Survey Research Methods (2017)
Vol. 11, No. 4, pp. 361-368
doi:10.18148/srm/2017.v11i4.7145

ISSN 1864-3361
http://www.surveymethods.org

Reduction of Measurement Error due to Survey Length: Evaluation of the Split Questionnaire Design Approach

Andy Peytchev
RTI International
Research Triangle Park, USA

Emilia Peytcheva
RTI International
Research Triangle Park, USA

Long survey instruments can be taxing to respondents, which may result in greater measurement error. There is little empirical evidence on the relationship between length and measurement error, possibly leading to longer surveys than desirable. At least equally important is the need for methods to reduce survey length while meeting the survey's objectives. This study tests the ability to reduce measurement error related to survey length through split questionnaire design, in which the survey is modularized and respondents are randomly assigned to receive subsets of the survey modules. The omitted questions are then multiply imputed for all respondents. The imputation variance, however, may overwhelm any benefits to survey estimates from the reduction of survey length. We use an experimental design to further evaluate the effect of survey length on measurement error and to examine the degree to which a split questionnaire design can yield estimates with less measurement error. We found strong evidence for greater measurement error when the questions were asked late in the survey. We also found that a split questionnaire design retained lower measurement error without compromising total error from the additional imputation variance. This is the first study with an experimental design used to evaluate split questionnaire design, demonstrating substantial benefits in reduction of measurement error. Future experimental designs are needed to empirically evaluate the approach's ability to reduce nonresponse bias.

Keywords: Split Questionnaire Design; Matrix Sampling; Survey Length; Measurement Error; Multiple Imputation;

1 Introduction

Researchers have long suspected the positive association between survey length and measurement error, creating a tension between the amount of information that can be collected from respondents and the quality of these data (e.g. Weisberg, 2005, p. 129). Measurement error is difficult to quantify, yet there are two types of evidence that supports a link between survey length and measurement error: (1) when faced with a long interview, respondents become increasingly more likely to say "no" to questions that allow them to skip out of additional questions during the course of the interview, as well as across reinterviews (Biemer, 2000; Shields & To, 2005; Silberstein & Jacobs, 1989), and in self-administered surveys (Backor, Golde, & Nie, 2007; Peytchev, Couper, McCabe, & Crawford, 2006), and (2) the predictive ability of the questions decreases from apparently increased random measurement error as the survey length increases (Peytchev, 2007). Less direct evidence is based on indicators that can only be asserted to be related to measurement error, such as finding questions placed towards the end of a self-administered survey to be linked to shorter answers (Galesic & Bosnjak, 2009), faster responding and less variability across questions (Galesic & Bosnjak, 2009; Peytchev, 2007), and extreme straight-lining across items, i.e., selecting the same response option (Herzog & Bachman, 1981). Although beyond the scope of this study, survey length can lead to other consequences, such as higher nonresponse found in longer surveys (for summaries, see Bogen, 1996; Heberlein & Baumgartner, 1978) and the resulting greater potential for nonresponse bias (especially in estimates related to who is unwilling to complete a longer survey).

A survey instrument can be reduced in length, but that may fail to meet the descriptive and analytic objectives. Alternatively, multiple versions of the survey can be administered to different respondents (an idea dating back over half a century Hill, 1951), but that may not meet the demands for multivariate analyses (as noted in Weisberg, 2005, p. 129). To satisfy a survey's key descriptive and analytic objectives, reduction of the survey content may be highly undesirable, leaving a choice between compromising on the objectives of the survey, and achieving all objectives with potentially

Contact information: Andy Peytchev, RTI International, 3040 Cornwallis Rd., Research Triangle Park, NC 27709, (E-mail: andrey@umich.edu)

greater error in the estimates. Since the error resulting from survey length is seldom, if ever, quantified, many studies likely err on the side of collecting more information regardless of the impact on the properties of the survey estimates. Thus, there is the need to quantify error resulting from survey length, and to demonstrate effective methods that mitigate the tradeoff between demands for data and error in survey estimates.

2 Split Questionnaire Design

A potential solution that can be applied to a wide range of surveys is split questionnaire design (Raghunathan & Grizzle, 1995), although the benefits in reduction of measurement error have never been demonstrated – perhaps a foremost barrier to its widespread adoption. In this approach, the instrument is divided into distinct sections and several combinations of these sections are created. Ideally, the combinations are constructed so that each possible pair of questions is observed in at least one of the forms (splits). Respondents are randomly assigned to one of these reduced survey instruments. The missing variables (different across respondents) are then imputed. The imputation is repeated several times and each resulting dataset is used in the estimation process to incorporate the uncertainty in not having fully observed the data. There are two important reasons for (multiple) imputation of the omitted variables: it creates a rectangular analytic data structure allowing for multivariate analysis using standard procedures (only with an additional step of combining the results from the multiply imputed datasets), and it can lead to greater precision (smaller variance estimates) compared to using only observed data (e.g. Peytchev, 2012). Such an approach including questionnaire splits with multiple imputation was proposed by Raghunathan and Grizzle (1995), but components of this design have been used earlier, such as dividing the questionnaire into versions without combining all the data in estimation (Hill, 1951), and matrix sampling in which questions are sampled from a pool of questions and a sample is presented to each individual (Munger & Lloyd, 1988; Shoemaker, 1973). The main appeal of split questionnaire design as a unified framework that combines attention to survey measurement (e.g., question order) with efficient statistical estimation through multiple imputation.

There has been limited research on split questionnaire design since it was proposed. In the absence of experimental designs, focus has been predominantly on optimal methods for creating the splits (Adigüzel & Wedel, 2008; Rässler, Koller, & Mäenpää, 2002; Thomas, Raghunathan, Schenker, Katzoff, & Johnson, 2006; Wacholder, Carroll, Pee, & Gail, 1994). Such studies are based on simulations with generated data (e.g. Bunting, Adamson, & Mulhall, 2002; Graham, Hofer, & MacKinnon, 1996, 2006), and as such, they cannot examine the measurement properties of split questionnaire design (e.g., varying context effects due to different questionnaire structures), nor can they evaluate the potential reduction in measurement error through the use of a shorter questionnaire.

The potential benefits and drawbacks to the use of split questionnaire design have not been well studied. The main objective of the approach is reduction of the instrument length in order to minimize the burden on sample members, and as a result, reduce measurement error and nonresponse. Unfortunately, neither benefit has been evaluated and this is likely a reason that has prevented the widespread adoption of the approach. One could only speculate that the benefits from implementing split questionnaire design will be even greater in particular survey designs, such as those that are among the longest in administration time, which use small or no incentives, and those that have a panel design in which sample members are reinterviewed.

In addition to potentially reducing error in survey estimates, split questionnaire design may cut down costs by reducing the survey length and borrowing statistical strength from the overlap in content across respondents. This result cannot be general and should be survey specific – invariably, it depends on the survey design (e.g., length, mode, incentives), content (e.g., questions and correlational structure), and implementation of split questionnaire design (i.e., creation of the splits, imputation methods, and estimation procedures). Like nonresponse and nonresponse error, cost reduction cannot be evaluated with the data available for this study, but all three – nonresponse, nonresponse error, and cost – are important potential benefits in addition to measurement error reduction.

There are also several limitations of the split questionnaire design approach, mostly related to the dependence on additional statistical models and shifting of burden from the respondents to the survey organization. Even though the latter seems like a benefit, at least initially survey organizations may see it as an increased burden on the entire survey process, from questionnaire development, to data collection, and to postsurvey processing and estimation. With regard to increased dependence on statistical models, imputation models can be optimized for various planned analyses, but data users may come up with analytic models that have not been anticipated and for which the current imputation models may not yield optimal results. There are also errors that could occur in the imputation, particularly when the questionnaire structure is highly complex with convoluted conditional logic – questions that are asked depending on responses to other questions. The area is certainly ripe for additional research, such as improving the statistical models and designing methods to test the sensitivity of survey results to the specification of these models.

The goal of this study is to further investigate the presence of greater measurement error due to survey length and then evaluate the use of split questionnaire design to address this

problem. We address three related research questions:

1. Whether there is greater measurement error when questions are asked late in the survey;
2. Whether a split questionnaire design can provide estimates with lower measurement error than the use of a full questionnaire; and
3. Whether split questionnaire design can yield efficient estimates as to not increase total error (MSE) in this setting.

3 Data and Methods

Data. We use data from a manipulation in which the location of two sets of questions in the survey was randomized.[1] Although the original intent was to minimize the effect of order on other experiments rather than to evaluate split questionnaire design, this manipulation is ideal for the isolation of measurement error resulting from survey length. Most importantly, it provides the ability to evaluate the use of a split questionnaire design to reduce bias in survey estimates due to measurement error resulting from survey length, and any tradeoffs with increased variance from imputation of part of the data.

Data come from an experiment within a set of web survey experiments conducted in November 2004, described by Couper, Conrad, and Tourangeau (2007) and Peytchev, Conrad, Couper, and Tourangeau (2010). Two sampling strategies were used in an attempt to test the sensitivity of the results to a particular set of respondents, and specifically, their prior experience in taking web surveys. Half of the respondents came from SSI's Survey Spot web survey panel, while the other half were recruited through AOL's river sampling in which respondents are recruited by pop-up messages on AOL's web site (and since they are not members of a panel, in general have lower prior experience with web surveys).

The embedded experiment randomized the placement of eight questions, four on diet and another four on exercise, appearing either 2.7 minutes or 10.2 minutes into the survey, on average.[2] A programming error caused the loss of data for one of the diet and one of the exercise questions. The remaining questions asked whether the respondent ate vegetables, fruits, or sweets and whether the respondent walked, exercised indoors, or engaged in other physical activity on a seven-point scale, ranging from "much less than I should" to "much more than I should." In addition, towards the end of the survey respondents were asked to report their height and weight.

Experimental Design. The diet and exercise questions were randomly assigned to appear either earlier or later into the survey instrument. Two other manipulations were administered on these questions, in a full factorial experimental design: framing as being on one topic (health) or multiple topics (diet and exercise), and whether the questions appeared

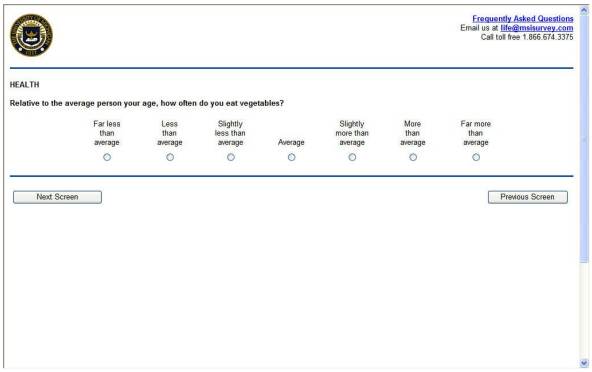

Figure 1. Screen Capture of a Question in the Experiment

all on the same page in a grid (questions in the rows and response options in the columns), all listed on the same page (each question followed by its response options), or each question listed on a separate page. The survey was designed to evaluate any context effects of similar layout experiments being implemented on two separate sets of questions. Since the three layout conditions produced significantly different results, affecting means and correlations (Peytchev, 2007), in this study we include the condition that was found to have the least measurement error based on criterion validity – each question listed on a separate page, shown in Figure 1.

Because the other experiments affected two thirds of the completed surveys and due to additional item nonresponse, we included 732 of the completed 2,587 surveys into the analysis. Of these, 369 received the experimental questions late and 363 saw them early into the survey (Figure 2), which took approximately 18 minutes to complete[3].

In order to ask all the questions in a survey, some questions are invariably going to be asked after many other questions. We label this the "full questionnaire" condition in Figure 2. The diet and exercise questions appear as the fourth set of questions in the instrument, which for ease of schematic representation we describe as having five sets.

Although impossible, ideally, all questions would be

[1] In order to evaluate the full impact of the length of the survey on error in the resulting estimates and any benefit of using split questionnaire design to reduce these errors, both measurement error and nonresponse need to be experimentally manipulated and evaluated. In a full factorial design, this would include the manipulation of the location of the questions in the survey and the survey length. No such data were available and the focus of this study is on measurement error alone.

[2] Twelve respondents who took over one hour to reach the experimental questions were excluded from the estimate of time.

[3] The survey invitation stated that it should take less than 20 minutes.

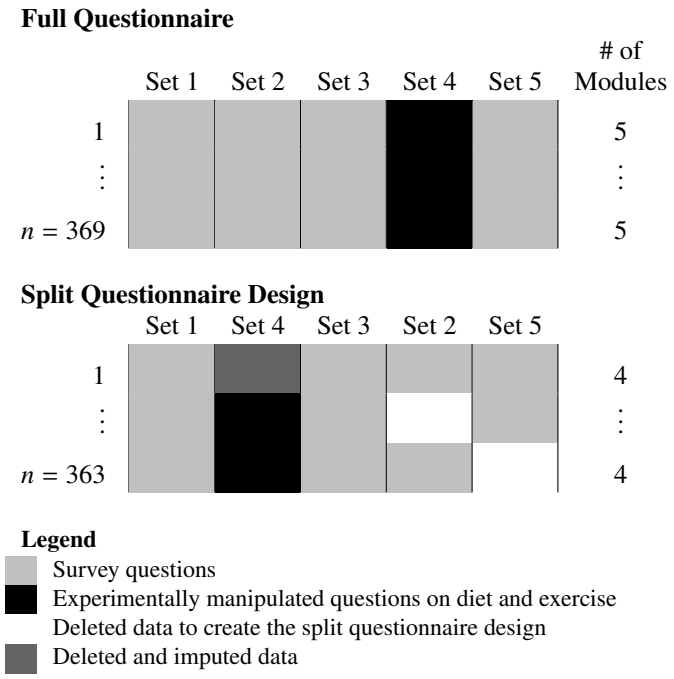

Figure 2. The Full Questionnaire and the Split Questionnaire Design

asked early in the survey to the extent that asking them much later yields different results due to respondent fatigue and increased burden. Asking the diet and exercise questions earlier in the instrument is this ideal condition, which we refer to as the "gold standard" – the "split questionnaire design" in Figure 2, prior to deletion of any of the data.

A feasible way to ask all the questions early in the instrument is through a split questionnaire design by not asking all questions from each respondent. To simulate this design, we use data from when the diet and exercise questions were asked early, and delete the responses for a random third of the respondents – as if three modules were manipulated and only two of them were administered to any given respondent. Deleted data were then imputed 25 times, using variables from question sets 1 and 3. For practical reasons, fewer imputed datasets have been recommended in the past – as few as 5 – but that can lead to larger estimates of between-imputation variances, especially when data for a large proportion of the sample are imputed, as is often the case in split questionnaire design. Sequential regression multiple imputation was implemented in IVEware (Raghunathan, Lepkowski, van Hoewyk, & Solenberger, 2001). In this process, the variable with the least amount of missing data is imputed first, and then the imputed variable can be used as a covariate in the imputation models for the next variable being imputed. In addition, multiple iterations were performed on each imputed dataset, so that the variables imputed towards the end of the process could be used to inform the imputations of the first variables. To preserve associations of analytic interest, key interactions were included in the imputation models. We label this condition "split questionnaire design" in Figure 2.

Evaluation. To address whether there is greater measurement error when questions are asked late in the survey, we evaluate differences in means, expected relationships between variables, and the predictive power of the six diet and exercise questions of interest. A slightly more elaborate set of measures are used to examine if the simulated split questionnaire design leads to lower measurement error compared to the full questionnaire. First, we compare bias[4] and mean square error (MSE). We expect that some bias may be evident in the full questionnaire design to the extent that survey fatigue may lead to a systematic bias. Such bias can be reduced, but at the expense of collecting fewer data from each respondent. Some of this information can be regained through imputation, while reflecting the uncertainty in the imputed values in the variance estimates through multiple imputation (see Rubin, 1978, 1987). The additional source of variance is also included in the MSE estimates, which are computed as the sum of the bias squared and the variance.

Second, we exploited expected association through an artifact of the experiment – two of the diet questions were healthy behaviors, eating vegetables and fruits, while the third was an unhealthy behavior, eating sweets. Thus, a negative correlation can be expected. To the extent that some respondents fail to notice that the question about sweets asks about an unhealthy behavior, similarly to the use of reverse-worded questions to capture inattentiveness in scales, the negative associations will become smaller in magnitude.

Lastly, a more direct use of criterion validity, is the ability of the diet and exercise questions to explain the variability in respondents' body mass index (BMI).[5] Diet and exercise are key inputs and outputs determining a person's BMI and have been shown to be highly predictive of BMI (for review, see Miller, Koceja, & Hamilton, 1997). Greater measurement error in the responses to the diet and exercise questions can attenuate this relationship. The location of the weight and height questions that define BMI was fixed, among the demographic questions at the end of the survey. Indeed, Peytchev (2007) found that the proportion of variation in BMI explained by the diet and exercise questions was lower when the diet and exercise questions were asked later in the survey instrument. A critical question, however, is whether the use of imputation in a split questionnaire design will be able to preserve higher criterion validity, or will attenuate these relationships even more than the full (long) questionnaire. This is evaluated by fitting an ordinary least squares model in

[4] Computed as the difference between the estimate under the full questionnaire or the split questionnaire design, and the gold standard.

[5] BMI is computed as the ratio of weight in kilograms to the square of height in meters.

Table 1
Means and Standard Errors for the Six Diet and Exercise Questions under the Gold Standard Condition, under a Long Questionnaire Condition, and under Split Questionnaire Design

	Gold Standard (Questions Asked Early)		Long Questionnaire (Questions Asked Late)		Split Questionnaire Design Questions Asked Early	
	Mean	Std. Err.	Mean	Std. Err.	Mean	Std. Err.
Eat vegetables	4.63	0.08	3.92	0.07	4.64	0.09
Eat fruit	4.20	0.08	3.46	0.08	4.14	0.10
Eat sweet foods	3.97	0.08	3.93	0.09	3.88	0.09
Walk	4.15	0.09	3.86	0.09	4.04	0.10
Exercise indoors	3.33	0.10	3.04	0.09	3.40	0.11
Engage in other physical activity	4.34	0.08	4.07	0.09	4.32	0.10
n	363		369		363	
n_{observed}	-		-		252	
n_{imputed}	-		-		111	

Measurement on a 7-point scale where 1=much less than I should and 7=much more than I should.

Table 2
Bias and Mean Square Error for the Six Diet and Exercise Questions under the Long Questionnaire and the Split Questionnaire Design

	Long Questionnaire		Split Questionnaire Design	
	Bias	MSE	Bias	MSE
Eat vegetables	-0.71^{*}	0.50	0.01	0.01
Eat fruit	-0.74^{*}	0.55	−0.06	0.01
Eat sweet foods	−0.04	0.01	−0.08	0.02
Walk	-0.29^{*}	0.09	−0.11	0.02
Exercise indoors	-0.29^{*}	0.09	0.07	0.02
Engage in other physical activity	-0.27^{*}	0.08	−0.02	0.01
Average absolute value	0.39	0.22	0.06	0.01

Standard Errors in parentheses. Measurement on a 7-point scale where 1=much less than I should and 7=much more than I should.
* $p < .05$ ** $p < .01$ *** $p < .001$

Table 3
Criterion Validity Measured by (1) Expected Correlations (and their Confidence Intervals), and (2) the Explanatory Ability of the Six Diet and Exercise Questions based on R^2 from an OLS Model Explaining the Variability in the Body Mass Index

	Gold Standard	Long Questionnaire	Split Questionnaire Design	Split Questionnaire (Multinomial Logistic Regression Models)
Correlation of Vegetables with Sweet Foods	-0.29 (-0.38, -0.20)	-0.15 (-0.25, -0.05)	-0.22 (-0.33, -0.10)	-0.25 (-0.64, 0.24)
Correlation of Fruits with Sweet Foods	-0.12 (-0.22, -0.02)	-0.04 (-0.15, 0.06)	-0.06 (-0.18, 0.06)	-0.08 (-0.35, 0.20)
Criterion Validity: Square Multiple Correlation with BMI (R^2)	0.25	0.17	0.18	0.22

which the three diet and three exercise questions are used to predict respondents' BMI. Uncorrelated measurement error in the responses to the diet and exercise questions will attenuate the association with BMI, or if used in a multivariate linear regression model, will decrease the computed R^2.

4 Results

The means and standard errors for each of the six questions are presented in Table 1. We compare three conditions – a gold standard condition, when the questions are asked early in the survey; a long questionnaire condition, when the questions are asked later in the survey; and split questionnaire design, when the questions are asked early for approximately two thirds of the respondents and data are multiply imputed for the remaining third. On average, the means under the split questionnaire design are closer to the gold standard than the means under the long questionnaire, but as expected, at the expense of slightly larger standard errors. The split questionnaire design condition uses data from the same respondents as the gold standard condition – thus reducing sampling variance – but more importantly, the imputation models were able to reproduce the missing data very well. The Fraction of Missing Information (Rubin, 1987), which is defined as the ratio of the between imputation variance to the total variance (with values approaching 0 indicating high certainty in the imputed values and values closer to 1 indicating uninformed imputations) was below 0.02 for all six variables.

These differences are presented in Table 2 as bias estimates with statistical significance tests, along with combined bias and variance (MSE). Five of the six bias estimates are significant in the long questionnaire, while none is significant in the split questionnaire design. The average absolute bias is six and a half times larger, 0.39 compared to 0.06. Despite the slightly larger standard errors in the split questionnaire design shown in Table 1, the average MSE is overwhelmingly larger in the long questionnaire compared to the split questionnaire design: 0.22 and 0.01, respectively.

Next, we turn to associations between variables. Recall that one of the diet variables is, in a sense, reverse worded – eating more vegetables and fruits is generally more healthy, but eating more sweets is less healthy. If respondents fail to pay sufficient attention to the questions, the expected negative association between vegetables and fruits with consumption of sweets is going to be attenuated – reduced in magnitude due to measurement error. Indeed, Table 3 shows that the correlation between eating vegetables and eating sweets is −0.29 in the gold standard (when they appear early), but it is about half as large in the long questionnaire when they appear late, −0.15. Despite the use of imputation models, the split questionnaire design yields a correlation far closer to the gold standard, −0.22. This pattern is the same for the correlation between eating fruits and eating sweets, with −0.12, −0.04, and −0.06, respectively.[6]

The diet and exercise variables, which were all measured on a 7-point scale, were imputed as continuous variables with values bounded between 1 and 7. The analysis was repeated by treating them as multinomial categorical variables. The latter model specification does not rely on the same normality assumptions, but requires larger sample sizes to avoid extreme variation across imputed values. The correlations were computed and presented in the last column in Table 3. Indeed, the correlations were preserved slightly better, −0.25 and −0.08, but at the expense of substantially larger confidence intervals.

A key criterion in the evaluation of the split questionnaire design was examination of the predictive power of the set of questions on diet and exercise of Body Mass Index (BMI). We compared R^2 for the model based on the split questionnaire design and the long questionnaire relative to the gold standard. The split questionnaire design performed only slightly better with R^2 of 0.18 for the split questionnaire design compared to 0.17 for the long questionnaire; the R^2 is 0.25 for the gold standard (Table 3). When multinomial regression is used for the imputation, however, the multivariate correlation is preserved substantially better, with $R^2 = 0.22$. These results are promising for split questionnaire design and underscore the importance of anticipating how the data will be used in order to optimize the imputation for these objectives.

5 Summary and Discussion

Greater survey length was associated with higher measurement error, and split questionnaire design was found to be a viable solution to reduce both survey length and measurement error. We found further support for the notion that longer survey instruments can lead to greater measurement error, by evaluating expected associations between variables. When questions appeared later in the instrument, respondents seemed less likely to differentiate between dietary behaviors that are opposite in nature. More importantly, the criterion validity, measured by the questions' predictive ability of the respondent's BMI, was substantially lower when the questions were asked later in the survey instrument.

We found the results from this first experimental evaluation of split questionnaire design for reduction of measurement error to be particularly encouraging. All four types of metrics employed in the evaluation – estimates of means, total error (MSE), ability to preserve associations among related questions, and maintaining multivariate relationships

[6] As noted earlier, we deliberately restricted analysis to the condition where the questions were presented only one per page; presenting them on the same page was found to induce correlated measurement error (i.e., through straight-lining and other non-differentiation) and led to higher correlations (Peytchev, 2007).

that can be attenuated by measurement error – faired substantially better in the split questionnaire design than in the full questionnaire design. Even with respect to MSE, the full questionnaire design did not outperform the split questionnaire design, despite the relatively small sample sizes and imputation of approximately one third of the data in the latter design.

Split questionnaire design relies on multiple imputation. Imputation models perform best when the analytic objectives are known, so that the models are specified to preserve all important associations. We add that the type of model is also of critical importance and can involve tradeoffs between intended uses of the data. For example, using linear least-squares regression to impute data for the diet and exercise questions was superior in minimizing bias and variance in estimates of means, yet a nonparametric approach of using multinomial regression yielded the highest criterion validity and was best able to preserve bivariate (and multivariate) associations in the data. Researchers should be mindful of both the type and specification of the imputation models in optimizing split questionnaire design to survey objectives.

We could not address questions related to unit nonresponse and survey length, although we were able to completely exclude unit nonresponse through our experimental design and evaluation approach, to focus on measurement error. It would be important to evaluate nonresponse bias in estimates on various topics, including topics expected to be related to the respondents' willingness to complete a longer survey. Since survey length has been found to affect response rates, split questionnaire design may help reduce this source of survey error as well. Such research should manipulate the length of the survey in addition to the random assignment of modules.

There are also many aspects of split questionnaire design that could benefit from further research, such as the tradeoff between measurement invariance and imputation variance – creating splits to retain the same context for all questions (keeping related questions in the same modules and in the same order, even though respondents will not get asked about all major topics in the survey) versus creating splits that aid the imputation models (dividing questions so that each module covers each topic, thus asking each respondent at least some questions on every topic). In this evaluation, we avoided potential measurement differences from drawing samples of questions, and sampled entire modules. There has been research on how to create the splits, but more is needed, particularly from a cognitive perspective.

Replication would also be beneficial, to evaluate the sensitivity and generalizability of the findings. This study used only a self-administered survey mode and nonprobability samples. The sample sizes were also quite small, leading to two unfavorable conditions for split questionnaire design in this evaluation: imputation models perform better for large samples, and MSE estimates for total survey error are less-influenced by variance as opposed to bias as sample size increases. Optimization of a design with respect to survey error is invariably survey- and estimate-specific. Thus, in addition to replication, one area in need of future attention is the development of methods to optimize split questionnaire designs.

Despite some design limitations, this study provides the first experimental evidence that split questionnaire design can yield estimates with less measurement error than using a full (longer) instrument and has potential to become an important addition to the survey practitioner's toolbox.

6 Acknowledgments

The authors thank Roger Tourangeau, Mick Couper, Fred Conrad, and Reg Baker for making data available from Web survey experiments supported in part by National Institutes of Health grant R01 HD041386-01A1. Any opinions, findings and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the NIH.

References

Adigüzel, F. & Wedel, M. (2008). Split questionnaire design for massive surveys. *Journal of Marketing Research*, *45*(5), 608–617.

Backor, K., Golde, S., & Nie, N. (2007). *Estimating survey fatigue in time use study.* Paper presented at the 29th Annual Conference of the International Association of Time Use Research, Washington, DC.

Biemer, P. P. (2000). *An application of markov latent class analysis for evaluating reporting error in consumer expenditure survey screening questions.* Technical Report for the US Bureau of Labor Statistics, RTI International, Research Triangle Park, NC, .

Bogen, K. (1996). The effect of questionnaire length on response rates – a review of the literature. Survey Research Methods Section of the American Statistical Association. Retrieved from https://www.census.gov/srd/papers/pdf/kb9601.pdf

Bunting, B. P., Adamson, G., & Mulhall, P. K. (2002). A Monte Carlo examination of an mtmm model with planned incomplete data structures. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*(3), 369–389.

Couper, M. P., Conrad, F. G., & Tourangeau, R. (2007). Visual context effects in web surveys. *Public Opinion Quarterly*, *71*(4), 623–634.

Galesic, M. & Bosnjak, M. (2009). Effects of questionnaire length on participation and indicators of response quality in a web survey. *Public Opinion Quarterly*, *73*(2), 349–360.

Graham, J. W., Hofer, S. M., & MacKinnon, D. P. (1996). Maximizing the usefulness of data obtained with planned missing value patterns: an application of maximum likelihood procedures. *Multivariate Behavioral Research*, *31*(2), 197–218.

Graham, J. W., Taylor, B. J., Olchowski, A. E., & Cumsille, P. E. (2006). Planned missing data designs in psychological research. *Psychological Methods*, *11*(4), 323–343.

Heberlein, T. A. & Baumgartner, R. (1978). Factors affecting response rates to mailed questionnaires: a quantitative analysis of the published literature. *American Sociological Review*, *43*(4), 447–462.

Herzog, A. R. & Bachman, J. G. (1981). Effects of questionnaire length on response quality. *Public Opinion Quarterly*, *45*(4), 549–559.

Hill, B. (1951). The doctor's day and pay. *Journal of the Royal Statistical Society. Series A (General)*, *114*(1), 1–34.

Miller, W. C., Koceja, D. M., & Hamilton, E. J. (1997). A meta-analysis of the past 25 years of weight loss research using diet, exercise or diet plus exercise intervention. *International Journal of Obesity*, *21*(10), 941–947.

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

Peytchev, A. (2007). *Participation decisions and measurement error in web surveys.* Survey Methodology. Ann Arbor, University of Michigan. Ph.D.

Peytchev, A. (2012). Multiple imputation for unit nonresponse and measurement error. *Public Opinion Quarterly*, *76*(2), 214–237.

Peytchev, A., Conrad, F. G., Couper, M. P., & Tourangeau, R. (2010). Increasing respondents' use of definitions in web surveys. *Journal of Official Statistics*, *26*(4), 633–650.

Peytchev, A., Couper, M. P., McCabe, S. E., & Crawford, S. (2006). Web survey design: paging vs. scrolling. *Public Opinion Quarterly*, *70*(4), 596–607.

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

Raghunathan, T. E., Lepkowski, J. M., van Hoewyk, J., & Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology*, *27*, 85–95.

Rässler, S., Koller, F., & Mäenpää, C. (2002). *A split questionnaire survey design applied to German media and consumer surveys.* Friedrich-Alexander-University Erlangen-Nuremberg, Chair of Statistics and Econometrics Discussion Papers.

Rubin, D. B. (1978). Multiple imputations in sample surveys – a phenomenological baysian approach to nonresponse. Survey Research Methods Section of the American Statistical Association. Retrieved from https://ww2.amstat.org/sections/srms/Proceedings/papers/1978_004.pdf

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

Shields, J. & To, N. (2005). *Learning to say no: conditioned underreporting in an expenditure survey.* Paper presented at the American Association for Public Opinion Research Annual Conference, Miami Beach.

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

Silberstein, A. R. & Jacobs, C. A. (1989). *Symptoms of repeated interview effects in the Consumer Expenditure Survey.* Panel Surveys. D. Kasprzyk, G. Duncan, G. Kalton and M. P. Singh, Wiley.

Thomas, N., Raghunathan, T. E., Schenker, N., Katzoff, M. J., & Johnson, C. L. (2006). An evaluation of matrix sampling methods using data from the National Health and Nutrition Examination Survey. *Survey Methodology*, *32*(2), 217–232.

Wacholder, S., Carroll, R., Pee, D., & Gail, M. (1994). The partial questionnaire design for case-control studies. *Statistics in Medicine*, *13*(5–7), 623–634.

Weisberg, H. F. (2005). *The total survey error approach: a guide to the New Science of Survey Research.* Chicago, IL: University Of Chicago Press.