

Web Appendix 1

MODIFIED FEDOROV ALGORITHM

Generating Between-Block Designs

We start describing the procedure that is used to generate the between-block designs. We assume that if there are N individuals and Q questions, then N/K individuals will be assigned randomly to each of the K splits. Each alternative split questionnaire design then consists of an $N \times Q$ 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 (R , a $N_S \times Q$ matrix), which is a list of all potential splits contained in its rows. N_S is equal to 2^Q . Potential restrictions are reflected in the candidate split set, i.e. inadmissible designs such as one where only a single block of questions is administered are removed from R .

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 after eliminating values based on D_0 , i.e. $\mu_{Q \times 1}$ and $\Sigma_{Q \times Q}$. Compute the KL-measure for the starting design $KL(D_0)$ based on these estimates, using (4).

Step 3. Take the first split (first N/K rows) in the starting design D_0 . Exchange that with the candidates, $\ell = 1, \dots, N_S$, i.e. each of the rows in R , in turn. For every exchange, compute the KL-distance in (4), i.e. $KL(D_0^\ell)$. Ensure that the design is fully identified by checking off-diagonals of the $(D_0^{\ell'} D_0^\ell)$ matrix, and reject splits that cause zero off-diagonal values. Keep that split that minimizes the KL-distance, i.e. retain design D_1 such that $KL(D_1) = \min_{\ell} KL(D_0^\ell)$, 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 $\ell = 1, \dots, N_S$ in R , ensuring the design is fully identified, and replacing the design matrix D_1 by D_2 if $KL(D_2) = \min_{\ell} KL(D_1^\ell)$.

Step 5. 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 R , cycling through steps 3 and 4, until no improvement is possible.

Step 6. To avoid local optima, the whole process is restarted with different (random) starting designs and the final design is selected, i.e. the one that yields the lowest KL-distance.

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. The procedure then operates as follows.

Step 1. Build a candidate split set (R_b), $b=1,\dots,B$, for each block. Inadmissible designs are removed from R . 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 R_1 using the modified Fedorov algorithm as described in Steps 3-5 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 R_2 , as described in Steps 3-5 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 3-5 above given the optimal designs of the previous blocks, and assuming the remaining blocks complete, thus obtaining $D_{b,1}|D_{b-1,1},\dots,D_{1,1}$, for $b=1,\dots,B$.

Unfortunately, it proves impossible to produce fully identified within-block designs using the “greedy” approach just 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 complete 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 also be imposed by only considering admissible designs in the candidate split set R_b . One important class of constraints may reflect 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.

REFERENCES

Sudman, Seymour and Norman M. Bradburn (1989), *Asking Questions*. Oxford, Jossey-Bass.

Web Appendix 2

PERFORMANCE OF THE MODIFIED FEDOROV ALGORITHM

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 Fedorov algorithm in finding the optimal design. Let Y_{ij} denote the answer of respondent $i \in \{1, \dots, N\}$ to question $j \in \{1, \dots, Q\}$, 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 20 questions. We generate Y from a multivariate normal distribution with given $\mu_{Q \times 1}$ and $\Sigma_{Q \times Q}$. The matrix R is a $N_S \times B$ matrix containing N_S candidate splits, 1 denoting an included block and 0 denoting an excluded block. There are 32 candidate splits contained in the matrix R , but unrealistic or undesirable combinations such as one where none of the questions is asked (a row with only zeros in the design matrix R) or where just one block of questions is asked, are excluded. Even under the external constraint that fixes the number of desired splits (K), there are many possible designs. For example, there are in total 5,311,735 ($= 26!/(16!10!)$) different designs for $K = 10$ splits with 26 candidate splits. We choose K splits from the candidate split matrix and distribute these splits evenly to one hundred subjects. We do this both with the modified Fedorov algorithm, as well as 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 design with the minimum $KL(D)$ 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 Fedorov algorithm needed to find the optimal questionnaire design with $K=5$, 10 or 15 splits is compared to that for complete enumeration in Table 1. All calculations are done with a Pentium 3 computer, using the GAUSS software. For the Fedorov algorithm, we used 10 iterations, and 1000 different random starts. All 1000 random starts produced the true optimal design in all three cases in $1/10^{\text{th}}$ or less of the computation time of complete enumeration, as shown in Table 1. This indicates that the performance of the Fedorov algorithm as applied to the problem of split questionnaire design is highly satisfactory.

Table 1
PERFORMANCE OF THE MODIFIED-FEDOROV ALGORITHM

K	# of Possible designs (N_S)	Complete Enumeration (sec.)	Modified Fedorov Algorithm (sec.) ¹
5 splits	65,780	260	20
10 splits	5,311,735	10,456	50
15 splits	7,726,160	13,343	78

¹ The modified Fedorov algorithm results are based on 1000 random starts.

Web Appendix 3

GIBBS SAMPLING FOR MULTIPLE IMPUTATION

1. Gibbs Sampling for Continuous Data

We start describing the Gibbs sampler for estimation and imputation when the data from the questionnaires can be assumed to follow a multivariate Normal distribution: $Y_{\text{obs}} \sim N(\mu_{\text{obs}}, \Sigma_{\text{obs,obs}})$. 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 $\Sigma_{\text{obs,obs}}$, $\Sigma_{\text{mis,mis}}$, and $\Sigma_{\text{mis,obs}}$ 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 multivariate normal with mean $\mu_{\text{mis}} + \Sigma_{\text{obs,mis}} \Sigma_{\text{obs,obs}}^{-1} (Y_{\text{obs}} - \mu_{\text{obs}})$ and variance $\Sigma_{\text{mis,mis}} - \Sigma_{\text{obs,mis}} \Sigma_{\text{obs,obs}}^{-1} \Sigma_{\text{mis,obs}}$. The Gibbs sampler iterates between (“rest” denotes the values that are being conditioned on other than the argument of the posterior density):

Step 1. $Y_{\text{mis}} \mid \text{Rest} \sim N\left(\mu_{\text{mis}} + \Sigma_{\text{obs,mis}} \Sigma_{\text{obs,obs}}^{-1} (Y_{\text{obs}} - \mu_{\text{obs}}); \Sigma_{\text{mis,mis}} - \Sigma_{\text{obs,mis}} \Sigma_{\text{obs,obs}}^{-1} \Sigma_{\text{mis,obs}}\right)$,

where Y is $Y = (Y_{\text{obs}}, Y_{\text{mis}})$.

Step 2. $\Sigma \mid \text{Rest} \sim \text{IW}(n_{\text{obs}} + \rho, (n_{\text{obs}} - 1)S + \rho \times \Sigma_{\text{pr}} + S_{\text{m}})$,

where S is the sample covariance matrix, and $S_{\text{m}} = \frac{n_{\text{obs}} \times n_0}{(n_{\text{obs}} + n_0)} (\bar{Y} - \mu_{\text{pr}})(\bar{Y} - \mu_{\text{pr}})'$

Step 3. $\mu \mid \text{Rest} \sim \text{N}\left(\frac{(n_{\text{obs}} \bar{Y} + n_0 \mu_{\text{pr}})}{n_{\text{obs}} + n_0}, \frac{\Sigma}{n_{\text{obs}} + n_0}\right)$

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).

2. Gibbs Sampling for Rank Ordered Data

Since our empirical data consists of only Likert-scale data, we discuss the Gibbs sampling for the rank ordered data. The model is a cut-point extension of the Multivariate Probit model (Albert and Chib 1993). The data is $V_{N \times q}$ matrix of ordered $V = (V_1, \dots, V_k, \dots, V_q)$ variables and we observe ordinal $V_{i,k}$, where $V_{i,k} = j$ if for the latent continuous variables $V_i^* \sim \text{N}(\mu_V, \Sigma_V)$, $\theta_{k,j-1} \leq V_{i,k}^* < \theta_{k,j}$ ($\theta_{k,0} = -\infty$ and $\theta_{k,J} = \infty$). Note that Σ_V should be in correlation form¹ for identification purposes (a diagonal matrix $D^{-1/2} = \text{diag}(\Sigma)^{-1/2}$ will be used to write the covariance matrix as a correlation matrix: $\Sigma_V = D^{1/2} R_V D^{1/2}$, where R_V is a correlation matrix). A uniform prior is assumed for θ_j .

Step 1. Draw cut-points for the ordinal data

¹ Although sampling of a covariance matrix (Σ) from the Wishart distribution is well defined and easy, the sampling of a correlation matrix is more complex. Chib and Greenberg (1998) suggested a Metropolis-Hastings method that generates candidate draws that obey positive-definiteness of the correlation matrix; an alternative is the Griddy-Gibbs sampler, which sequentially generates draws for each element of the correlation matrix subject to the positive-definiteness constraint.

$$f(\theta | V^*) \sim \text{unif}[\max(\theta_{k,j-1}, \max\{V_{i,k}^* : V_{i,k}^* = j\}), \min(\theta_{k,j+1}, \min\{V_{i,k}^* : V_{i,k}^* = j+1\})]$$

That is, the conditional distribution of each cutoff θ_j is uniform on the interval from the highest $V_{i,k}^*$ below the cutoff to the lowest $V_{i,k}^*$ above the cutoff, subject to the order constraints on cutoffs.

Step 2. Draw latent values V^* given the other parameters as follows:

$$\text{Step2a: Draw } f(V_{i,k}^* | \mu_Y, \Sigma_Y, \theta_V) \sim \text{TN}(\mu_V, \Sigma_V); \quad \theta_{k,j-1} < V_{i,k}^* < \theta_{k,j}$$

The truncated normal draw can be obtained through a series of univariate draws using the inverse cumulative distribution function method (Geweke 1991).

Step 3. We draw the means from a Normal distribution and draw the covariance matrix from the Inverse-Wishart and post-process inside the Gibbs chain to obtain a correlation matrix with $R_V = D^{-1/2} \Sigma_V D^{-1/2}$ where $D^{-1/2}$ is a diagonal scale matrix, $D^{-1/2} = \text{diag}(\Sigma)^{-1/2}$ (Edwards and Allenby 2003; Boscardin and Zhang 2004).

$$\text{Step 3b. } \mu_V | \text{Rest} \sim N\left(\frac{(n_{\text{obs}} \bar{V}^* + n_0 \mu_{\text{pr}(V)})}{n_{\text{obs}} + n_0}, \frac{\Sigma_V}{n_{\text{obs}} + n_0}\right)$$

$$\text{Step 3a. } \Sigma_V | \text{Rest} \sim \text{IW}(n_{\text{obs}} + \rho, (n_{\text{obs}} - 1)S + \rho \times \Sigma_{\text{pr}} + S_m)$$

where S is the sample covariance matrix, and

$$S_m = \frac{n_{\text{obs}} \times n_0}{(n_{\text{obs}} + n_0)} (\bar{V}^* - \mu_{\text{pr}(V)}) (\bar{V}^* - \mu_{\text{pr}(V)})'$$

$\mu_{\text{pr}(V)}$ and $\Sigma_{\text{pr}(V)}$ are from prior data.

Step 4. Draw V_{mis} from

$$V_{\text{mis}}^* | \text{Rest} \sim N\left(\mu_{\text{mis}} + \Sigma_{\text{obs,mis}} \Sigma_{\text{obs,obs}}^{-1} (V_{\text{obs}}^* - \mu_{\text{obs}}); \Sigma_{\text{mis,mis}} - \Sigma_{\text{obs,mis}} \Sigma_{\text{obs,obs}}^{-1} \Sigma_{\text{mis,obs}}\right)$$

If $\theta_{k,j-1} \leq V_{i,k,\text{mis}}^* < \theta_j \Rightarrow V_{i,k,\text{mis}} = j$

REFERENCES

- Albert James H. and Siddhartha Chib (1993), "Bayesian Analysis of Binary and Polychotomous Response Data", *Journal of the American Statistical Association*, Vol. 88 (422), 669-679
- Boscardin, John W. and Xiao Zhang (2004), "Modeling the Covariance and Correlation Matrix of Repeated Measures", *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*, John Wiley & Sons, 215-226.
- Chib, Siddhartha and Edward Greenberg (1998), "Analysis of Multivariate Probit Models", *Biometrika*, 85(2), 347-361
- Edwards, Yancy D. and Greg Allenby (2003), "Multivariate Analysis of Multiple Response Data", *Journal of Marketing Research*, 40(3), 321-334
- Gelfand, Alan E. and Adrian F.M. Smith (1990), "Sampling Based Approaches to Calculating Marginal Densities", *Journal of the American Statistical Association*, 85, 398-409.
- Geweke, John (1991), "Efficient Simulation from the Multivariate Normal and Student-t distributions subject to linear constraints" in E.M. Keramidas (ed.), *computing Science and Statistics: Proceedings of the 23rd Symposium Interface*, 571-578. Fairfax, VA: Interface Foundation of North America
- Rubin, Donald B. (1987), *Multiple Imputation for Nonresponse in Surveys*, J. Wiley & Sons, New York
- Schafer, Joe L. (1997), *Analysis of Incomplete Multivariate Data*, London: Chapman & Hall.

Web Appendix 4

**COMPARISON OF SPLIT QUESTIONNAIRE DESIGNS AND
RELATED STUDIES IN THE LITERATURE**

Table 2 below provides a comparison of our approach to related studies in the literature, in particular Kuhfeld et al. (1994), and Raghunathan and Grizzle (1995), on a variety of aspects of the design and modeling procedures.

Table 2
COMPARISON OF OUR APPROACH TO RELATED STUDIES

Criterion	RG95	KTG94	Our study
Designs Split Questionnaires	No	No	Yes
Model used as basis for design	None	Aggregate Multinomial	Location-scale Model
Optimization criterion	None	Logit Fisher information matrix	Kullback- Leibler distance
Optimization space	None	Attribute levels of product attributes in choice sets	Blocks of questions, and questions in blocks
Within and between-block designs	No	No	Yes
Constraints on the design	No	No	Yes
Optimization method	None	Fedorov	Fedorov
Missing data imputation	Yes	No	Yes
Model used as basis for imputation	Location- scale model	None	Mixed Data Model
Estimation method	MCMC	ML	MCMC

Note: RG95: Raghunathan and Grizzle 1995;

KTG94: Kuhfeld, Warren F., Randall D. Tobias, and Mark Garratt 1994

REFERENCES

Kuhfeld, Warren F., Randall D. Tobias, and Mark Garratt (1994), “Efficient Experimental Design with Marketing Research Applications”, *Journal of Marketing Research*, 31 (4), 545-557.

Raghunathan, Trivellore E. and James Grizzle (1995), “A Split Questionnaire Survey Design”, *Journal of the American Statistical Association*, 90 (429), 54-63.

Web Appendix 5

OPTIMAL SPLIT QUESTIONNAIRE DESIGNS

Figure 1

OPTIMAL UNCONSTRAINED BETWEEN-BLOCK DESIGNS

A. 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 Q26-32	Block 6 Q33-40	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									

B. 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 Q26-32	Block 6 Q33-40	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.

Note: 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 people 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 2

THE OPTIMAL CONSTRAINED BETWEEN-BLOCK DESIGNS

A. 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 Q26-32	Block 6 Q33-40	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									

B. 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 Q26-32	Block 6 Q33-40	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.

Note: 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 people 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

THE OPTIMAL WITHIN-BLOCK DESIGNS

A. 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 Q26-32	Bl. 6 Q33-40	Bl. 7 Q41-50	Bl. 8 Q51-59	Bl. 9 Q60-65
10010	00010010	00110	0101000	1100000	00111000	0101000000	001101100	011000
11000	01000010	00011	0100100	0000101	01010000	1000010000	001001000	110000
00110	00001001	10100	0001100	0100001	10010000	0000110111	001010000	101000
00011	00001010	01100	0000110	1000001	00001010	1000000010	011000000	010100
10001	00100010	10001	0001001	0010010	10000010	1000001000	101000010	100001
01100	00011000	01001	0100011	1110000	00010111	0100100011	110001111	100010
00101	01001000	11000	1010101	0011101	01100010	1000000100	001100000	110101
10100	10010111	01010	1110010	0101110	00010010	1111011100	011110001	011110
01001	01100001	00101	0011010	1001000	10000001	1011001001	000011000	000011
01010	11111100	10010	1001000	1000111	11101101	1011101010	100110110	001001

B. 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 Q26-32	Block 6 Q33-40	Block 7 Q41-50	Block 8 Q51-59	Block 9 Q60-65
00110	10000100	00110	1101010	0100100	00111000	0101111110	011001101	011000
00011	01011110	00011	1010010	1011101	11001001	1011000101	110110011	001011
01001	00100101	00101	1000111	0001100	00001010	1000001000	100101000	101000
11110	11111011	11010	0000110	1101011	00001100	1110111011	101110110	001100
10101	00010010	11101	0111101	0110110	11110111	0000011000	000011010	110111