		
Absolute average deviation (in perc. points)	2.16	2.14
Proportion dev. (share of sig. t statistics in %)	0.96	1.59
Chi2 tests (share of sig. Chi2 statistics in %)	1.92	2.88
Bivariate distributions (split vs. split)		
Chi2 tests (share of sig. Chi2 statistics in %)	52.56	20.50

The descriptive statistics yielded satisfying results for both data sets, with slightly smaller average deviations for the imputed data set. However, the results for the t- and Chi^2-tests are only comparable to a limited extent, because they are based on different sample sizes. For the first two cases the tests for the AC data set are based on roughly half the original sample size. In the third case (bivariate distributions between two split variables) the sample size is reduced to one sixth of the original number of observations (see Fig. 1.1).

The data set of the consumer survey we examined included 4027 observations where 327 observations were used to calculate the component structure. These cases were discarded afterwards leaving 3700 observations for the split data set. We summarise the results for this data set in Table 4.3.

Table 4.3: Deviations between complete and imputed data sets, Chi^2-tests ($\alpha =$ 0.05) for marginal and bivariate distributions between variables of the core component and split variables.

	Consumer survey (3700 observations)		
	AC	Best imputation	Selected imputation
Marginal distributions			
Absolute average deviation (in %-points)	0.53	0.59	0.64
Chi2-test (share of significant Chi2-statistics)	5%	16%	23%
Bivariate distributions (split vs. core component)			
Absolute average deviation (in %-points)	2.33	1.76	1.76
Chi2-test (share of significant Chi2-statistics)	8%	9%	10%

Although the average deviations are small for both the marginal and the conditional distributions, the Chi^2-statistics yielded, compared to the media survey, more significantly deviating distributions because of the higher total number of observations. This might also explain the lower share of significantly deviating distributions for the available case data set (approx. 1850 observations), especially for the conditional distributions where the average deviations are higher but the share of significantly deviating distributions is lower than for the imputed data set (3700 observations).

5 Conclusions

Split questionnaire surveys have turned out to show several positive effects: They are especially useful for cost-cutting reasons if the interview time is taking a high share of the total interview costs. This is clearly the case for CATI and other telephone based surveys, where CATI has an additional benefit of easily implementing split designs into the survey. A second benefit of conducting studies with split questionnaire designs is the reduced respondent burden which should lead to less unit non-response and therefore to a better sample quality. The other way round, instead of reducing the respondent burden, split designs also allow to include more questions without increasing it.

The results of the previous section suggest that it is possible to reduce the respondent burden while retaining at least marginal distributions and bivariate distributions of split and core component variables. Even better results are to be expected when proper multiple imputation methods are applied. However, the reduction cannot be extended indefinitely because the remaining information is also getting less. This may explain the worse results we found for the bivariate distributions between the split variables. Besides, while more components do result in further cost-cuttings and an even lower respondent burden, they also increase the complexity of the questionnaire design and reduce the sample size for every single questionnaire. Hence, an appropriate balance has to be found for the trade-off between these effects.

Because the first survey we examined is conducted using CATI with a randomised order of questions, the answer behavior should not be affected by the split questionnaire design. However, it was beyond the scope of this project to conduct additional field tests where split questionnaire results are compared with the corresponding complete questionnaire results.

References

Holland, P.W. and Rubin, D.B. (1982) *Test Equating*. Academic Press, New York.

Koller, F. (2001) *Multiple Ergänzung im Rahmen des Fragebogensplittings*. Master Thesis, Nuremberg, Germany.

Little, R.J.A. (1988) Missing-Data Adjustments in Large Surveys, *Journal of Business & Economic Statistics,* **6**, 3, 287-297.

Little, R.J.A. and Rubin, D.B. (1987) *Statistical Analysis with Missing Data.* John Wiley and Sons, New York.

Mäenpää, C. (2001) *Blockbildung und Methodenevaluation im Rahmen des Fragebogensplittings.* Master Thesis, Nuremberg, Germany.

Munger, G.F. and Lloyd, B.H. (1988) The use of multiple matrix sampling for survey research, *The Journal of Experimental Education,* **56**, 187-191.

Rässler, H. (2001) Split Questionnaire Survey. *Funktionale Spezifikation zur Software SQS 1.0*, raessler automation & consulting.

Raghunathan, T.E. and Grizzle, J.E. (1995) A Split Questionnaire Survey Design, *Journal of the American Statistical Association*, **90**, 54-63.

Rubin, D.B. (1976) Inference and Missing Data, *Biometrika,* **63**, 581-592.

Rubin, D.B. (1986) Statistical Matching Using File Concatenation with Adjusted Weights and Multiple Imputations, *Journal of Business and Economic Statistics*, **4**, 87--95.

Rubin, D.B. (1987) *Multiple Imputation for Nonresponse in Surveys.* John Wiley and Sons, New York.

Schafer, J.L. (1997) *Analysis of Incomplete Multivariate Data.* Chapman and Hall, London.

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

Tennstädt (1987) Are Interviews too Long? *Readership Research: Theory and Practice*, 361-369.