Three-Form Split Questionnaire Design for Panel Surveys

Paul M. Imbriano

IASS Webinar

April 27, 2022

FDA Disclaimer

- The presenter is an employee of the US Food and Drug Administration (FDA) and has no conflict of interest to report.
- The views expressed in this presentation are those of the presenter and should not be construed to represent official FDA views or policies.
- The information, methods, analyses, and examples included in this presentation are provided for academic research and educational purposes only and should not be considered FDA recommended approaches.

Introduction

- A planned missing data design is when a study is purposely constructed to include missing values
- Generally used to either reduce study costs or improve the quality of collected data
- Researchers directly control missing data resulting from the study design
  - As such, planned missing data are either missing completely at random or missing at random
- One commonly implemented planned missing data design is the split questionnaire design

Split Questionnaire Survey Design

- In split questionnaire survey design, a survey is broken into several components and each participant receives a subset of the total components
- Used to reduce survey length, and decrease burden and fatigue on participants
- Which is beneficial because
  - Longer surveys generally have higher item nonresponse rates, with more non-response towards the end of the questionnaire (Raghunathan & Grizzle 1995)
  - Obtained responses are less likely to be valid (Gonzalez 2007)
- Due to the cost of conducting and recruiting participants into studies, survey questions may be pooled from several investigators with different research interests
- Split questionnaire reduces survey length without reducing the number of variables collected in a study

Split Questionnaire Survey Design

- The split questionnaire design was an extension of multiple matrix sampling
- Multiple matrix sampling is a method in which questions are divided into subsets and one or multiple subsets are randomly given to each participant
- Multiple matrix sampling has frequently been used to reduce the testing burden on students for educational assessment, where students are evaluated on several subjects and evaluating a single student on every subject would take a great deal of time (Childs 2003)
- Selecting subsets at random does not guarantee that all item correlations will be estimable
- Raghunathan & Grizzle (1995) placed constraints on item assignment for split questionnaire designs so that all population quantities of interest were estimable

Three-form Design in Cross-Sectional Studies

One common split questionnaire design, the three-form design divides the survey into 4 components

- A shared component X is given to all participants (may be used for demographic information or primary variables of interest)
- The remainder of the survey is divided into three parts (A, B, and C)
- Participants are divided into three groups

Group	Components
1	(X,A,B)
2	(X,A,C)
3	(X,B,C)

- This particular design reduces the survey length by approximately 25% (if variables are allocated equally)
- Modifications can be made to the number of total components and fraction given to each participant

Number of Possible Split Questionnaires

There are usually a large number of possible ways to assign questions to splits. Assuming we have an equal number of variable blocks in each split, we could calculate the number of ways to assign variable blocks using the following equation (assuming the order of splits do not matter):

$$\frac{\binom{b}{q}\binom{b-q}{q}\dots\binom{2q}{q}\binom{q}{q}}{s!} \tag{1}$$

- b represents the total number of variable blocks, where each block consists of one or more variables
- s represents the number of splits or survey components into which the variable blocks are divided
- q denotes the number of blocks per split
- With 18 blocks and 3 splits, there are 2,858,856 potential ways to assign blocks

Assigning Questions to Splits

There are several papers that describe methods to assign variables in cross-sectional studies

- Raghunathan & Grizzle (1995) assigned variables using correlations, where variables with high partial correlations are placed in different components
- Thomas et al. (2006) focused on assigning split questionnaires so observed variables are more predictive of missing values
- Chipperfield & Steel (2009) considered maximizing the efficiency of estimated population totals for a fixed cost or minimizing the cost for a fixed variance
- Adigüzel & Wedel (2008) proposed minimizing the Kullback–Leibler (KL) divergence between the observed and complete data likelihoods for determining an optimal split questionnaire design

Split Questionnaire Designs in Longitudinal Studies

Very few papers discuss administering split questionnaire designs in longitudinal studies, one exception being Jorgensen et al. (2014), which explored longitudinal three-form split questionnaire designs in a latent variable setting

- Limiting survey length may be even more important for longitudinal studies
- Lengthy surveys have the same potential problems with data quality as in cross-sectional studies
- Longitudinal studies also have to worry about dropout
- Longer surveys could result in higher dropout rates
- Zabel (1998) found that a planned reduction in the length of interviews for the Panel Study of Income Dynamics led to a decrease in the attrition rate

Split Questionnaire Designs in Longitudinal Studies

There are more possible ways to assign split questionnaires in longitudinal studies

- Can consider reassigning questions to splits at each wave
- For simplicity we consider scenarios where questions assigned to splits remain unchanged over time
- However, the splits assigned to each participant can vary over time
- We propose six different methods for assigning a three-form split questionnaire design for consideration
    - For simplicity we are omitting the X component

Three-form Design in Longitudinal Studies

Same Forms Design

Design	Group	Wave 1	Wave 2	Wave 3
	1	AB	AB	AB
1	2	AC	AC	AC
(SF)	3	BC	BC	BC

Different Forms Design (Single Rotation)

Design	Group	Wave 1	Wave 2	Wave 3
	1	AB	AC	BC
2	2	AC	BC	AB
(DF1)	3	BC	AB	AC

Three-form Design in Longitudinal Studies

Different Forms Design (Multiple Rotations)

Design	Group	Wave 1	Wave 2	Wave 3
	1	AB	AC	BC
	2	AB	BC	AC
3	3	AC	BC	AB
(DF2)	4	AC	AB	BC
	5	BC	AB	AC
	6	BC	AC	AB

Three-form Design in Longitudinal Studies

Same Form and Different Form Combination

Design	Group	Wave 1	Wave 2	Wave 3
	1	AB	AB	AB
	2	AB	AC	BC
4	3	AC	AC	AC
(CF1)	4	AC	BC	AB
	5	BC	BC	BC
	6	BC	AB	AC

Three-form Design in Longitudinal Studies

Same Form and Different Form (Multiple Rotations) Combination

Design	Group	Wave 1	Wave 2	Wave 3
	1	AB	AB	AB
	2	AB	AC	BC
	3	AB	BC	AC
	4	AC	AC	AC
5	5	AC	BC	AB
(CF2)	6	AC	AB	BC
	7	BC	BC	BC
	8	BC	AB	AC
	9	BC	AC	AB

In addition, we consider a design which assign forms randomly at each wave (Design 6 denoted as R)

Simulation Setup

We performed simulations to evaluate the performance of these designs

- We simulated three variables at three waves to represent the A, B, and C components of the survey
- Variables were drawn from a multivariate normal distribution
- We used multiple imputation and maximum likelihood estimation to analyze the data obtained from these designs (maximum likelihood results were largely the same as multiple imputation and are not included in this presentation)
- Interested in estimating 3 things:
  - Mean
  - Variance-covariance
  - Change in mean over time

Simulation Setup

Table: Structure of the variance-covariance matrix used in simulations.

		Wave 1			Wave 2			Wave 3		
		A	B	C	A	B	C	A	B	C
Wave 1	A	1	ρ_1	ρ_1	ρ_2	ρ_4	ρ_4	ρ_3	ρ_5	ρ_5
	B	ρ_1	1	ρ_1	ρ_4	ρ_2	ρ_4	ρ_5	ρ_3	ρ_5
	C	ρ_1	ρ_1	1	ρ_4	ρ_4	ρ_2	ρ_5	ρ_5	ρ_3
Wave 2	A	ρ_2	ρ_4	ρ_4	1	ρ_1	ρ_1	ρ_2	ρ_4	ρ_4
	B	ρ_4	ρ_2	ρ_4	ρ_1	1	ρ_1	ρ_4	ρ_2	ρ_4
	C	ρ_4	ρ_4	ρ_2	ρ_1	ρ_1	1	ρ_4	ρ_4	ρ_2
Wave 3	A	ρ_3	ρ_5	ρ_5	ρ_2	ρ_4	ρ_4	1	ρ_1	ρ_1
	B	ρ_5	ρ_3	ρ_5	ρ_4	ρ_2	ρ_4	ρ_1	1	ρ_1
	C	ρ_5	ρ_5	ρ_3	ρ_4	ρ_4	ρ_2	ρ_1	ρ_1	1

Simulation Setup

Table: Different correlation values used in simulations.

	ρ_1	ρ_2	ρ_3	ρ_4	ρ_5
Structure 1	0.50	0.00	0.00	0.00	0.00
Structure 2	0.00	0.50	0.50	0.00	0.00
Structure 3	0.00	0.00	0.00	0.50	0.50
Structure 4	0.80	0.50	0.50	0.40	0.40
Structure 5	0.50	0.70	0.70	0.30	0.30
Structure 6	0.50	0.70	0.49	0.25	0.125

- Also included randomly generated correlation structures
- Values were set to missing based on what would have been observed had the split questionnaire design been implemented
- We estimated the quantities of interest using multiple imputation analysis

Simulation Results

Table: Average percent increase in variance from complete data for mean and variance-covariance components using MI

Structure	1	2	3	4	5	6	Random
Design 1: SF	90.7	134.0	109.7	38.0	67.6	79.7	68.3
Design 2: DF1	90.4	92.2	111.2	32.3	46.3	37.7	66.7
Design 3: DF2	86.2	87.9	52.8	27.7	30.1	34.3	53.3
Design 4: CF1	86.9	99.9	35.5	29.7	40.3	45.0	53.7
Design 5: CF2	85.9	94.1	33.4	28.0	32.4	39.3	49.0
Design 6: R	86.6	92.0	18.6	28.7	30.8	37.0	50.0

Simulation Results

Table: Average percent increase in variance from complete data for repeated measures regression change in mean over time for MI

Structure	1	2	3	4	5	6	Random
Design 1: SF	51.5	73.0	89.7	24.1	42.0	44.6	44.3
Design 2: DF1	52.9	102.1	50.5	41.2	73.5	61.7	44.7
Design 3: DF2	52.8	103.0	31.2	36.3	50.1	58.9	37.4
Design 4: CF1	51.9	82.8	28.4	30.2	39.3	45.5	36.5
Design 5: CF2	51.2	90.7	23.0	30.5	36.6	50.2	33.4
Design 6: R	53.9	94.4	12.2	33.7	46.6	57.4	34.4

Example with Survey Data

We took data from the 2002, 2004, 2006, and 2008 waves of the Health and Retirement Study (HRS) and set values to missing to mimic what would have been obtained using the proposed split questionnaire designs

- We took variables measuring blood pressure and diabetes (component A); heart disease, stroke, and weight (component B), cancer wealth and income (component C) from each year
- In addition, we selected basic demographic information (age, gender, race, height, education) and past health behavior information (smoking and drinking history/status) to represent the X component measured in all participants
- In general, variables were most highly correlated with themselves over time
- The correlations between different variables within the same wave were usually very similar to the correlations between different variables measured at different waves

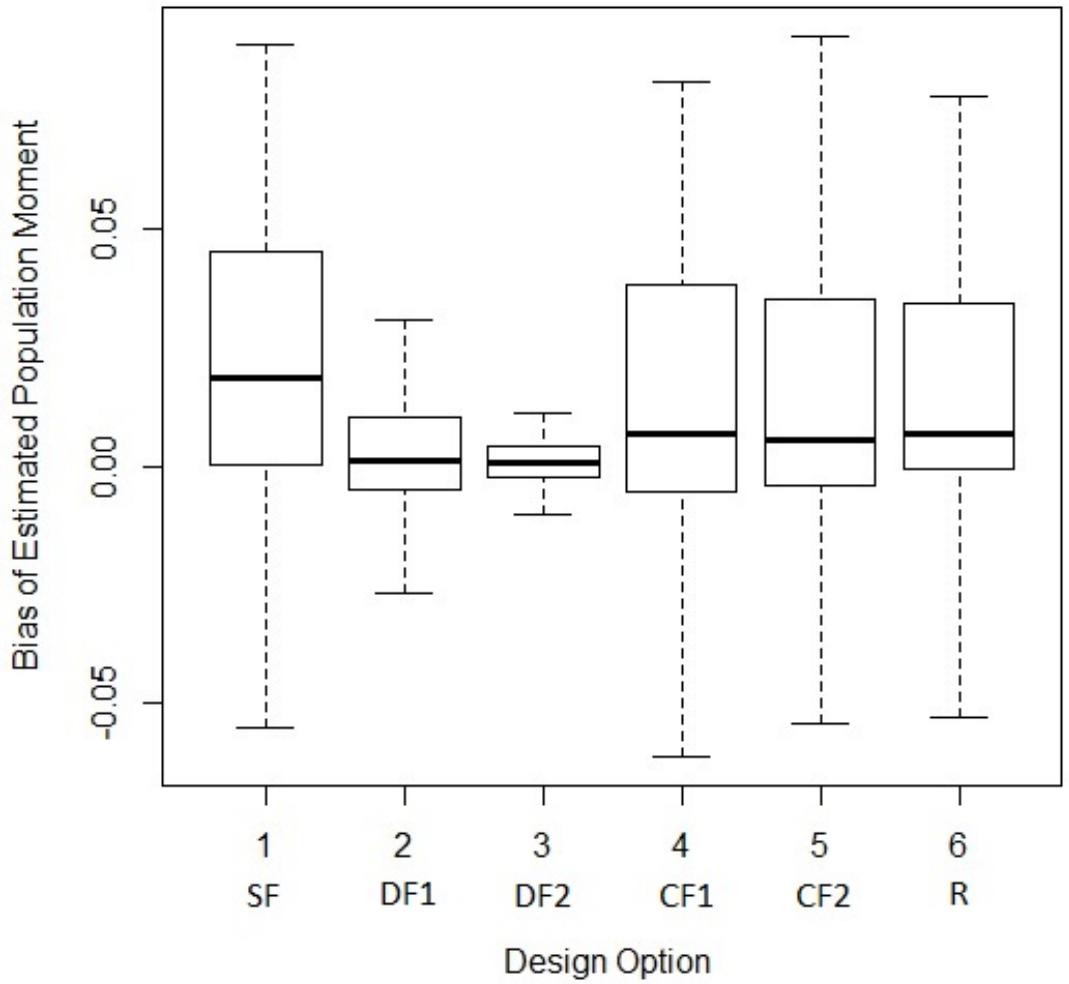
Analysis of Survey Data

- We performed multiple imputation and analysis under each design
- First we analyzed the population moments (mean, median, and quartiles)
- We computed the differences between the estimated population moment from the split questionnaire design and the population moment using complete data divided by the estimated population standard deviation under complete data
- The ratio of the standard errors for the estimate of the mean divided by the complete data standard errors were also computed

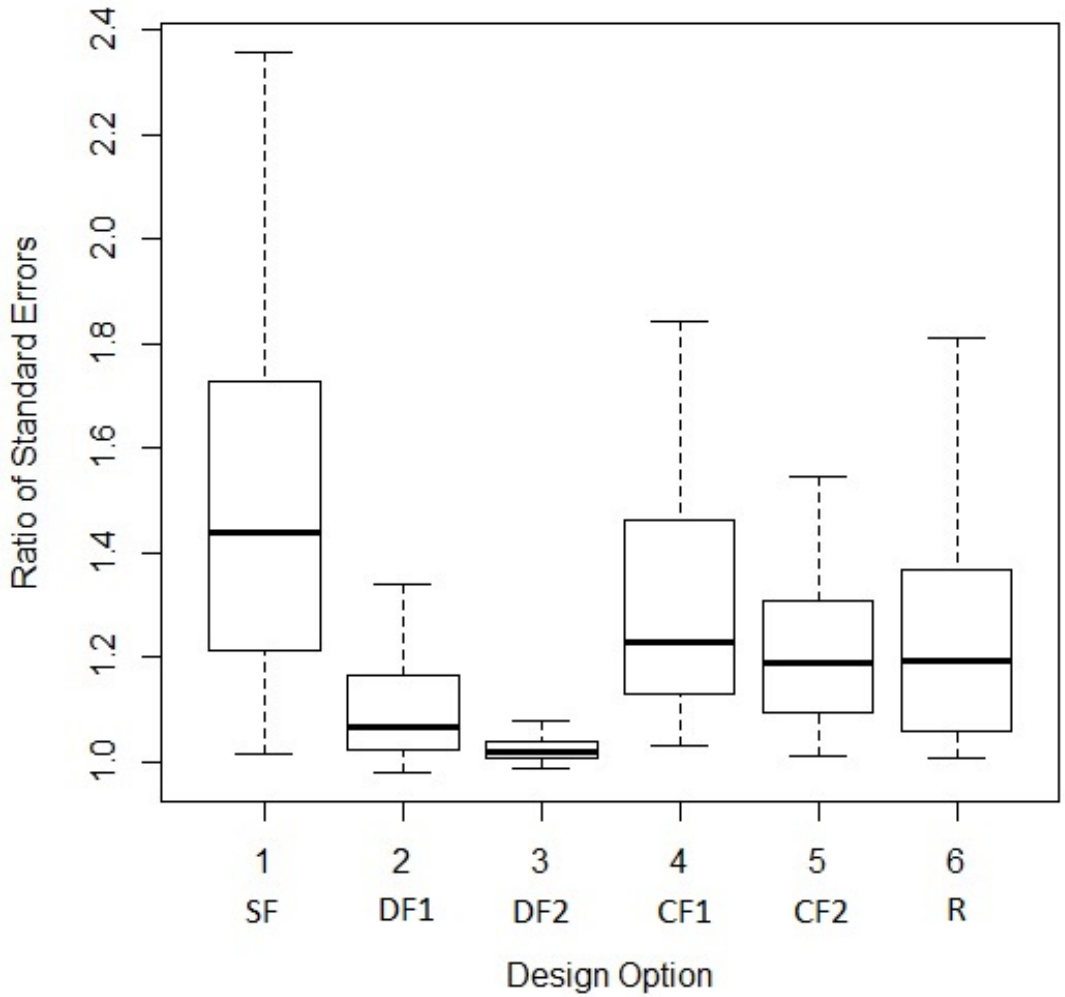
Analysis of Survey Data

- Next we performed regression analyses
- Regressions were based on models used in published papers by Mary Bowen (2010); Best, Hayward, and Hidajat (2005); and Avendano and Glymour (2008) that used the HRS data
- We computed the differences between the regression estimate from the split questionnaire design and the regression estimate using complete data divided by the estimated standard error under complete data
- The ratio of the standard errors for the regression estimates divided by the complete data standard errors were also computed

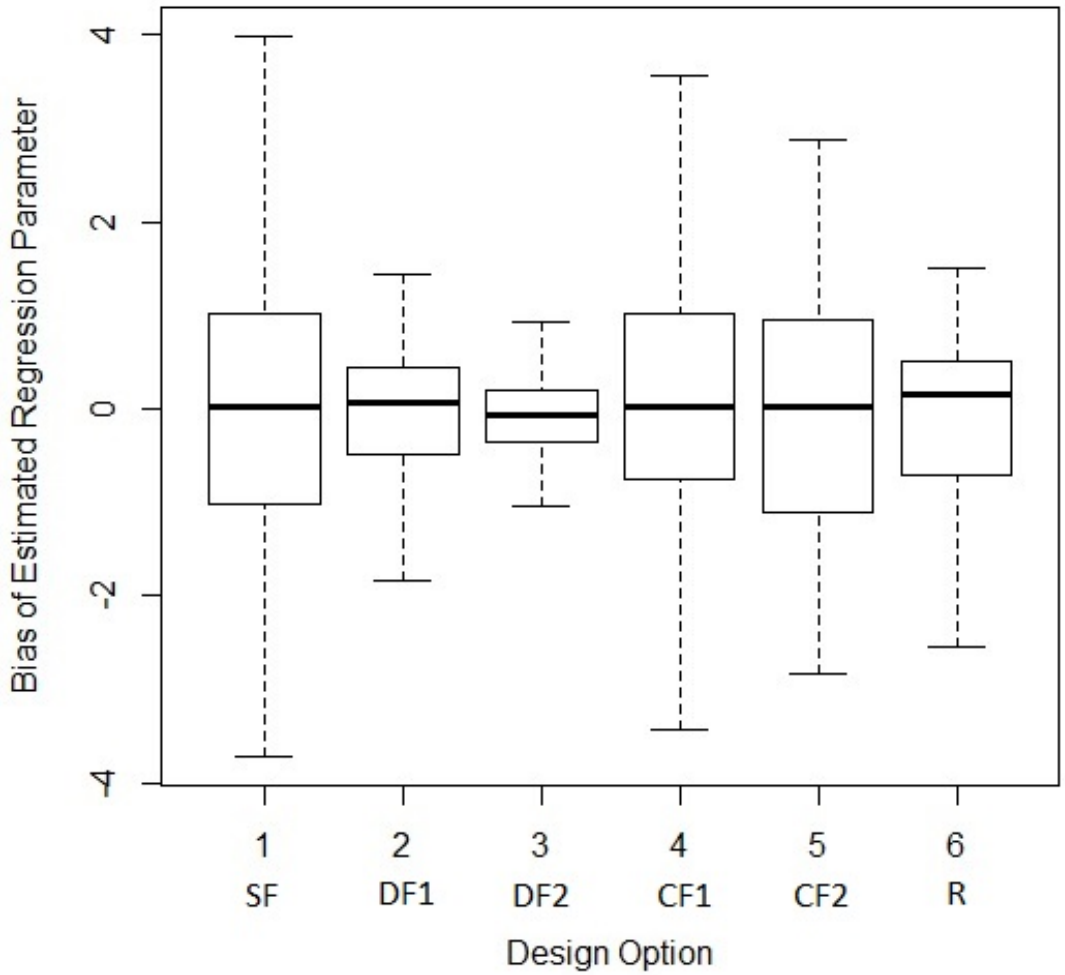
Univariate Parameter Estimates



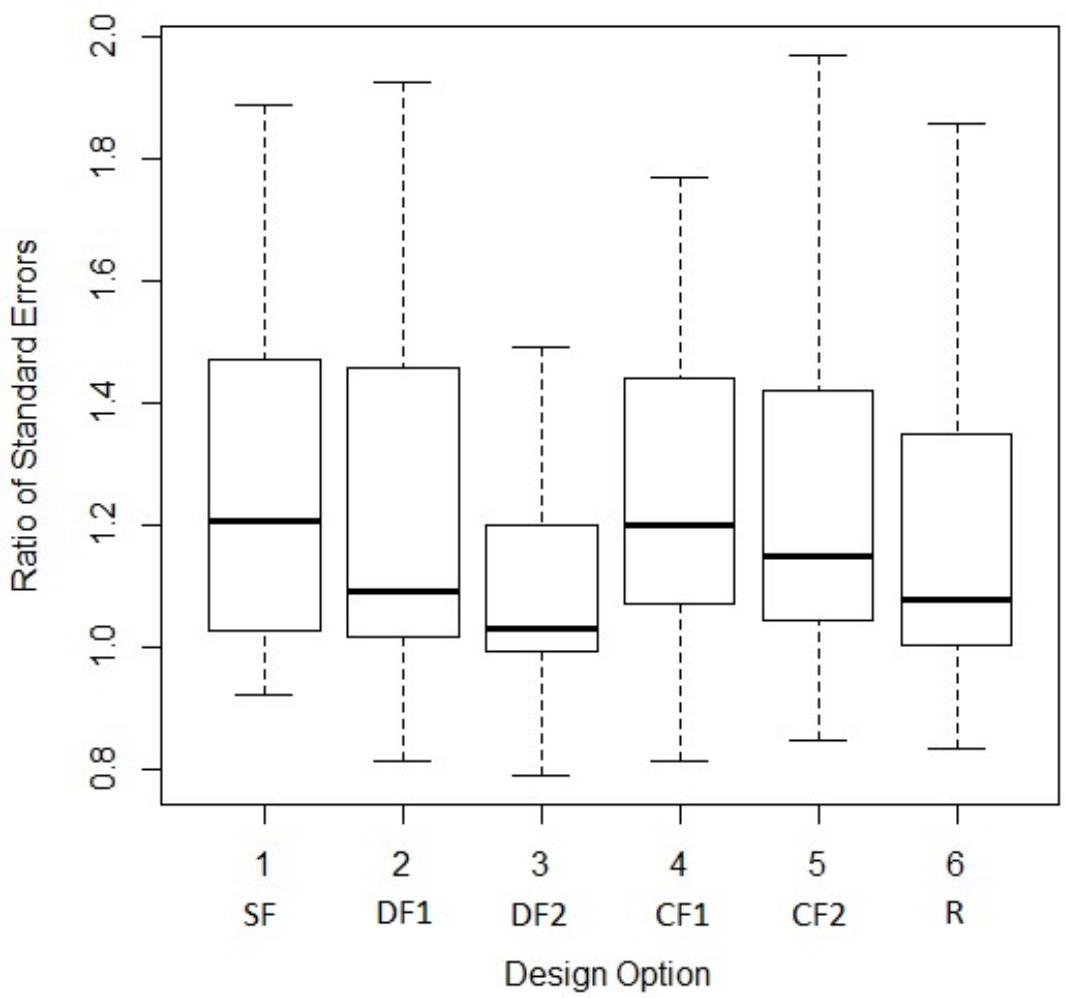
Standard Errors for Univariate Parameter Estimates



Regression Parameter Estimates



Standard Errors of Regression Parameter Estimates



Discussion

The optimal longitudinal design depends on the correlation structure of the data and what quantities we are interested in estimating.

- If investigators are primarily concerned with estimating the mean, regression of one variable on a subset of variables, or cross-sectional properties of the data, we recommend implementing Design 3 (DF2)
- The more complex designs generally outperformed the simpler designs in terms of estimating the means and variance-covariance in our simulations, especially under unusual correlation structures
- If investigators are only interested in the change in a variable over time, Design 1 (SF) will probably be the best choice; however, designs 4 (CF1) and 5 (CF2) should still be considered as they performed fairly similarly to Design 1 in terms of estimating change over time but were generally better with respect to other quantities

Limitations

- No consideration was given to higher-order interaction term, which may not be directly estimable depending on the design
- Although the HRS data contained both binary and continuous variables, we did not examine any joint distributions of variables other than multivariate normal for our simulations
- We did not consider changing the variables assigned to splits at each visit
  - Perhaps reassigning items could have improved efficiency
- We did not consider the effect that complex survey designs might have on our proposed methods
  - We assumed simple random sampling was used, which is rarely the case for large surveys

References

Adigüzel, F. and Wedel, M. (2008)

Split Questionnaire Design for Massive Surveys.

Journal of Marketing Research 45(5), 608-617.

Avendano, M., and Glymour, M. M. (2008)

Stroke disparities in older americans: Is wealth a more powerful indicator of risk than income and education?

Stroke 39, 1533-1540.

Best, L. E., Hayward, M. D., and Hidajat, M. M. (2005)

Life course pathways to adult-onset diabetes.

Social Biology 52, 94-111.

Bowen, M. E. (2010)

Coronary heart disease from a life-course approach: findings from the health and retirement study, 1998-2004.

Journal of Aging and Health 22, 219-241.

References

Childs, R.A. and Jaciw, A.P. (2003)

Matrix Sampling of Items in Large-Scale Assessments.

Practical Assessment, Research and Evaluation 8 (16).

Chipperfield, J.O., and Steel, D.G. (2009)

Design and estimation for split questionnaire surveys.

Journal of Official Statistics 25(2): 227 – 244.

Gonzalez, J., and Eltinge, J. (2007)

Multiple matrix sampling: A review.

Proceedings of the Survey Research Methods Section (December), 3069-3075.

Imbriano, P.M., and Raghunathan, T.E. (2020)

Three-Form split questionnaire design for panel surveys.

Journal of Official Statistics 36 (4): 827-854.

References

Jorgensen, T. D., Rhemtulla, M., Schoemann, A., and et. al (2014)

Optimal assignment methods in three-form planned missing data designs for longitudinal panel studies.

International Journal of Behavioral Development 38, 397-410.

Raghunathan, T. E. and Grizzle, J. E. (1995)

A Split Questionnaire Survey Design.

Journal of the American Statistical Association 90, 54-63.

Thomas, Neal and Raghunathan, Trivellore E. and Schenker, Nathaniel and et. al (2006)

An evaluation of matrix sampling methods using data from the National Health and Nutrition Examination Survey.

Survey Methodology 32 (2), 217-231.

Zabel, J.E. (1998)

An Analysis of Attrition in the Panel Study of Income Dynamics and the Survey of Income and Program Participation with an Application to a Model of LaborMarket Behavior.

The Journal of Human Resources 33(2): 479 – 506.