ABSTRACT

Title of dissertation: THE USE OF RESPONSIVE SPLIT QUESTIONNAIRES IN A PANEL SURVEY

Jeffrey M. Gonzalez, Doctor of Philosophy, 2012

Dissertation directed by: Professor Richard Valliant
The Joint Program in Survey Methodology

Lengthy surveys may be associated with high respondent burden, low data quality, and high unit nonresponse. To address these concerns, survey designers may reduce the length of a survey by eliminating questions from the original questionnaire, but this means that some information would never get collected. An alternative may be to divide a lengthy questionnaire into subsets of survey items and then administer each subset to distinct subsamples of the full sample. This is referred to as a split questionnaire design and has the benefit of collecting all of the original survey information.

We identify a significant deficiency in the current set of split questionnaire methods, namely, the incomplete use of prior information about the sample unit in the design. In most contemporary applications of split questionnaires, generally only characteristics of the survey items (e.g., content, cognitive burden) are used to inform the design; however, if joint consideration is given to characteristics on the survey items as well as the sample unit when designing a split questionnaire, then there may be the potential to improve the split questionnaire's utility. In this

dissertation, we explore the extent to which, if any, jointly considering both types of information at the design stage will yield more efficient split questionnaires.

We propose various methods for incorporating prior information about the sample unit into the split questionnaire using features of responsive design. We highlight how this specific application of a responsive split questionnaire can be used to address the concerns present in a major federal survey. Finally, we draw from the literature pertaining to survey design, experimental design, and epidemiology to develop and implement a framework for evaluating the proposed new elements of our split questionnaire design.

THE USE OF RESPONSIVE SPLIT QUESTIONNAIRES IN A PANEL SURVEY

by

Jeffrey Mark Gonzalez

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2012

Advisory Committee:
Dr. Richard Valliant, Chair/Advisor
Dr. Francis B. Alt
Dr. John L. Eltinge
Dr. Frauke Kreuter
Dr. Nathaniel Schenker

Preface

This dissertation details and evaluates potential methods for designing a specific class of split questionnaire surveys. These methods will contribute to the ongoing research involving split questionnaire designs as well as the redesign of the U.S. Consumer Expenditure Quarterly Interview Survey (CE) at the U.S. Bureau of Labor Statistics (BLS). The primary objective of this dissertation is to develop and propose various extensions of split questionnaire methods and to examine their utility in addressing some of the challenges that the current CE presents.

The current CE instrument is challenging for both the interviewers and respondents and the challenging nature of the instrument can adversely affect the survey products. The primary challenges are as follows. First, the interview is long. Depending on the amount and type of expenditures reported, it takes on average 65 minutes to complete (BLS *Handbook of Methods*, 2007). Lengthy surveys are often viewed as burdensome and subsequently hypothesized to be negatively correlated with response quality.

Second, the questions are detailed. In particular, the respondent is asked to report information (e.g., what was purchased, the amount of the purchase, and when it was purchased) for about 60 to 70 percent of their household's expenses made in the previous three months. These generally include significant purchases, such as those for property, automobiles, and other large durable goods, as well as recurring expenses, such as mortgage/rent and utility bill payments. Even though these are thought to be the types of expenditures that respondents can recount over a three-month period or longer, the nature of the reporting task may still pose problems

for respondents and thus undermine the quality of the data collected. Furthermore, there are external researchers who claim that CE estimates of consumer expenditure shares are biased (Garner et al., 2006). They base their claim on the disparity between expenditure shares calculated from CE data and those calculated from the Bureau of Economic Analysis's Personal Consumption Expenditures (PCE) of the National Income and Product Accounts, a data source that is often cited as being the gold standard for comparison.

Finally, there is also concern over declining unit response rates. Although this trend is not unique to the CE (de Leeuw and de Heer, 2002), the unit response rate for the CE was about 80 percent in the early 2000s but around 74 percent in 2009. The Office of Management and Budget (OMB) has emphasized that every attempt should be made to achieve and maintain an acceptable unit response rate (OMB *Standards and Guidelines for Statistical Surveys*, 2006). Therefore, to potentially reduce respondent burden while improving data quality and the unit response rate, this dissertation will examine extensions of the current set of split questionnaire methods and their feasibility for use as a redesign option for the CE.

Two special features of the current CE (which are unlikely to change during the redesign effort) make it an appropriate data source for investigating extensions of split questionnaire methods. These features are (1) the panel survey design and (2) the measurement goal of the survey – collecting detailed household expenditures and information about the types of consumers who make those purchases. The CE employs a rotating panel survey design in which a particular consumer unit (CU), which can be thought of as being equivalent to a household, is interviewed once

every quarter for five calendar quarters. Although the initial interview is a bounding interview and the data are not used in any official published estimates, similar categories of expenditures, in general, are asked in all five interviews. The implementation of split questionnaire designs that will be explored in this dissertation is for the first interview to remain as is, but the second would be subject to the split questionnaire design.

The second feature is household expenditure patterns. We can leverage our understanding of purchase behavior across time and how some demographic characteristics are predictive of those purchases to collect those information from an initial data gathering effort (e.g., first interview). Once we have the necessary information, we can input it into a model that describes purchase behavior, and use the outputs from the model to design the split questionnaire for the second interview. Thus, we make explicit use of both the characteristics of the survey items and the sample units to design the split questionnaire. We refer to this type of split questionnaire as a *responsive* split questionnaire design because we draw on features of responsive designs (see Groves and Heeringa [2006] for a discussion of responsive designs). Responsive split questionnaire designs will be the focus of this dissertation research.

This dissertation is organized into the following chapters. Chapter 1 briefly introduces the topic of study and provides some motivation for and applications of the dissertation research. Chapter 2 delves deeper into the motivation by providing a literature review of relevant research in survey methodology and other fields (e.g., epidemiology). Chapter 3 presents the results of preliminary data analyses. These highlight relationships in the data that could potentially be used in the development

of decision rules (by a decision rule we mean whether or not to ask a particular question of a sample unit), and demonstrate the utility of extending the current set of split questionnaire methods. We also use these preliminary analyses to identify areas of the research problem that warrant particular focus. In Chapter 4, using historical CE data, we examine a variety of responsive split questionnaire designs based on different techniques for incorporating prior information on sample units to develop decision rules. We also employ mathematical programming methods to ensure that various survey constraints are met when a responsive split questionnaire design is utilized. We then evaluate the performance of our decision rules by exploring hypothetical scenarios of collected data had the methods been implemented. In Chapter 5, we conclude with a discussion of the lessons learned and some of the limitations of this research, offer guidelines for survey programs wishing to utilize the methods developed in this research, and identify areas for future investigation.

Dedication

I dedicate this dissertation to my parents, Joseph and Barbara Gonzalez.

Acknowledgments

This dissertation is a culmination of more than 25 years of schooling and because of this it is important to acknowledge many people for their general support and encouragement and helpful discussions on the topics covered in this dissertation. I am forever indebted to everyone who has helped along the way.

First, I would like to thank Rick Valliant, my dissertation advisor and chair. Rick's guidance throughout the whole PhD process (and while I was a graduate student at Michigan) was tremendously valuable. I would like to thank John Eltinge, my former supervisor at the Bureau of Labor Statistics (BLS) and member of my dissertation committee, as he gave me the general topic of this dissertation and spent countless hours with me discussing issues related to split questionnaire designs.

I would also like to thank the remaining members of my committee – Frank Alt, Frauke Kreuter, and Nat Schenker – for their thought-provoking questions, comments, and suggestions regarding this research. At the time of my prospectus defense, they wanted more specificity in what I was planning to do in my research. I kept this thought in the back of my mind while completing the dissertation and this comment as well as many others greatly improved the final product. I also want to thank Frauke for pushing me to set a date for the defense since once that was set everything seemed to fall into place.

I would also like to thank Steve Heeringa and Jill Montaquila, in addition to Frauke and Nat, for serving on my Comprehensive Exam committee. I would like to thank the faculty (Bob, Partha, Roger, and Fred), staff, and fellow students (espe-

cially Jenna) at the Joint Program in Survey Methodology for positively contributing to the atmosphere of the PhD program. Bob and Partha deserve special recognition for listening to countless mini-proposals on matrix sampling. I am tremendously grateful for Sarah because of her infectious happy disposition, lending an ear when I needed to vent, and keeping the candy jar and snack closet constantly filled. She is one of the people I will miss most from the program.

I would like to say a special thanks to Ashley Bowers. She is by far the most important and significant contributor to my being in the field of survey methodology. She first introduced me to surveys, taught me about all aspects of the survey process, and encouraged me to pursue a career in survey methodology. I finally believe her – it is my "bread and butter."

I would like to thank Steve Cohen for offering me a position in the Office of Survey Methods Research (OSMR) and encouraging me to finish my PhD. I want to thank my other colleagues (MoonJung, Jennifer, Scott, Steve, Bill, Polly, and Clyde) in OSMR for listening to and giving feedback on my numerous presentations on split questionnaires over the past five years. I would like to acknowledge staff in the Consumer Expenditure Survey Program as they allowed me to work with their data and answered questions I had pertaining to the data structures.

I have had many teachers and professors during high school, my undergraduate career, and graduate studies at Michigan who have contributed either indirectly or directly to this dissertation. They are Robert Belton, Tom Braun, Tyler Curtain, Mrs. Foley, Dr. K., Bill Kalsbeek, Jim Lepkowski, Rod Little, Mary Ellen Lively, Mr. Loveland, Noreen Mitchell, Barbara Poole, Raghu, Marcia Rooney, Mrs. Sandhu,

and Doug Schaubel.

There is an additional person that undoubtedly needs acknowledgment, but this person does not like to be recognized. So without mentioning a name, I will say that I could not have done this without you. You have read (and edited) so many versions of this dissertation. You encouraged me throughout the whole process and were there with me through the highs and (very) lows, even when you were an ocean away.

I want to thank all of my friends for putting up with me throughout the years, especially Cereba, Claudia, Jenny, Paula, and Rich. I would also like to thank the DC Front Runners. Running was a great way to de-stress throughout the process and I am thankful for the support and friendships I have gained through my involvement in the club (even though several members constantly reminded me how long I was taking to finish this dissertation).

Finally, I want to thank my family (including my brother, sister, and Aunt Pat). To my parents, Joseph and Barbara, although I questioned your methods at times, I could not have done this without the love and support you have given (and continue to give) me throughout the years.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Many survey organizations are concerned with declining response rates, low data quality, and high respondent burden. Survey methodological research on these issues suggests that each may be related to the length of a survey. More specifically, several studies provide evidence that lengthy survey questionnaires tend to have low response rates (Adams and Darwin, 1982; Bogen, 1996; Love and Turner, 1975; to mention a few). There is research supporting the claim that a lengthy questionnaire can have adverse effects on data quality (Herzog and Bachman, 1981; Johnson et al., 1974; Kraut et al., 1975; among others). Finally, Bradburn (1978) and more recent work by Fricker et al. (2011) identify length of the interview as one of the key dimensions of respondent burden – suggesting that lengthier surveys are more burdensome.

To address the concerns regarding response rates, data quality, and burden, one approach may be to administer a shorter questionnaire to each sample member. If length is one of the underlying causes of these problems, then the hope is that a shorter questionnaire may improve the response rate, improve data quality (at least from the questions asked), and decrease respondent burden. If the original questionnaire is to be shortened by eliminating some questions, a challenge is that the

survey designers, in collaboration with the various stakeholders[1], must decide which questions to eliminate from the questionnaire. This is a daunting task especially if each question is viewed as important by some primary stakeholder.

An alternative to eliminating questions from the original questionnaire is to divide the original questionnaire into subsets of survey items, and then administer each subset to distinct subsamples of the full sample. This is often referred to as a split questionnaire (Raghunathan and Grizzle, 1995) or multiple matrix sampling (Munger and Lloyd, 1988; Shoemaker, 1973a). Chipperfield and Steel (2009) identify three advantages that split questionnaires may have over the typical single-phase design (i.e., the design in which every sample unit is administered every survey item). The advantages are: (1) increased efficiency with which design objectives can be met by allowing the number of survey items administered to each sample unit to vary (i.e., the sample sizes required for each characteristic measured in a survey often differ); (2) improved efficiency in estimation by exploiting the correlation among the survey items collected (i.e., leveraging information can enhance the design and analysis, this is especially true if imputation methods are used to analyze the collected data); and, (3) flexibility to restrict the maximum number of survey items collected from a sample unit to be less than that of the full set of survey items (i.e., common sense dictates that a shorter questionnaire should be less burdensome than a longer one).

Although the above advantages may sufficiently motivate the exploration of

[1]For the purposes of this discussion, a stakeholder is essentially any entity (e.g., person or organization) with a vested interest in the survey program and/or any products subsequently produced from the collected survey data (Gonzalez and Eltinge, 2010).

split questionnaire designs for use in future survey endeavors or redesign efforts, it is necessary to explore the extent to which, if any, additional benefits can be realized by identifying and investigating extensions of the current set of split questionnaire methods. Thus, split questionnaires would have broader applicability and make them a more attractive alternative to standard solutions for addressing the problems associated with lengthy surveys (e.g., declining response rates, low data quality, and high respondent burden). To reiterate, an example of a standard solution would be to form a shorter questionnaire by only eliminating questions from the original survey.

One facet of split questionnaires that warrants further investigation is the assignment of subsets of questions to sample members. In previous applications of split questionnaire designs (Raghunathan and Grizzle, 1995; Thomas et al., 2006; to name a few), the primary focus was on allocating survey items to form subsets of questions (i.e., blocks or splits). Although various methods were used to form the subsets of questions, for instance, item stratification on question content and/or difficulty to ensure that they were homogeneous with respect to the stratification classes, the assignment of the subsets of questions to sample members were generally made with equal probabilities and without regard to prior information on the sample unit. There may be situations, however, in which the survey designer might want to assign the subsets of questions, or even individual questions, to sample members with unequal probabilities. This might occur under a pre-specified estimation plan for which assigning subsets with unequal probabilities yields a smaller theoretical variance of the estimator of the population quantity of interest than an

equal probability random assignment. This is similar to the motivation behind using stratified sampling, specifically, Neyman or optimum allocation, for sampling from some populations as opposed to simple random sampling (Cochran, 1977). Furthermore, these probabilities of subset assignment could be informed by some prior knowledge about the sample unit. This situation is consistent with the framework of multi-phase (Cochran, 1977; Särndal et al., 1992) and responsive survey designs (Groves and Heeringa, 2006) and can be likened to adaptive treatments in medical settings (e.g., clinical trials).

This dissertation explores methods for and the use of a responsive split questionnaire design. In general terms, we refer to a responsive split questionnaire as a split questionnaire that incorporates prior information about the sample unit into the question (or subset of questions) assignment process. It is responsive in the sense that we are tailoring the set of questions to a specific respondent. This particular implementation of a split questionnaire design may have several benefits over current split questionnaire designs. First, because the set of methods we develop would allow for the possibility for sample members to be administered a questionnaire that is tailored to him/her, rather than a standardized one that is administered to a diverse set of sample members. This customization of the survey instrument may help address issues related to surveying highly heterogeneous target populations. Conrad and Schober (2000) concluded that standardized instruments (i.e., surveys for which every sample member is asked the same set of questions in the same way) are sometimes suboptimal because not every sample member understands or interprets questions the same. This may be due, in part, to the fact that sample

members' experiences and situations differ. So, by applying conclusions from their research to our problem, it makes intuitive sense that customization of the survey might elicit more thoughtful responses to the survey questions administered since the topics covered may be more relevant to that individual. Furthermore, because the survey is tailored to the individual it might increase interest in the survey and thereby combat other negative outcomes (e.g., refusal to participate).

Although we are not using a reduction in survey costs as a primary motivating factor for this research (and design decisions always involve quality-cost tradeoffs), the outcomes of this research may provide a framework for considering extensions that would result in substantial cost savings. We recognize that shortening the length of the survey per se would not likely substantially decrease data collection costs since those are generally dominated by the cost of locating and contacting sample units (Sudman, 1967). However, shorter surveys may be more amenable to other design modifications that would yield a significant reduction in survey costs, such as switching the primary mode of data collection from personal visit to telephone.

This dissertation represents a significant contribution to the survey methodological field, particularly in the area of survey design. Motivated by the goal of reducing respondent burden while attempting to improve aspects of the measurement process, we draw on concepts from multiple disciplines – survey methodology, statistics, and epidemiology – and illustrate a novel application of responsive designs to split questionnaire surveys. We use the existing literature on responsive designs and extract features of those designs to propose methods for constructing

split questionnaire surveys in which the questions a respondent receives is dependent on their personal characteristics (e.g., demographics) and our understanding of the behaviors the survey is collecting information on. We demonstrate the use of these methods in a major federal survey collecting information on consumer expenditures. Finally, since several countries around the world are concerned with measuring the expenditure patterns of their respective inhabitants, these methods have extremely broad applicability. Specifically, many countries have in-person household consumer expenditure surveys, some of which have similar features to the U.S. Consumer Expenditure Survey. Among them are Australia, Canada, Denmark, Finland, Norway, and Singapore (To et al., 2011). The methods proposed and developed in this dissertation could be considered viable options for collecting expenditure data in those surveys as well.

This dissertation investigates various issues associated with designing a responsive split questionnaire for a panel survey. Specifically, we investigate various methods for incorporating prior information about the sample unit into the question assignment mechanism and comment on the relative merits of each. For each method, we address the following issues: (1) how to incorporate prior sample unit information in the design of question assignment for the second interview (i.e., modeling information collected from one phase of data collection and using that model to determine which questions to ask sample members at a subsequent phase of data collection); (2) evaluating the impact on estimation efficiency (bias and variance); and, (3) evaluating the methods on their ability to tailor the survey to the individual sample unit. Finally, we demonstrate how to ensure that various survey objectives

are met under the developed methods by imposing constraints on the system (and by system, we mean survey design).

Chapter 2

Literature review

2.1 Overview

This research is motivated from an integration of concepts and methods from both sociological and statistical areas of survey methods research. In this chapter, we review relevant literature from multiple areas of survey methodology and an area of public health (e.g., epidemiology) and identify how each is applicable to our research problem. The literature that we utilize comes from: (1) the social science aspects of survey methodology; (2) split questionnaire designs; (3) responsive/adaptive designs; (4) optimal survey designs; and, (5) epidemiology.

We use survey methods literature to provide motivation for reducing the length of a questionnaire. We provide evidence that lengthy surveys may be associated with high respondent burden, low data quality, and high unit nonresponse. One way to reduce the length of a survey is to employ a split questionnaire. We describe split questionnaire methods in-depth and identify an important deficiency in the current set of methods, namely, the incomplete use of known prior information about the sample unit in the design. If joint consideration is given to characteristics of the survey items and prior information on the sample unit, then there may be the potential to improve the efficiency of split questionnaire designs. A reasonable framework for incorporating prior information into a split questionnaire is a responsive design. We

demonstrate how this framework aligns with our problem by describing key components of a responsive design and then relating those components to the CE survey. We draw on the optimal survey design literature to (1) provide a framework for evaluating the proposed new elements (e.g., responsiveness or success of the procedures in tailoring the survey to the respondent) of the split questionnaire design; and, (2) suggest that the utility of the proposed new procedures may be judged differently by various stakeholders. Finally, we identify possible metrics from sampling design and epidemiology to use in the evaluation of our methods.

2.2 Survey methodological motivation

Justifying this dissertation from a broad survey methodological context is necessary because it would demonstrate that the outcomes of this research are applicable to any future survey endeavors and/or redesign efforts where survey length is a major concern. In this section, we expand on the notion that survey questionnaire length can be adversely related to data quality, response rates, and respondent burden.

Survey methodological literature suggests that the statistical properties of an estimate, e.g., mean square error, can be functions of the quality of the data collected. Data quality can, in turn, be correlated with characteristics of the survey instrument, such as, length. In the case of survey length, the correlation tends to be negative, i.e., longer surveys generally have poorer data quality (Johnson et al., 1974; Kraut et al., 1975). One study concludes that the likelihood of providing accurate

responses tends to decrease when the surveying process extends beyond some optimal length (Herzog and Bachman, 1981). The authors attribute this phenomenon to a decrease in motivation to continue to comply with the survey request. As a consequence, less-motivated respondents take shortcuts and look for easier ways to respond to questions, such as straight-lining[1].

This conclusion is further supported by research from Shields and To (2005). They provide anecdotal evidence from CE Survey interviewers who claim that respondents learn to report not taking any trips or vacations, so they will not get asked a subsequent battery of questions about expenses incurred during those events. It is important to point out that the CE is divided into approximately 22 sections, asked in sequence, and the section inquiring about vacation expenditures is the 18^{th}. Thus, respondents have ample time during the course of the interview to "learn to say no." However, if the respondents, at the beginning of the survey, are only asked relevant questions for which the true response is "yes," then the respondents may be less likely to acquire/learn this negative reporting behavior. Thus, customization of the survey instrument to the respondent's situation might counteract the behavior of learning to say no.

Kreuter et al. (2011) expanded on this line of research by examining the effects of asking filter questions in interleafed versus grouped format. They defined interleafed format as administering follow-up questions immediately after the relevant filter and grouped format as administering follow-up questions after multiple filters. One finding from their research suggests that respondents are more likely

[1]Straight-lining is an increased tendency to give the same responses to similar survey questions.

to answer filter questions affirmatively when they are asked in grouped format as opposed to interleafed. Furthermore, their results demonstrate that the effect of the filter format grows as the number of filters increases and that the filter effect may be influenced by cognitive burden, i.e., the effect size was larger for filters that resulted in more burdensome follow-up questions.

A more recent study by Creech et al. (2011) actually investigated the effect of a split questionnaire design on data quality. They modeled their split questionnaire survey after the current CE Survey instrument and administered it over the telephone in a small-scale field test. They found that indirect measures of data quality, e.g., amount of expenditure dollars reported and "don't know/refusal" responses, moderately improved under the split questionnaire treatment relative to the control.

Survey length can also potentially affect the participation decision of a sample member (Bogen, 1996; Groves et al., 1992). A potential respondent may be less inclined to participate in a survey if he/she lacks an intrinsic interest in the survey topic (Groves et al., 2000). Furthermore, if he/she includes interview length as a factor in the participation decision, then it may be in the survey organization's best interest to administer a shorter questionnaire to sway that potential respondent to participate. A specific example of the association between length and response rate can be found in Roskowski and Bean (1990). They found that the response rate for sample members receiving a shortened version of a questionnaire was about 28-percentage points higher than the response rate for sample members receiving a full questionnaire. In addition, Creech et al. (2011) found that attrition rates were lower with the split questionnaire condition relative to the control condition.

Sharp and Frankel (1983) took this notion one step further and related length to burden perception (which is consistent with Bradburn's [1978] multi-dimensional definition of respondent burden). They found that instrument length was statistically significantly associated with a respondent's perception of burden of the interview and that a greater proportion of long-interview respondents declared that they would not agree to a reinterview a year later. Fricker et al. (2011) analyzed data from the Creech et al. (2011) small-scale field test and came to a similar conclusion. Specifically, they used recursive partitioning to show that a respondent's judgment of the appropriateness of survey length was the most important dimension associated with perception of survey burden.

We acknowledge that some of the literature described in this section is dated. However, previous studies involving split questionnaire designs have been motivated from the same survey methodological perspective as ours (Raghunathan and Grizzle, 1995; Adiguzel and Wedel, 2008). Since the literature review conducted by Bogen (1996) on the effect of questionnaire length on response rate, few additional studies have examined the topic with the exception of the recent study by Creech et al. (2011) relating survey length to attrition. It is often the case that many researchers apply "common sense" and simply assert this fact despite the lack of empirical evidence supporting the claim (Bogen, 1996). Of the few additional studies, the current focus seems to be concerned with the effect of length on participation in web surveys (Galesic and Bosnjak, 2009). In motivating their research, Galesic and Bosnjack (2009) cite similar literature to ours and test the hypothesis that expected length of a web survey is negatively associated with the initial willingness

to participate. Ultimately, the authors found evidence supporting their claim.

In addition to the effect of length on response rates, there is a small body of relatively new literature relating survey length to response quality. As their second hypothesis, Galesic and Bosnjak (2009) examined the quality of responses for questions placed earlier in the web-based survey compared to those placed later in the same survey. They found that responses to survey items placed later in the survey were faster, shorter, and more uniform (i.e., identical answers to different questions) than those placed near the beginning – suggesting that response quality is negatively correlated with length. A study by Krosnick et al. (2002) examined whether there were differences in the amount of "no-opinion" responses based on the location of a question in a telephone survey. They found that there was an increase in the propensity to choose the "no-opinion" option for questions asked later in the survey. Effectively, this is an undesirable outcome if the respondent really does have an opinion about the question being asked; thus, this demonstrates again that response quality can be affected by survey length.

Even if one believes that the research conclusions of previous studies are tenuous or not entirely applicable to the CE surveying environment (due, in part, to the different topics covered in the survey, modes of data collection, etc.), it can still be argued that the effect of length on many aspects of data quality, including unit nonresponse, might be more pronounced in the surveys conducted today. Evidence from the American Time Use Survey (ATUS) suggests that respondents feel busier than they once did even though their data suggest that they actually have more "free time" (Robinson and Godbey, 1997). The authors attribute this to the relentless

multitasking of the respondent so it gives him/her the perception of being busier. Based on this perception, however, respondents may be more reluctant to give up portions of their free time. As a consequence, longer surveys might be viewed as a greater infringement on free time, resulting in greater reluctance to comply with the survey request and/or an increased propensity to respond less thoughtfully and thoroughly.

Finally, data quality concerns are particularly relevant for the CE since there is a vocal group of external researchers who claim that CE estimates of consumer expenditure shares are biased (Garner et al., 2006). This group bases their claim on the incongruence between expenditure shares calculated from CE data and corresponding shares calculated from the Personal Consumption Expenditures (PCE). Comparing survey data from one survey to another might provide a basis for judging the quality of the collected data; however, differences in collection modes, survey definitions, and estimation procedures may contribute to any discrepancies observed between the two sources. Regardless of the true comparability of the two data sources, this group is outspoken enough to result in resources being allocated to an ongoing evaluation of the disparities between the two data sources and members of this critical group are involved in the redesign of the CE.

Regardless of the amount of literature (old or new), it is evident that reduced length questionnaires have the *potential* to improve the measurement (data quality) and nonresponse (response rates) error properties of a survey. Furthermore, previous studies exploring reduced length questionnaires and more explicitly, split questionnaires, have justified their research using these arguments. Therefore, it seems

appropriate to motivate our further study of these methods from that perspective. In the next section, we provide a review of the research on split questionnaire designs and identify some remaining questions that this dissertation will address.

2.3 Split questionnaire methods

Since we have provided justification for being concerned about survey length from a broader survey methodological context, we next focus our review on the use of split questionnaire methods as an approach to reduce survey length, and the issues that warrant study.

2.3.1 Previous research and identification of gaps in the current methods

Split questionnaire methods can be viewed as an extension of multiple matrix sampling, which was first used by researchers at the Educational Testing Service to sample items and to estimate the normative distribution of standardized tests (Shoemaker, 1973a). Matrix sampling designs have a wide range of application in research and evaluation (Askegaard and Umila, 1982). Pugh (1971) demonstrated the superiority of matrix sampling over examinee sampling, which is equivalent to sampling individuals from a defined population, for estimating means and standard deviations of Likert-type attitude items. Sirotnik (1970) also used matrix sampling to show that responses to matrix sampled Likert-type items did not differ significantly from responses when the entire set of items was provided. Multiple matrix

sampling has also been used in estimating scale values obtained by the method of paired comparisons (Askegaard and Umila, 1982). Askegaard and Umila also empirically studied the applicability of these designs to the method of rank ordering – where subjects were asked to rank stimuli from highest to lowest with respect to some attribute. Finally, others have extended these ideas and developed statistical procedures for estimating population moments and other quantities in a sample survey setting (Shoemaker, 1973a). In the sample survey setting, we refer to matrix sampling designs as split questionnaire designs.

Formally defined, split questionnaire methods involve dividing a questionnaire into subsets of survey items, possibly overlapping, and then administering these subsets to subsamples of a full sample. While these designs ensure that every survey item is asked of at least a portion of the sample, they may result in a loss of efficiency. This is a consequence of the reduced sample size receiving each item and the resulting increase in sampling variance.

Compensating for this loss of efficiency became the focus of much of the research on split questionnaires. There are three aspects of split questionnaire methods to study when investigating the reduction in efficiency. They are (1) the allocation of survey items to split questionnaire forms, or blocks of items; (2) the allocation of forms to sample members; and, (3) estimation and inference. The first two deal with design issues, which in very flexible applications can be combined into one process. The third focuses on estimation issues, but these are partly informed by the specific design decisions made. This latter point will be made clearer when we illustrate various split questionnaire designs later in Section 2.3.2 and in Section 2.5 where we

review relevant literature on optimal survey design.

Finding the optimal procedure for allocating survey items to forms involves determining the configuration of survey items on forms such that information loss is minimized when compared to the full questionnaire (Adiguzel and Wedel, 2008). In most applications, a purely random allocation to subsets is least preferred when compared to procedures that make use of item content and other statistical criteria (Shoemaker, 1973b). For instance, some designs employ an item stratification sampling procedure. Under this approach, survey items are stratified by content, difficulty, etc. and then standard stratified sampling techniques (Cochran, 1977) are used to distribute the items among a prespecified number of forms (Shoemaker, 1973b). This ensures that each form is "balanced" with respect to the stratification classes. An empirical evaluation of this allocation method by Shoemaker (1973b) concluded that item stratification was, indeed, preferred over random allocation. One feature to highlight from this application is that allocation was informed by specific characteristics of the survey item.

Other allocation methods require access to prior information on the survey items, possibly coming from a previous administration of the survey or an external data source. One technique is to examine correlations among survey items from the previous administration of the survey and identify those that are most related (Raghunathan and Grizzle, 1995; Thomas et al., 2006). Survey items with high correlations would be allocated to different forms and then the forms would be randomly assigned to individual sample members. The rationale for allocating survey items with high correlations to different forms was based on the proposed inferential

procedures used to analyze the collected data, namely imputation methods. With these methods, survey items could be predicted (imputed) by other items not asked on a particular form (Raghunathan and Grizzle, 1995), and then desired quantities, beyond univariate statistics, (e.g., such as correlations among survey items on two different forms) can be estimated.

It is important to understand how alternative allocations of survey items to forms and then forms to sample members might affect the missing data mechanism. This is because modifications to the post-survey adjustments procedures would be needed to compensate for the missing data mechanism. Alternative allocations also have a direct impact on the type of statistics that can be calculated from the split questionnaire. This is because certain allocations may result in some combinations of survey items never appearing on the same form. So, if methods like imputation were not used in the analysis, then statistics like correlations between items on different forms could not be computed. However, since our survey of application is the CE Survey and one of their primary interests lies in estimating the average spent for a particular item among all consumers[2], we do not focus on the effect of these designs on the estimation of correlations, but acknowledge that it is an important issue and one that needs to be addressed in future research after we demonstrate the feasibility of these designs and the success of the procedures for estimating means.

[2]We refer to this as an unconditional mean expenditure. We use the modifier *unconditional* because the average is computed from all sample units and not just those reporting the purchase. The average computed from only the purchasers is referred to as a *conditional* mean or a domain mean.

2.3.1.1 Gap 1: Prior information on sample unit often ignored

Just as efficiency gains can be obtained from incorporating information on characteristics of survey items in the allocation of items to split questionnaire forms, efficiency gains may also be achieved when information on sample units (e.g., demographic characteristics) are utilized in the allocation of split questionnaire forms. This feature was lacking in the methods of Raghunathan and Grizzle (1995) and Thomas et al. (2006) because they only based their allocation on an aggregate summary of the survey items (e.g., correlations). In other words, their methods ignored known prior information about the individual sample unit. It was also absent from the investigation by Chipperfield and Steel (2009), who took an optimal survey design approach (subject to constraints on fixed costs and variances), in that the only prior information used about the individual sample unit was per unit interviewing costs while other characteristics were disregarded.

A few studies have explored the use of prior information on an individual sample unit to inform the design of a split questionnaire (Gonzalez and Eltinge, 2008; Hinkins, 1984). In particular, Gonzalez and Eltinge (2008) considered five methods for allocating items to sample members. In their study the process of allocating items to forms and then forms to sample members was combined into one process. Their allocation methods are summarized in Table 2.1. In the table, $y_{Int1,ik}$ refers to the i^{th} sample unit's first interview expenditure value on item k and $\bar{y}_{Int1,k}$ is the estimated mean expenditure per sample unit for expenditure k from the first interview. They assumed that the assignment of the question on expenditure k to a sample unit

assignment was based on a random process and denoted the probability associated with this process as p_{ik} for $k = 0, 1 \ldots, K$. These probabilities corresponded to either receiving one of five expenditure categories (clothing, insurance, medical, miscellaneous items, and utilities) or the full set of expenditure categories (when $k = 0$). In each of the five methods investigated, every sample unit had a one-sixth probability of receiving the full questionnaire[3]. As a final note, they used the term subsampling probability to refer to the probability by which the sample unit is administered (only) one of the five survey questions or the full questionnaire.

Allocation method	Subsampling probabilities
Equal	$p_{ik} = 1/6, \forall k$
Squared Deviation	$p_{ik} \propto (y_{Int1,ik} - \bar{y}_{Int1,k})^2$
Squared Relative Deviation	$p_{ik} \propto [(y_{Int1,ik} - \bar{y}_{Int1,k})/\bar{y}_{Int1,k}]^2$
Absolute Deviation	$p_{ik} \propto \lvert y_{Int1,ik} - \bar{y}_{Int1,k} \rvert$
Absolute Relative Deviation	$p_{ik} \propto \lvert y_{Int1,ik} - \bar{y}_{Int1,k} \rvert / \bar{y}_{Int1,k}$

Table 2.1: Summary of subsampling methods (from Gonzalez and Eltinge [2008])

The first method they considered was equal allocation. This means that every sample unit had an equal probability of being assigned to any one of the six subsamples (i.e., $p_{ik} = 1/6$, for $k = 0, 1, \ldots, 5$). This method can be viewed as a baseline procedure because it is the simplest and most similar to a random, uninformative (in the sense that no information on the sample unit or survey item are used in the assignment process) allocation. The second method, squared deviation, required that the subsampling probabilities were proportional to the squared mean deviation of i^{th} sample unit's first interview expenditure value. The third method accounted

[3]This feature is often referred to as a full questionnaire subsample.

for the possibility that some expenditures categories would naturally produce large deviations, so they computed squared relative mean deviations by dividing the i^{th} sample unit's deviation by the mean expenditure value from the first interview. The subsampling probabilities for the fourth and fifth methods were constructed in a similar manner, but instead of using the squared deviation and the squared relative deviation they used the absolute deviation and the absolute relative deviation, respectively. A constraint for all five methods was that the subsampling probabilities for each sample unit sum to one. This constraint ensured that the sample unit would fall into only one of the six subsamples.

One of the implications of subsampling using the methods in Table 2.1 (with the exception of the equal probability method) is that they are attempting to oversample units that would likely have very different expenses from the mean expenditure per sample unit (to the extent that deviant expenditures in the first interview are an indication of deviant expenditures in the second interview). This is similar in spirit to optimum allocation (Cochran, 1977). Under the version of optimum allocation, in a given stratum, the rules dictate that a larger proportion of the sample would come from strata where (1) the stratum is larger; (2) the stratum is more variable internally; and, (3) sampling is cheaper in the stratum. Having an expenditure different from the mean expenditure is similar to the situation of being more variable internally.

Their goal was to study the effect of each allocation method on the properties of $\bar{y}_k$, the mean spent on category k for $k = 1, 2, \ldots, K$; thus, they conducted a series of five simulations, one corresponding to each allocation method. They

ran each simulation $M = 1,000$ times and during each iteration, they randomly assigned sample members to one of the six subsamples based on their subsampling probabilities, p_{ik}. Also during each iteration, they computed the mean expenditure for each of the K expenditure categories, denoted as $\hat{\bar{y}}_{mk}$ for $m = 1, 2, \ldots, M$, using the following design-based estimator of $\bar{y}_k$

$$\hat{\bar{y}}_{mk} = \left(\sum_{i \in S} w_i^* (\alpha_{i0} + \alpha_{ik}) \right)^{-1} \left(\sum_{i \in S} w_i^* (\alpha_{i0} + \alpha_{ik}) y_{ik} \right) \tag{2.1}$$

where $\alpha_{ik} = 1$, if unit i is in the k^{th} subsample; and, $p_{ik} = P(\alpha_{ik} = 1)$. It is worth noting that a sample unit received survey item k for $k = 1, 2, \ldots, 5$ if and only if $\alpha_{i0} = 1$ or $\alpha_{ik} = 1$. The probability that the i^{th} unit was administered the k^{th} item was the sum of p_{i0} (the probability that the unit received the full questionnaire) and p_{ik}. For $k > 0$, the overall probability of receiving the k^{th} survey item was denoted as p_{ik}^* (i.e., $p_{ik}^* = p_{i0} + p_{ik}$). The full sample inclusion probability for the i^{th} sample unit was given as π_i and $w_i = \pi_i^{-1}$ denoted the inverse-probability, or design, weights. Finally, the set of modified design weights were $w_i^* = w_i / p_{ik}^*$.

Using criteria frequently used in optimal experimental design, specifically A- and D-optimality, they identified the allocation method that resulted in the smallest loss of information. Essentially, A- and D-optimality are covariance-minimizing criteria in which the effect of a design on the estimated covariance of key parameters is reduced to a univariate functional. In traditional applications of these metrics, the A-optimality criteria would select the design with the smallest trace, denoted as $tr(\mathbf{V})$, or sum of the diagonal elements of the covariance matrix while the D-

optimality criteria would select the design with smallest determinant of the covariance matrix, denoted as $det(\mathbf{V})$ (Cornell, 1990).

In their study however, A- and D-optimality were used to summarize the simulation covariance matrix and to compare the loss of precision/efficiency, equivalently, the variance inflations across the five allocation methods. So, choosing an allocation method based on the A-optimality criteria was equivalent to selecting the allocation method with the smallest sum of the mean expenditure simulation variances. On the other hand, choosing an allocation method based on the D-optimality criteria was equivalent to choosing the design with the smallest determinant of the simulation covariance matrix. It is also worth noting that A- and D-optimality were appropriate metrics for comparing the losses in efficiency across the simulation conditions because the burden, as measured by the number of questions given to the sample members, was, on average, the same across conditions. For methods that result in different numbers of questions being administered to the sample members, comparisons using the trace and/or determinant are flawed[4].

We present selected results of their key findings in Table 2.2, specifically comparisons of A-optimality and D-optimality for each of the allocation methods. Based on the two optimality criteria, the allocation method that resulted in the smallest loss of efficiency for mean expenditure estimates was the one in which the subsampling probabilities were made proportional to the absolute relative mean deviation. Also in Table 2.2, we present an additional calculation from Gonzalez and Eltinge

[4]The methods we develop in this dissertation will lead to different numbers of questions being administered to sample; therefore, we will not use these criteria in our evaluation.

(2008) in the last column. This quantity is the variance inflation (VIF) and was computed by dividing the sum of the trace for an allocation method and the trace for the estimated full sample balanced repeated replication (BRR) variance by the trace for the full sample BRR variance[5]. This was calculated so that we could assess the relative variance inflation due to the allocation method. It is worth noting that estimated full sample BRR variance is the variance that would have been achieved assuming full response on the expenditure category. From this calculation, we see that their allocation methods resulted in variance inflations from about 125% to 150%. We know that losses in efficiency will occur due to the subsampling, but we hope that our modifications to their methods will either result in smaller losses in efficiency or improve some other aspect of the design (e.g., responsiveness).

Allocation method	$tr(\mathbf{V})$	$det(\mathbf{V})$	VIF
Equal	712.34	34.3×10^9	2.425
Squared Deviation	722.73	47.7×10^9	2.446
Squared Relative Deviation	740.72	52.8×10^9	2.482
Absolute Deviation	660.96	30.0×10^9	2.322
Absolute Relative Deviation	626.72	22.8×10^9	2.254
Estimated Full Sample (BRR)	499.89	3.9×10^9	$\cdots$

Table 2.2: Comparison of allocation methods using optimal design criteria (modified from Gonzalez and Eltinge [2008])

We plan to modify their methods to address their study's limitations. The limitations are as follows. First, they only based the subsample assignment on the expenditure amount from the first interview. This is a very naive perspective of

[5]This is the variance estimation method currently used by the Consumer Expenditure Survey Program (BLS *Handbook of Methods*, 2007).

purchase patterns across successive time periods because it assumes that a large expenditure in one quarter is likely to occur in the next quarter. While this may be true for some expenditure categories, it is not true for all. So, it is possible that a more accurate prediction, or decision, as to which expenditure items should be administered at the second interview, could be obtained by accounting for other characteristics of the sample unit (e.g., demographics) in a model that more thoroughly explains/represents purchase patterns.

Another limitation of their study was that they only considered designs in which one expenditure category was allocated to a sample unit as opposed to multiple categories. In reality, surveys solicit information regarding several survey items from sample members, so it is essential to explore the extensions of these types of split questionnaire methods that solicit information on multiple survey items from respondents. Furthermore, due to the restriction of one item per sample unit, they could not assess burden reduction (when burden is measured by the number of survey items asked) because the average burden across all sample units was the same.

Finally, they did not give an adequate discussion as to whether the methods they developed were "successful" since they only presented and discussed simulation means and covariances. By "successful" we mean how well the surveys were customized or tailored to the sample unit.

2.3.1.2 Gap 2: Ineffective split questionnaire designs for surveys with questions on rare events

An additional reason for why we might want to extend the current set of split questionnaire methods is related to the survey for which we develop and test our methods. In developing an algorithm for creating split questionnaire forms, Thomas et al. (2006) considered the situation of allocating survey items to forms after a "core" set of items had already been chosen, as is often done in practice. This core contained survey items that were highly predictive of several other items, of special interest, and/or their estimates had certain precision requirements. Precision of estimates and, in particular, sample size is often cited as a primary concern when implementing a split questionnaire design (Raghunathan and Grizzle, 1995; Thomas et al., 2006). Since the goal of designing split questionnaires is to minimize the amount of information lost, it seems reasonable that certain items might be designated to a core. An example of a type of item for which precision becomes an issue is a rare event. This is because these events occur with relatively low prevalence in the survey population. We use the term "rare event" very loosely and examples of these could be health events, crimes, or expenditures of a certain type. In standard applications of split questionnaires, questions about rare events are designated to the core. The CE survey solicits some information on relatively low prevalence events, i.e., some items are rarely purchased over a three-month time period.

In Table 2.3 we provide an example of the role of sample size in estimating the

Expenditure[6]	Prevalence (%)	Sample size
Survey methods books	0.5	19,900
Calculators	1	9,900
Cash contributions	2	4,900
Blenders	5	1,900
Bathroom linens	10	900
Men's clothing	20	400

Table 2.3: Sample size required for estimating the prevalence of a characteristic with a 10% CV (assuming SRS)

prevalence of a characteristic with a 10% coefficient of variation (CV), a normalized measure of dispersion, when simple random sampling (SRS) is used. Suppose that the true prevalence is simply the percent of the population purchasing a particular item. Based on this example, it is apparent that special attention should be given to sample size issues when making inferences about rare events from surveys and in particular split questionnaires because the effective sample size for measuring a characteristic will be reduced substantially. Said differently, the increase in sample size needed to acheive the same level of precision becomes quite large as the prevalence of the characteristic being measured decreases.

We continue the discussion of the effect of prevalence on estimates. In Table 2.4, we present the results of a simulation study we conducted examining the effect of prevalence on the CV for the estimated mean of a characteristic. We simulated 10 independent random variables, in a population of $N = 100,000$, each from the distribution $Y \sim N(400, 250^2)$, with the added restriction that $Y \geq 0$ to mimic expenditure data. After we simulated these values and restricted them to the non-

[6]These expenditures are not based on any real data. They are for illustration purposes only.

Prevalence (%)	Sim Mean	Sim SD	Pop Mean	CV
0.05	0.20	0.34	0.21	1.66
0.5	1.95	0.98	1.95	0.50
1	4.28	1.55	4.21	0.37
5	20.56	3.08	20.46	0.15
10	41.18	4.52	41.05	0.11
20	83.04	6.15	83.02	0.07
50	206.42	8.45	206.49	0.04
75	309.36	7.98	309.07	0.03
90	370.48	8.03	370.55	0.02
100	410.72	7.28	411.52	0.02

Table 2.4: Effect of prevalence on coefficient of variation from a population $N = 100,000$ with a sample size of $n = 1,000$

negative half of the real-line, we "zero filled" a random proportion of the y-values according to the prevalences reported in the first column of Table 2.4. What we observe is that, for a simple random sample without replacement of a fixed size of $n = 1,000$, as the prevalence of the characteristic decreases, the CV for the estimated mean increases. For characteristics in which the prevalence of occurrence is 10% or less, the CV is greater than 0.1. A CV greater than 0.1 may be regarded as undesirable for certain stakeholders. For example, an overwhelming majority of CE published estimates of annual expenditure means for various demographic categories (e.g., age of reference person, composition of consumer unit) exhibit CVs less than this value (BLS *CE Current Standard Error Tables*, 2010).

As a final point, for surveys collecting information on rare events, it is common for them to be structured such that the presence of the event is first screened for, and then a sequence of follow-up questions about characteristics of that event is asked.

Thus, if we are designing a split questionnaire for a survey that contains several questions about rare events, then automatically designating them to the core does not entirely solve the allocation problem. This is because if all screener questions about rare events are allocated to the core and conditional on the presence of the event, the follow-up questions are still asked, then ultimately no split questionnaire design has been implemented. Furthermore, asking too many irrelevant questions in a core might have other adverse effects in terms of measurement error (e.g., respondents might straight-line or exhibit other satisficing behaviors). Therefore, it is necessary to develop more efficient methods for split questionnaire designs for surveys that contain many questions about rare events.

2.3.2 Illustrations of split questionnaire designs

We illustrate six representations of various split questionnaire designs in Figure 2.1 and comment on their respective advantages and disadvantages. In each design, we also include some special features that may be useful when attempting to satisfy certain survey objectives. If we let y represent the full vector of survey questions and S represent an initial, full sample, then the rows, denoted by S_i for $i = 1, 2, \ldots, n$ represent subsamples of S and the columns, denoted by y_k for $k = 1, 2, \ldots, K$ represent specialized subsets of questions of the full questionnaire. Furthermore, we have the following two relationships: (1) $y = \bigcup_{k=1}^{K} y_k$; and (2) $S = \bigcup_{i=1}^{n} S_i$. As a final note, in each of the designs presented the shaded squares correspond to data that are collected, while the open squares correspond to data

that are not collected, or missing by design.

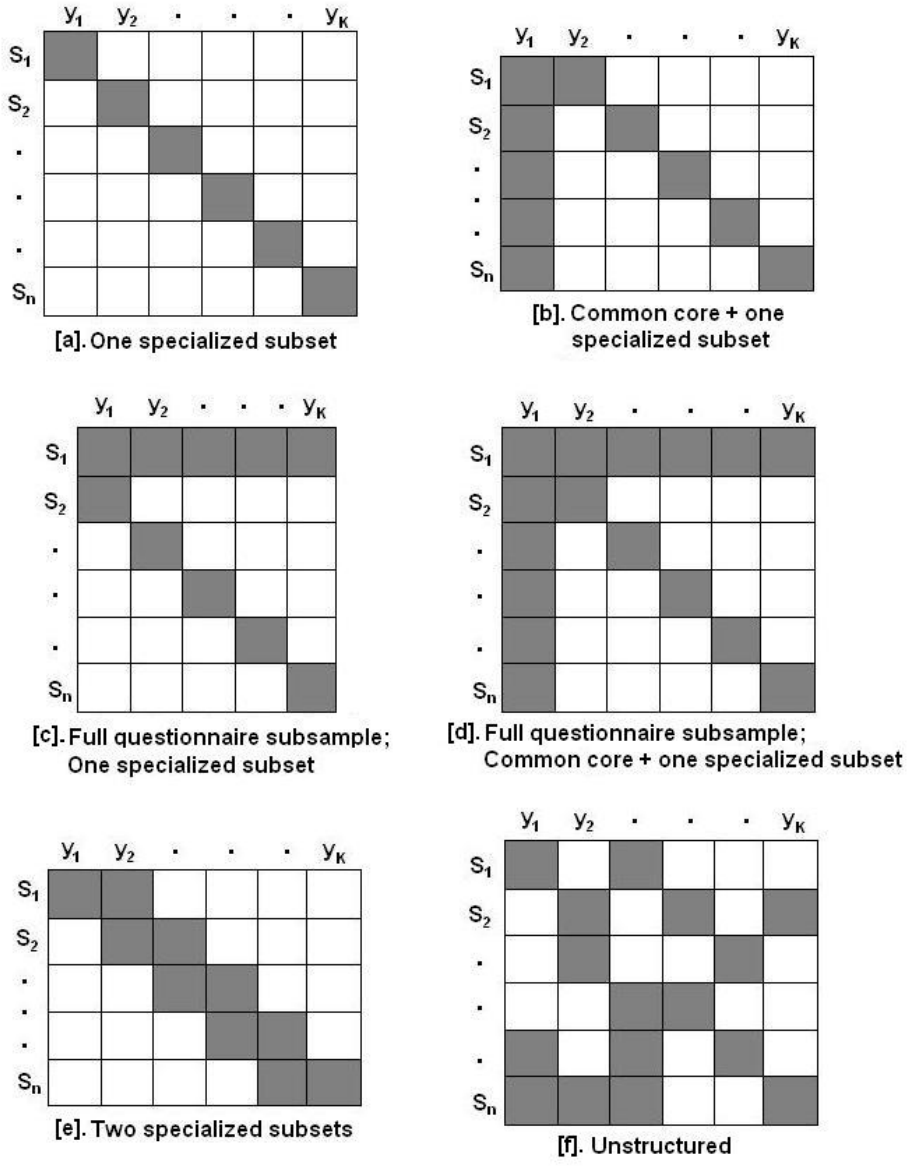

Figure 2.1: Various representations of split questionnaire designs

The split questionnaire design depicted in Figure 2.1[a] is adequate for estimating univariate statistics and other parameters from univariate distributions. A disadvantage of this design is that depending on the full, initial sample size, the sample size used to estimate desired paramters about each y_k from the split ques-

tionnaire may be deemed too small for some purposes. For instance, assuming a simple random sample of size n from a population of N units, we know that the theoretical sampling variance of the estimated mean, $\hat{\bar{y}}$, is

$$V(\hat{\bar{y}}) = (1 - f)S^2/n \tag{2.2}$$

where $f = n/N$ is the sampling fraction; and, S^2 is the population element variance (Cochran, 1977). Thus, if we let n^* be a reduced sample size, for instance a 50% reduction in the size of the sample on which we measure y, then we have the following (ignoring finite population corrections):

$$\frac{V(\hat{\bar{y}}_{new})}{V(\hat{\bar{y}}_{old})} \doteq \frac{S^2/n^*}{S^2/n} = \frac{S^2/0.5n}{S^2/n} = 2 \tag{2.3}$$

So, our sampling variance would increase by a factor of 2, or equivalently by 100%, if the split questionnaire design was used. We can apply these types of relationships to all split questionnaire designs, i.e., the reduction in sample size for each question will result in an increase in the sampling variance.

In Figure 2.1[b] we illustrate the design in which we have a "core" set of questions, denoted by y_1, that every sample member receives, regardless of subsample membership. In addition each subsample receives one specialized subset of questions. Previously, we mentioned that a "core" might contain high priority survey items and/or items in which specific precision requirements are to be met. Another advantage of this design is that correlations between any y_k for $k = 2, 3, \ldots, K$

and any of the elements of y_1 can be computed. In terms of deficiencies of this design, for y_k for $k = 2, 3, \ldots, K$, the same sample size issues that are pertinent to Figure 2.1[a] are relevant for this design as well. In reality, most, if not all split questionnaire designs will involve some type of core as core sets of questions tend to include demographic characteristics as well.

In Figure 2.1[c] we have designated a subsample to receive the full questionnaire, S_1, and the remaining subsamples to only receive one specialized subset of survey items. Using the data collected from the full questionnaire subsample, S_1, any analytic objective from the original survey can be met such as providing a full microdata file to data users, without the data processor having to utilize imputation methods to fill in the missing data. The design in Figure 2.1[d] combines the two special features of 2.1[b] and 2.1[c], the common core and the full questionnaire subsample, so the benefits of each of these designs would be realized if 2.1[d] was implemented.

The design depicted in Figure 2.1[e] represents the split questionnaire design in which adjacent pairs of survey items are administered to a subsample. Extensions of this design can easily be realized so that all $\binom{K}{2}$ pairs of survey items are administered to a subsample. This design may be viewed as an improvement over Figure 2.1[a] because the sample size receiving each subset of survey items increases by a factor of 2 (if the same number of sample members is included in each subsample). Additionally, this design is useful for computing estimates of parameters from bivariate distributions and regression analyses in which no higher than first order interactions are included in the model.

Finally, in the last design Figure 2.1[f] we depict an unstructured split questionnaire design. For this design the number of survey items administered to each subsample can vary dramatically. One of the advantages of Figure 2.1[f] is that each subsample can be administered a tailored survey in the sense that we administer only the elements of y that are relevant to S_i. This type of design, however, requires a commitment from the field interviewers to actively administer the survey since the instrument may be quite different for each sample member. Another potential deficiency of this design is that some survey objectives might not be met. For instance, without imposing certain constraints on the design, a pair of characteristics might never be administered to the same sample members; thus, certain quantities, such as correlations between pairs of survey items, cannot be estimated from the collected data using standard techniques (i.e., without imputation). Despite its potential limitations, the type of design depicted in Figure 2.1[f] will be the focus of this dissertation. We chose to focus on this type of design because it is the design that addresses the goal of tailoring the survey to the respondent. By focusing our efforts on this design, we can also explore how well certain design objectives are met (e.g., design unbiasedness of key estimates and potential variance inflations) and discuss the tradeoffs associated with attempting to tailor the survey.

As is evident from these illustrations, the choice of the split questionnaire design should depend on some combination of the objectives of the survey and what errors or issues associated with the data collection we are trying to address. Furthermore, the preceding discussion was given under the assumption that the desired quantities would only be estimated from the collected data. However, if

other techniques, like imputation (Little and Rubin, 2002), are used to recover the information not collected from sample members, then many, if not all, analytic objectives can be met with any split questionnaire design.

To recapitulate, in Section 2.3, we reviewed some existing methods for allocating survey items to split questionnaire forms and then allocating the forms to sample members. We also suggested that these two processes can be combined into one. We highlighted that it is important to consider characteristics of both the survey items and sample units in the design of split questionnaires as this joint consideration may alleviate some of the issues associated with losses in efficiency in current split questionnaire designs or improve some other aspect of the design. We also discussed why the existing methods are inadequate for surveys with many questions about rare or low prevalence events and as a direct consequence of our data source, we will study this issue in-depth. Furthermore, we conveyed that design issues cannot be considered without thinking about the desired inferences from the collected data, or the survey objectives. Finally, we provided illustrations of various split questionnaire designs and discussed their relative merits.

2.4 Responsive survey design procedures

As previously mentioned, we aim to develop split questionnaire methods in which we assign survey items to sample units based on prior information about characteristics of the unit and by leveraging our understanding of the behavior our survey is trying to solicit information on. To accomplish this we use features from

a responsive design (Groves and Heeringa, 2006). Groves and Heeringa use the term responsive survey design to describe making mid-course decisions and unit-level survey design changes based on accumulating process and survey data. These decisions are meant to improve the error (and cost) properties of the resulting survey statistics. Responsive designs can be tied closely to multi-phase designs.

2.4.1 Components of a responsive survey design

The formal definition of a responsive survey design has five components: (1) identify survey design features that potentially affect the cost and error structures of the survey statistics; (2) identify indicators of the cost and error structures of those features; (3) monitor the indicators during some initial phase of data collection; (4) based on a decision rule, actively change the survey design features for the unit in the subsequent phases; and, (5) combine the data from the distinct phases to produce a single estimator of the desired quantity of interest.

In Figure 2.2 we provide a diagram of a three-phase responsive design. We adapted this figure from Groves and Heeringa (2006). In this figure, examples of Phase 1 design options may include the dollar amount of an incentive, nonresponse follow-up procedures, and different versions of the questionnaire. It is worth noting that Phase 1 does not have to include multiple design options. There may be only one design option that the entire sample receives, i.e., the same set of survey procedures are administered to every sample unit. If multiple design options are used, then each design option is administered to a replicate, or microcosm, of the

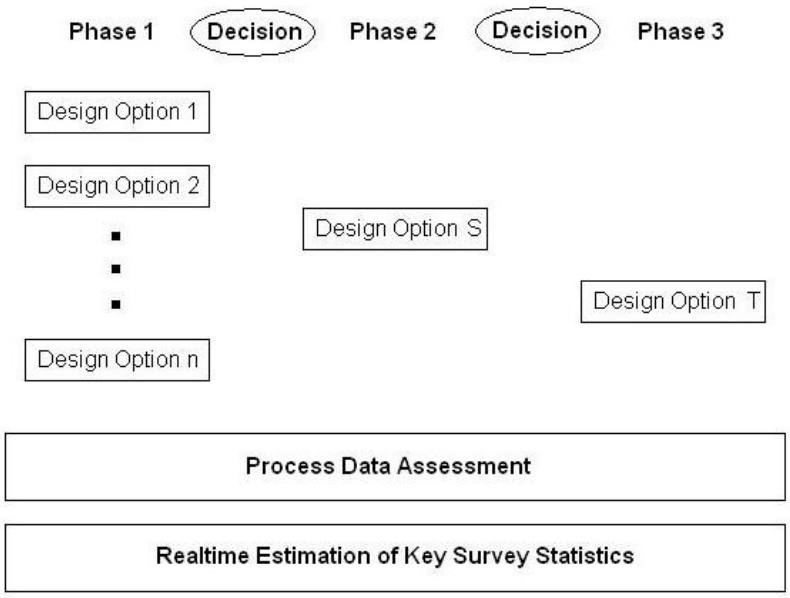

Figure 2.2: Illustration of a three-phase responsive design (from Groves and Heeringa [2006])

initial sample. During Phase 1, the survey organization begins to collect paradata, or process data, and other information that will inform the design decisions for subsequent phases. Certain quantities such as, key survey statistics, are computed on a regular basis, perhaps nightly, to monitor how they change as more data are collected. Based on the paradata and the estimated quantities from the collected data, the survey organization makes a decision as to which design option a particular sample unit will receive in Phase 2. At the end of Phase 2, the process data and survey statistics are reassessed and another decision is made about the design options that a sample unit will receive in the Phase 3. Not depicted in this figure but of importance is combining the data collected during each phase to produce the primary estimates of interest for the survey.

2.4.2 Responsive designs applied to split questionnaires

Given the general responsive design framework, it is important to demonstrate how components of a responsive design are applicable to this dissertation research. To do this we relate features of our research problem to each component of a responsive design.

2.4.2.1 Relating components of a responsive design to a responsive split questionnaire

The first component requires that we identify survey design features affecting the error structures of the survey statistics. In previous sections, we have provided evidence that data quality, specifically measurement and nonresponse error, can be (negatively) affected by survey questionnaire length. So, our primary design feature is questionnaire length, defined explicitly by the number of survey items a respondent is administered. Next, we specify a set of indicators of the error properties of those features. Our primary data source is the CE (extensive details of the CE are provided in Section 3.2.1 and Appendix A). For the research presented here, we focus on split questionnaire methods for the second interview. Thus, our set of indicators are collected during the first interview. The third component is likened to the actual data collection effort of the first interview (or during the process of locating and contacting sample units to solicit participation).

Based on the information collected in the initial interview, we want to make a decision about altering the specific survey feature, length. We first develop a decision

rule to determine which questions to ask each sample member during the second interview. This decision rule will be based on our understanding of the underlying behavior our survey is collecting data about and will incorporate prior information about the sample unit. Even in standard applications of responsive designs, paradata and the models relating the prior information to the design modifications in the subsequent phases need conceptual development and diagnostic assessment to ensure that the chosen modification is successful. Thus, we consider two perspectives on how to translate the prior information into split questionnaire design decisions. We identify these perspectives in Section 2.4.3 and discuss how decision rules will be developed under each perspective. Finally, according to the last component of a responsive design, we want to combine the data from the separate design phases into one single estimator.

In Figure 2.3 we illustrate how components of a responsive design align with our proposed split questionnaire methods. In Phase 1 of our problem, we have the first interview of the panel survey. Every sample unit is administered the same survey questionnaire so there is only one design option. During this time, we begin to assess our process data and other information collected during the initial interview and use it to develop our decision rules regarding whether to ask a sample member about the purchase of a particular item at Phase 2. Phase 2 in this research refers to the second interview of the panel. The data collection effort for the second interview will occur, with Phase 1 data (e.g., information on the sample unit) having been incorporated into the subsampling probabilities, and the subsampling probabilities will be combined with the second interview data to produce the desired survey

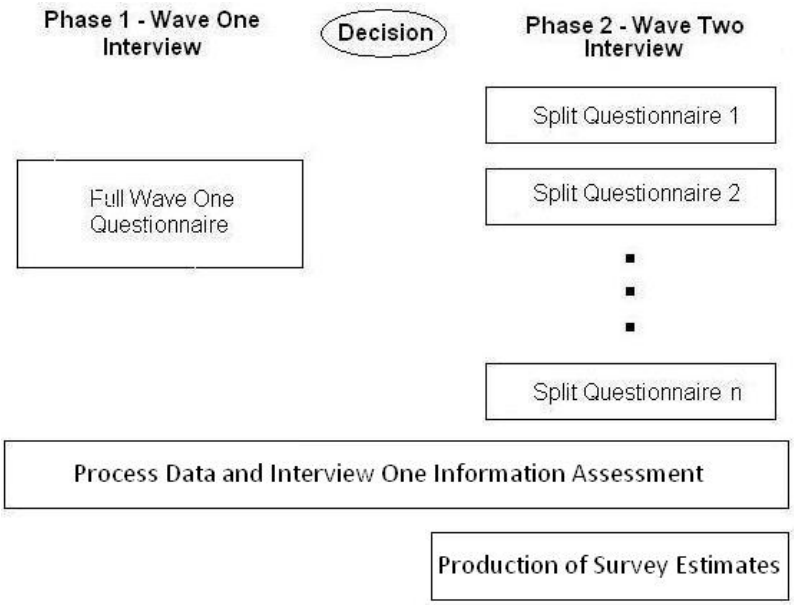

Figure 2.3: Illustration of a responsive split questionnaire design

statistics.

2.4.2.2 Estimation under a responsive split questionnaire

Given that we have laid the foundation for the general design of a responsive split questionnaire, we now discuss issues pertaining to estimation under a responsive split questionnaire. First, suppose that in the absence of a split questionnaire design we are interested in estimating the mean of some population characteristic, denoted by $\bar{y}$. If we have some general probability sample design, then we can use the standard design-based estimator for the mean given by

$$\hat{\bar{y}} = \left(\sum_{i \in S} w_i \right)^{-1} \left(\sum_{i \in S} w_i y_i \right) \tag{2.4}$$

where $\pi_i = P(i \in S)$ is the probability of inclusion into the sample, S, for unit $i \in U$ (where U denotes the set of N members in the target population); and, $w_i = \pi_i^{-1}$ is the inverse-probability, or design, weights for unit $i \in U$. This is famously known as the Horvitz-Thompson estimator of the population mean (Horvitz and Thompson, 1952).

The Horvitz-Thompson estimator of the population mean has the following properties. If we let $Z_i = 1$ if the i^{th} unit is included in the sample and zero otherwise, then we can rewrite $\hat{\bar{y}}$ as a summation over the full **population** as follows.

$$\hat{\bar{y}} = \left(\sum_{i \in U} Z_i w_i\right)^{-1} \left(\sum_{i \in U} Z_i w_i y_i\right) = \left(\sum_{i \in U} Z_i / \pi_i\right)^{-1} \left(\sum_{i \in U} Z_i y_i / \pi_i\right) \tag{2.5}$$

Using a Taylor Series Linearization, we can express $\hat{\bar{y}}$ as a linear function of **population** quantities

$$\hat{\bar{y}} \approx \bar{y} + \frac{1}{N}\left(\sum_{i \in U} Z_i y_i / \pi_i - \bar{y} \sum_{i \in U} Z_i / \pi_i\right) \tag{2.6}$$

then the approximate[7] design expectation (with respect to the full sample selection),

[7] This is the approximate design expectation because we are computing the expectation of the Taylor Series linearized quantity as opposed to $\hat{\bar{y}}$.

denoted as $E(\hat{\bar{y}})$, can be calculated as follows.

$$\begin{aligned}
E(\hat{\bar{y}}) &\approx E\left[\bar{y} + \frac{1}{N}(\sum_{i\in U} Z_i y_i/\pi_i - \bar{y}\sum_{i\in U} Z_i/\pi_i)\right] \\
&= \bar{y} + \frac{1}{N}\left[\sum_{i\in U} E(Z_i) y_i/\pi_i - \bar{y}\sum_{i\in U} E(Z_i)/\pi_i\right] \\
&= \bar{y} + \frac{1}{N}\left[\sum_{i\in U} \pi_i y_i/\pi_i - \bar{y}\sum_{i\in U} \pi_i/\pi_i\right] \\
&= \bar{y} + \frac{1}{N}\left[\sum_{i\in U} y_i - \bar{y}\sum_{i\in U} 1\right] \\
&= \bar{y} + \bar{y} - \frac{1}{N} N\bar{y} \\
E(\hat{\bar{y}}) &\approx \bar{y}
\end{aligned} \tag{2.7}$$

This implies that $\hat{\bar{y}}$ is approximately design-unbiased for $\bar{y}$. Furthermore, we have the approximate design variance, denoted as $V(\hat{\bar{y}})$, given below, with π_{ij} denoting the joint inclusion probability of units i and j.

$$\begin{aligned}
V(\hat{\bar{y}}) &\approx V\left[\bar{y} + \frac{1}{N}(\sum_{i\in U} Z_i y_i/\pi_i - \bar{y}\sum_{i\in U} Z_i/\pi_i)\right] \\
&= \frac{1}{N^2}\left[\sum_{i\in U} V(Z_i) y_i^2/(\pi_i)^2 + \bar{y}^2\sum_{i\in U} V(Z_i)/(\pi_i)^2 - 2\bar{y}\sum_i \sum_{j>i} \frac{y_i}{\pi_i\pi_j} C(Z_i, Z_j)\right] \\
&= \frac{1}{N^2}\left[\sum_{i\in U} \frac{\pi_i(1-\pi_i) y_i^2}{\pi_i^2} + \bar{y}^2\sum_{i\in U} \frac{\pi_i(1-\pi_i)}{\pi_i^2} - 2\bar{y}\sum_i \sum_{j>i} \frac{y_i}{\pi_i\pi_j}(\pi_{ij} - \pi_i\pi_j)\right] \\
&= \frac{1}{N^2}\left[\sum_{i\in U} \left(\frac{1-\pi_i}{\pi_i}\right) y_i^2 + \bar{y}^2\sum_{i\in U} \left(\frac{1-\pi_i}{\pi_i}\right) - 2\bar{y}\sum_i \sum_{j>i} y_i \left(\frac{\pi_{ij} - \pi_i\pi_j}{\pi_i\pi_j}\right)\right] \\
V(\hat{\bar{y}}) &\approx \frac{1}{N^2}\left[\sum_{i\in U} \left(\frac{1-\pi_i}{\pi_i}\right)(y_i^2 + \bar{y}^2) - 2\bar{y}\sum_i \sum_{j>i} y_i \left(\frac{\pi_{ij} - \pi_i\pi_j}{\pi_i\pi_j}\right)\right]
\end{aligned} \tag{2.8}$$

However, if we implement a split questionnaire design in the second interview,

then we would need to account for this in the estimator. Effectively, we would need to modify the design weights to compensate for the probability by which we administer a particular question, or equivalently, a set of questions, or split questionnaire form. If we denote p_{ik} as the probability with which we administer the k^{th} question, this subsampling probability is essentially the decision rule, then a natural candidate to estimate the quantity $\hat{\bar{y}}$ (and subsequently $\bar{y}$) from the sample members receiving the k^{th} question would be

$$\hat{\bar{y}}_{SQ} = \left(\sum_{i \in S_k} w_i^* \right)^{-1} \left(\sum_{i \in S_k} w_i^* y_i \right) \tag{2.9}$$

where S_k is the set of sample members receiving the k^{th} question[8]; and, $w_i^* = w_i / p_{ik}$ is the set of modified design weights. This type of modification is consistent with the theory for multi-phase sampling (Cochran, 1977; Särndal et al., 1992).

The estimator $\hat{\bar{y}}_{SQ}$ has the following feature. If we let $\alpha_{ik} = 1$ if $i \in S_k$ and zero otherwise, then we can rewrite $\hat{\bar{y}}_{SQ}$ as a summation over the full **sample**.

$$\hat{\bar{y}}_{SQ} = \left(\sum_{i \in S} \alpha_{ik} w_i^* \right)^{-1} \left(\sum_{i \in S} \alpha_{ik} w_i^* y_i \right) = \left(\sum_{i \in S} \alpha_{ik} w_i / p_{ik} \right)^{-1} \left(\sum_{i \in S} \alpha_{ik} w_i y_i / p_{ik} \right) \tag{2.10}$$

Using a Taylor Series Linearization, we can express $\hat{\bar{y}}_{SQ}$ as a linear function of **full-**

[8]We note that $S_k \subseteq S \subseteq U$.

sample quantities.

$$\hat{\bar{y}}_{SQ} \approx \hat{\bar{y}} + (\sum_{i \in S} w_i)^{-1} \left[\sum_{i \in S} \alpha_{ik} w_i y_i / p_{ik} - \hat{\bar{y}} \sum_{i \in S} \alpha_{ik} w_i / p_{ik} \right] \tag{2.11}$$

then the approximate design expectation, with respect to the split questionnaire design and conditional on the full, initial sample, S, denoted as $E_{SQ}(\hat{\bar{y}}_{SQ}|S)$, follows.

$$\begin{aligned}
E_{SQ}(\hat{\bar{y}}_{SQ}|S) &\approx E_{SQ} \left[\hat{\bar{y}} + (\sum_{i \in S} w_i)^{-1} \left\{ \sum_{i \in S} \alpha_{ik} w_i y_i / p_{ik} - \hat{\bar{y}} \sum_{i \in S} \alpha_{ik} w_i / p_{ik} \right\} |S \right] \\
&= \hat{\bar{y}} + (\sum_{i \in S} w_i)^{-1} \left\{ \sum_{i \in S} E_{SQ}(\alpha_{ik}|S) w_i y_i / p_{ik} - \hat{\bar{y}} \sum_{i \in S} E_{SQ}(\alpha_{ik}|S) w_i / p_{ik} \right\} \\
&= \hat{\bar{y}} + (\sum_{i \in S} w_i)^{-1} \left\{ \sum_{i \in S} p_{ik} w_i y_i / p_{ik} - \hat{\bar{y}} \sum_{i \in S} p_{ik} w_i / p_{ik} \right\} \\
&= \hat{\bar{y}} + (\sum_{i \in S} w_i)^{-1} \left\{ \sum_{i \in S} w_i y_i - \hat{\bar{y}} \sum_{i \in S} w_i \right\} \\
&= \hat{\bar{y}} + \hat{\bar{y}} - \hat{\bar{y}} (\sum_{i \in S} w_i)^{-1} \sum_{i \in S} w_i \\
E_{SQ}(\hat{\bar{y}}_{SQ}|S) &\approx \hat{\bar{y}}
\end{aligned} \tag{2.12}$$

This implies that $\hat{\bar{y}}_{SQ}$ is approximately design-unbiased for $\hat{\bar{y}}$, the full-sample estimator of $\bar{y}$. Furthermore, we have the approximate design variance, with respect to the split questionnaire design and conditional on the full, initial sample, S, denoted as $V_{SQ}(\hat{\bar{y}}_{SQ}|S)$, given as follows, with p_{ijk} denoting the joint inclusion probability of units i and j into S_k.

$$
\begin{aligned}
V_{SQ}(\hat{\bar{y}}_{SQ}|S) &\approx V_{SQ}\left[\hat{\bar{y}} + (\sum_{i\in S} w_i)^{-1}\left[\sum_{i\in S}\alpha_{ik}w_i y_i/p_{ik} - \hat{\bar{y}}\sum_{i\in S}\alpha_{ik}w_i/p_{ik}\right]|S\right] \\
&= (\sum_{i\in S} w_i)^{-2}\left[\sum_{i\in S} V_{SQ}(\alpha_{ik}|S)w_i^2 y_i^2/(p_{ik})^2 + \hat{\bar{y}}^2\sum_{i\in S} V_{SQ}(\alpha_{ik}|S)w_i^2/(p_{ik})^2\right] \\
&\qquad - (\sum_{i\in S} w_i)^{-2}\left[2\hat{\bar{y}}\sum_i\sum_{j>i}\frac{y_i w_i w_j}{p_{ik}p_{jk}}C_{SQ}(\alpha_{ik},\alpha_{jk}|S)\right] \\
&= (\sum_{i\in S} w_i)^{-2}\left[\sum_{i\in S}\frac{p_{ik}(1-p_{ik})w_i^2 y_i^2}{p_{ik}^2} + \hat{\bar{y}}^2\sum_{i\in S}\frac{p_{ik}(1-p_{ik})w_i^2}{p_{ik}^2}\right] \\
&\qquad - (\sum_{i\in S} w_i)^{-2}\left[2\hat{\bar{y}}\sum_i\sum_{j>i}\frac{y_i w_i w_j}{p_{ik}p_{jk}}(p_{ijk}-p_{ik}p_{jk})\right] \\
&= (\sum_{i\in S} w_i)^{-2}\left[\sum_{i\in S}\left(\frac{1-p_{ik}}{p_{ik}}\right)w_i^2 y_i^2 + \hat{\bar{y}}^2\sum_{i\in S}\left(\frac{1-p_{ik}}{p_{ik}}\right)w_i^2\right] \\
&\qquad - (\sum_{i\in S} w_i)^{-2}\left[2\hat{\bar{y}}\sum_i\sum_{j>i} y_i w_i w_j\left(\frac{p_{ijk}-p_{ik}p_{jk}}{p_{ik}p_{jk}}\right)\right] \\
V_{SQ}(\hat{\bar{y}}_{SQ}|S) &\approx (\sum_{i\in S} w_i)^{-2}\left[\sum_{i\in S}\left(\frac{1-p_{ik}}{p_{ik}}\right)w_i^2(y_i^2+\hat{\bar{y}}^2) - 2\hat{\bar{y}}\sum_i\sum_{j>i} y_i w_i w_j\left(\frac{p_{ijk}-p_{ik}p_{jk}}{p_{ik}p_{jk}}\right)\right]
\end{aligned}
\tag{2.13}
$$

The expectation with respect to the original sample selection of the quantity given in equation (2.13) reflects the *added* variance (to the overall sampling variance) attributable to the split questionnaire. It is important to note that $V_{SQ}(\hat{\bar{y}}_{SQ}|S)$ represents only one aspect of the total variance. In Section 2.6.1, equation (2.24), we provide the standard decomposition of variance formula reflecting both phases of sampling – initial sample selection and split questionnaire design.

Lastly, under certain types of designing the split questionnaire, a simplification of equation (2.13) is achieved. Specifically, if subsample membership was determined

by Bernoulli or Poisson sampling, then $p_{ijk} = p_{ik}p_{jk}$ for $i \neq j \in S$ and we have the following.

$$C_{SQ}(\alpha_{ik}, \alpha_{jk}|S) = p_{ijk} - p_{ik}p_{jk} = p_{ik}p_{jk} - p_{ik}p_{jk} = 0 \tag{2.14}$$

Thus, equation (2.13) simplifies to

$$\begin{aligned}
V_{SQ}(\hat{\bar{y}}_{SQ}|S) &\approx (\sum_{i\in S} w_i)^{-2} \left[\sum_{i\in S}\left(\frac{1-p_{ik}}{p_{ik}}\right) w_i^2(y_i^2 + \hat{\bar{y}}^2) - 2\hat{\bar{y}}\sum_i \sum_{j>i} y_i w_i w_j \left(\frac{p_{ijk} - p_{ik}p_{jk}}{p_{ik}p_{jk}}\right)\right] \\
&= (\sum_{i\in S} w_i)^{-2} \left[\sum_{i\in S}\left(\frac{1-p_{ik}}{p_{ik}}\right) w_i^2(y_i^2 + \hat{\bar{y}}^2) - 2\hat{\bar{y}}\sum_i \frac{y_i w_i^2}{p_{ik}^2} V_{SQ}(\alpha_{ik}|S)\right] \\
&= (\sum_{i\in S} w_i)^{-2} \left[\sum_{i\in S}\left(\frac{1-p_{ik}}{p_{ik}}\right) w_i^2(y_i^2 + \hat{\bar{y}}^2) - 2\hat{\bar{y}}\sum_i y_i w_i^2 \left(\frac{1-p_{ik}}{p_{ik}}\right)\right] \\
&= (\sum_{i\in S} w_i)^{-2} \left[\sum_{i\in S}\left(\frac{1-p_{ik}}{p_{ik}}\right) w_i^2(y_i^2 - 2y_i\hat{\bar{y}} + \hat{\bar{y}}^2)\right] \\
V_{SQ}(\hat{\bar{y}}_{SQ}|S) &\approx (\sum_{i\in S} w_i)^{-2} \left[\sum_{i\in S}\left(\frac{1-p_{ik}}{p_{ik}}\right) w_i^2(y_i - \hat{\bar{y}})^2\right]
\end{aligned} \tag{2.15}$$

Ultimately, since the estimator $\hat{\bar{y}}_{SQ}$ has desirable full-sample properties, e.g., design-unbiasedness, we consider these types of estimators in this dissertation.

From the preceding discussion, it is evident that the decision rule (i.e., sub-sampling probability) is a key component of a responsive design. This is because not only is the decision rule incorporated into the estimator, but also it dictates which design modification is given to a particular respondent in a subsequent phase. The success of a responsive design depends on how well the decision rule chooses the appropriate design modification; therefore, considerable effort should be devoted to developing decision rules. This fact is reinforced by Groves and Heeringa (2006) when they recommend that "the field needs to study how the survey statistician

should best model paradata from early phases."

These models need conceptual development, sensitivity analyses regarding alternative specifications, and diagnostic assessment so that survey researchers can fully understand the extent to which design modifications achieve the goals of the responsive design and how the design modifications affect the error properties of the resulting statistics (Groves and Heeringa, 2006). In the next section we offer two perspectives on how to incorporate prior information about the sample unit into formulating the decision rules. We operationalize these perspectives in the formal development of decision rules in Chapter 4.

2.4.3 Two perspectives on developing decision rules

In this section, we identify two perspectives for developing decision rules for a responsive split questionnaire. We motivate these perspectives from (1) multiphase sampling for stratification and (2) the methods used to compensate for nonresponse.

Before we describe these two perspectives, there are two issues that are relevant to this discussion. The first issue is that our responsive split questionnaire is based, in part, on the premise that we want to ask questions that are relevant to the sample unit. Said differently, if we knew (or were fairly confident) about the occurrence of a particular event for a sample unit before the next interview, we would want to ask about that event at the next interview. In the context of the CE, the event is the purchase of a particular item. Sample units that make a particular purchase could be considered to come from a subpopulation defined by that purchase. If our goal is to

ask about that purchase during the next interview without unnecessarily burdening those sample units who did not have that purchase, then we would like to do a more targeted sampling from the subpopulation with the purchase. Therefore, we need a mechanism by which we can predict the likelihood of purchasing the particular item and/or stratify or classify the sample units into categories based on that likelihood and then oversample from the group(s) that contains the likely purchasers.

To reiterate, the rationale for designing the split questionnaire in this way is two-fold. First it directly addresses one of the primary concerns of implementing a standard, or completely random, split questionnaire design, namely, simple random question assignment under-identifies low prevalence events. Second, the "tailoring" of questions minimizes the number of irrevelant questions for which the respondent has the opportunity to learn negative reporting behaviors.

The second issue is that any split questionnaire design creates missing data because not every sample member is asked every question. Unlike many nonresponse problems, however, the missing data mechanism is known because we are creating the missing data at the design stage by conditioning on the variables used in the decision mechanism. Thus, we can adjust for the mechanism at the analysis stage. Given these two issues, it seems reasonable that if there was a technique that is frequently used in classification and nonresponse problems, then that technique should be a prime candidate to explore when developing decision rules for the responsive split questionnaire design. Once the technique is identified, the next question for consideration is whether to use the technique directly or indirectly in developing the decision rule. So, the two perspectives are (1) to use the technique or model outputs

directly as the decision rule or (2) to use the technique or model outputs *indirectly* by treating them as inputs into the decision rule.

We expand on what we mean by these two perspectives. Take for example the technique of logistic regression. Logistic regression is often used to estimate the probability of the occurrence of an event (Agresti, 2002). In the context of the CE, we can use logistic regression to predict the probability that a sample unit has a certain type of purchase. One of the outputs of the logistic regression model is a predicted probability, or estimated propensity score. Once we have the propensity score for each sample unit, we can use it directly as the decision rule. In other words, the estimated propensity score becomes the probability with which we ask a particular question in the subsequent interview. This is a reasonable approach because the estimated propensity score represents our best guess of the likelihood that the sample unit will have the particular purchase. We refer to this way of developing decision rules as the *direct* method.

The other perspective is that we can *indirectly* incorporate the model outputs into the decision rule. In the survey setting, logistic regression has also been used to make response propensity score adjustments when unit nonresponse is present (Rosenbaum and Rubin, 1983). In some of these applications, the propensity scores are used to stratify the sample into groups based on their response propensities. Once the strata are formed, then the same adjustment is applied to each (responding) unit in each stratum. This formulation of the problem closely follows the ideas of multiphase sampling for stratification (Särndal et al., 1992) because the survey statistician uses information collected in a prior phase to stratify the sample into

groups (in the case of nonresponse adjustments, the groups would be based on the likelihood of responding). Applying these concepts to our research problem, we can use the propensity scores to stratify the sample based on their likelihood of purchasing a particular item. This now becomes a type of stratified sampling design, so we can use the theory behind optimal or Neyman allocation (Cochran, 1977) to devise the sampling fractions for each stratum. These sampling fractions become the decision rules for asking questions in the subsequent phase. We refer to this way of developing decision rules as the *indirect* method.

It is worth noting that logistic regression is not the only technique by which we can obtain propensity scores for use in a decision rule. Propensity scores may be directly obtained via other statistical methods or methods that mimic more traditional sampling techniques. One such method that mimics traditional sampling techniques is probability proportional-to-size (PPS) sampling (Cochran, 1977). PPS sampling is a sampling technique that assigns larger probabilities of inclusion to units that comprise a larger proportion of the total. In the context of CE, we use this method to assign higher subsampling probabilities to expenditure categories that comprise a larger share of that sample unit's total expenditures. In PPS sampling, it is standard practice to base the measure of size on a previous administration of the survey. In this research, we can base this measure of size on expenditure information collected in the first interview. Additionally, this method may serve as one reasonable method for comparison because it represents a traditional sampling technique. Therefore, in our development of decision rules in Chapter 4, we also consider a second direct method in which we obtain propensity scores based on this

type of PPS sampling design.

It is worth reiterating that the primary distinction between the direct method and the indirect method is that, regardless of the technique used to derive the propensity scores, with the indirect method, there is an additional step (or set of steps) after deriving the propensity scores, to obtain the decision rules. On the other hand, with the direct method, the propensity scores are the decision rules.

Regardless of which perspective and technique is used to obtain the decision rules, it is important to consider a range of models, methods, and ways of incorporating the prior information. This approach is consistent with the previously cited recommendation from Groves and Heeringa (2006) regarding responsive designs; specifically, that it is the duty of survey statistician to provide sensitivity analyses for alternative specifications of the models and diagnostic assessments of them. These diagnostic assessments not only include whether or not the responsive design is successful (in our case, whether the split questionnaire is truly being customized to the respondent) but also whether stakeholder needs are met. To address the former issue, we must identify a series of evaluation criteria that enable us to judge the success of the responsive design. We present these criteria in Section 2.6.2. To address the latter issue, it may be necessary to modify the decision rules by imposing constraints on the design to ensure that requirements are met. We discuss this issue in the next section.

As a final note, in this section, we only provided a theoretical motivation for the two perspectives. In Chapter 4, we provide explicit mathematical representations of the problem formulations under each perspective. Also, in Chapter 4, we formalize

the process of developing decision rules under each perspective and assess their relative merits.

2.4.4 Modifying decision rules

A diagnostic assessment of the decision rules may reveal that some key survey objectives (e.g., providing a full microdata file to data users) or certain precision requirements of primary stakeholders are not met under a given set of decisions. It may also be required that the split questionnaire design contain certain features and under a given set of decision rules those features are lacking. Therefore it may be necessary to modify the decision rules by imposing constraints on the system so that the key survey objectives or precision requirements are met or that the survey design contains the necessary features.

To demonstrate why this may be the case, recall that we motivated this research, in part, by wanting to reduce burden on the sample units. Suppose for a moment that instead of simply reducing burden on the sample members, we wanted to balance the burden reduction across sample units so that each sample unit would roughly be administered surveys that are more or less equivalent across the dimension of burden. A crude measure of burden may be the number of survey items administered to each sample unit. Say, that we wanted to bound the expected number of survey items administered to the sample unit to be between $N_{y,lb}$ and $N_{y,ub}$ (with $N_{y,lb} < N_{y,ub}$). First, define $\{p_{ik}\}$ as the current set of decision rules. If we let N_i be the number of survey items administered to the i^{th} sample unit; and,

$\alpha_{ik} = 1$ if sample unit i receives question k and 0 otherwise, then we can impose the constraint identified in equation (2.16) on $\{p_{ik}\}$.

$$N_{y,lb} \leq E(N_i) = E(\sum_{k=1}^{K} \alpha_{ik}) = \sum_{k=1}^{K} E(\alpha_{ik}) = \sum_{k=1}^{K} p_{ik} \leq N_{y,ub} \tag{2.16}$$

Essentially to impose the constraint, we would apply (and solve for) an adjustment factor, λ_{ik}, to p_{ik}, thus $\{\lambda_{ik}p_{ik}\}$ become the modified set of decision rules. This modified set would meet the constraint on the expected number of survey items administered to the sample units to be between $N_{y,lb}$ and $N_{y,ub}$.

Another measure of burden may be the interview length, measured in minutes. If we wanted to restrict the expected total interview length for the i^{th} sample unit to be less than T and we let T_i be the interview length (in minutes) for i^{th} sample unit and t_{ik} be the time spent (in minutes) on answering the question related to expenditure category k, and α_{ik} as defined above, then the constraint would take form.

$$E(T_i) = E(\sum_{k=1}^{K} \alpha_{ik}t_{ik}) = \sum_{k=1}^{K} E(\alpha_{ik})t_{ik} = \sum_{k=1}^{K} p_{ik}t_{ik} \leq T \tag{2.17}$$

Regardless of the type of constraint, imposing a constraint to meet stakeholder needs or satisfy requirements of the design may affect the original split questionnaire design's effectiveness in being "responsive" to the sample unit. In other words, by balancing burden across sample members we may negatively impact our ability to customize the survey to the *individual* respondent. Thus, there may be tradeoffs

between satisfying survey constraints and being successful with respect to the responsive design (or some other metric). We address some of the issues associated with striking a balance between satisfying constraints and being successful in terms of a responsive design in later sections.

In any case, to modify the decision rules, we impose constraints on the split questionnaire survey design (i.e., decision rules) and then maximize (or minimize) some objective function (by objective function we mean a function of one or more variables to be optimized), subject to the constraints. It is worth noting that finding an optimal solution to the set of adjustment factors (e.g., λ_{ik}) or decision rules (p_{ik}) may be impossible. This is because the combination of the size of the population and the number of questions makes the problem intractable. For these problems, standard software may not accommodate the number of decision rules thus, an alternative formulation of the problem may be required. We discuss one formulation of the design of a responsive split questionnaire in which we carry out the development of decision rules under a constrained system in Chapter 4. In the next section, we briefly identify relevant features of survey design and survey design optimization and discuss methods that survey designers may use to impose constraints on the decision rules.

2.5 Optimal survey design

In Section 2.3.2, we provided illustrations of several split questionnaire designs. We also discussed the relative merits of each by identifying which designs could

be used for particular purposes or analytic objectives. Then in Section 2.4.4, we suggested that decision rules for a responsive split questionnaire design may need to be modified in order to meet some survey objectives. In this section, we make a few general comments about survey design, outline the basic approach to survey design optimization, and provide an optimality framework that will be useful when evaluating a responsive split questionnaire. Although the comments on survey design are very general, they are an essential component of this discussion because they highlight the issues that need to be explored in this research and provide guidance on how to evaluate our proposed methodology.

2.5.1 Survey design

When exploring the use of alternative, and perhaps non-standard, methods for collecting survey data, it is important to understand why the survey is being conducted. This amounts to identifying, to the extent possible, the primary purposes and objectives of the survey. By identifying and prioritizing key survey objectives, the survey designer can make more informed design decisions about how best to collect the data in order to satisfy those objectives.

A convenient starting point for accomplishing this task is Kish's (1988) review of multipurpose surveys. In his review he provides a hierarchy of six primary purposes of surveys. They are: (1) calculation of diverse statistics; (2) characterization of diverse statistics; (3) collection of multiple variables; (4) multi-subject surveys; (5) continuation of survey operations; and (6) development of master frames. Most

surveys are intended to satisfy multiple objectives from this hierarchy. For instance, items (1) – (4) are particularly relevant for the CE. One of the primary objectives of the current CE is to meet the need for timely and detailed information on the spending patterns of different types of families (BLS *Handbook of Methods*, 2007). This one objective clearly involves the calculation and characterization of diverse statistics using inputs from multiple variables, collected on many sample members. Given this example, it is evident that the CE survey is inherently multipurpose. Therefore, when exploring the use of a responsive split questionnaire for implementation in a redesigned CE Survey (or any survey endeavor), consideration must be given to the multipurpose nature of the survey.

Kish (1965) provides a definition of survey design that has two key aspects – survey objectives and sample design. Survey objectives includes defining survey variables, identifying methods of observation, methods of analysis, utilization of results, and desired precision. He identifies two processes associated with sample design – selection of sample units and estimation from the sample units. It is important to recognize that these two key aspects are dependent on each other, i.e., survey design is a two-way process. In our case, the sampling mechanism is related to the objective of reducing burden while not missing important events (e.g., purchases of particular items). The method of sample selection (equivalently, asking questions in a responsive manner) will have a direct bearing on the methods of analysis (e.g., accounting for the sampling mechanism in the estimate).

When developing a responsive split questionnaire to meet some combination of the purposes enumerated previously, conflicts may arise because satisfying every

purpose is challenging. Kish (1988) identifies ten areas of conflict that may arise during this process. These ten areas provide guidance on what issues and comments we should make when evaluating a set of decision rules for a responsive split questionnaire. The ten areas are: (1) sample sizes; (2) relation of biases to sampling errors; (3) allocation of sample among domains; (4) allocation of sample among strata; (5) choice of stratification variables; (6) cluster sizes; (7) measures of size for clusters; (8) retaining sample units; (9) design over time; and, (10) sampling errors. Fortunately not every potential conflict needs to be considered because some conflicts are tied only to one purpose. However, Kish recommends that key considerations should always be given to the interplay among sample sizes and biases because those conflicts tend to be ubiquitous. We provided empirical evidence in Section 2.3.1 as to why sample size issues need to be considered in this research, but we will also investigate any potential biases in key estimates that may arise when implementing a set of decision rules for a responsive split questionnaire. Furthermore, we can also view members of the sample who purchase a particular item as comprising a domain of interest, so we investigate issues associated with characterizing members of this domain. With respect to methods for decision rules, specifically PPS, we have to be cognizant of the issues associated with our measures of size because the success of the set of decision rules directly depends on the utility of the measures of size. In sum, all of these issues are interrelated and require a balanced consideration in this research.

Recall that the goal of this dissertation is to explore the extent to which, if any, jointly considering information on the survey item as well as the sample

unit in the design of a split questionnaire will improve its efficiency. The survey of application's primary purpose is to produce estimates of means of several population characteristics (e.g., the mean expenditure on item k for $k = 1, 2, \ldots, K$ per sample unit). We hope to identify which method would be optimal under various conditions and constraints. For this research optimal means choosing the best element from some set of available alternatives.

In survey design, there are two perspectives on what is meant by optimal. The first is an "ideal" notion of optimal. This notion refers to the scenario of operating under an unconstrained system. As a simple example, survey research suggests that nonresponse rates tend to be lower with personal visit surveys than they are with either mail or telephone surveys. However, personal visit interviews are generally more expensive than telephone interviews (Groves et al., 2004). So, if the survey organization was unconstrained by budgets and if it wanted to optimize, or maximize, the response rate, then it should choose to administer a personal-visit survey.

Often, however, surveys must operate under constraints in particular, monetary constraints and, perhaps, those represented by equations (2.16) and (2.17). In practice, survey design must balance a wide range of factors with a finite set of resources to conduct a survey. Thus, there is a second notion of optimum which is referred to as a "practical" optimum. This notion of optimal can be thought of as the one that is achieved after meeting certain constraints and conditions imposed on the system. Our research explores the development of decision rules from both notions of optimal. The "ideal" notion of optimal corresponds to the situation in

which there was no modification to the decision rules while the "practical" optimum corresponds to the situation of constraining the decision rules to meet survey constraints.

2.5.2 Basic approach to optimization

One approach to characterizing constraints for a survey design objective is to express it as a mathematical optimization problem and then use mathematical programming methods to solve the optimization problem. Mathematical programming encompasses a variety of methods when solving optimization problems (by solving we mean choosing the "best" solution) subject to many constraints. The advantage of these methods is that they provide a formal way of solving complex allocation problems. There are four primary components to any optimization problem and these are relevant to survey design optimization. They are: (1) the objective function; (2) decision variables; (3) parameters; and, (4) constraints. In this section, we define each component and provide examples of each that have been used in other survey design problems.

The first component of an optimization problem is the objective function, or a function of one or more variables to be optimized, and by optimized we mean either maximized or minimized. An example of an objective function is the sampling variance of a population estimator. In equation (2.18) we provide the standard sampling variance formula for the stratified random sampling variance of the estimated mean (Cochran, 1977).

In equation (2.18), we are interested in the sampling variance of an estimated mean, $\hat{\bar{y}}$. We define the following: the subscript h is the stratum index; N_h is the total number of population units in stratum h; n_h is the number of units in the sample from stratum h; S_h^2 is the true population variance of y for the h^{th} stratum; and, $N = \sum_h N_h$.

$$V(\hat{\bar{y}}) = \frac{1}{N^2} \sum_h N_h(N_h - n_h)\frac{S_h^2}{n_h} \tag{2.18}$$

The second component of an optimization problem is the set of decision variables. These are the quantities that are adjusted in order to find a solution to the optimization problem. These quantities are what the survey designer is most interested in. Examples of decision variables may include full or stratum-specific sample sizes, denoted by n and n_h, respectively, (from equation [2.18]).

The third component is a set of fixed inputs, treated as constants, and are known parameters. Examples of these can be identified using equation (2.18). For instance, population stratum variances and stratum-specific population sizes, denoted by S_h^2 and N_h, respectively, could be considered parameters in a survey optimization problem. Two other important examples of parameters not contained in equation (2.18) are cost components, these include the "cost" of making a measurement on a specific sample unit[9], and timing data, or the amount of time required to complete a particular survey item or group of items. Stratum specific cost components can be denoted as c_h and item k completion time data can be denoted

[9]Cost can be broadly defined to not only refer to dollar amount, but other measures of cost like burden.

t_k.

The final component is the set of constraints. These are the restrictions on the decision variables. Recall that in Section 2.4.4 we identified possible constraints (see equations [2.16] and [2.17]) that might be relevant to a responsive split questionnaire design – bounding the expected total number of survey items administered to a sample unit or constraining the expected interview length. Other examples of constraints may include specifying interviewer workloads or "cost" constraints. It should be noted that cost contraints are similar in spirit to constraints on interview length. Survey operations are often constrained by budgets and equation (2.19) identifies a simple linear cost constraint in which the overall survey cost, denoted by C, is composed of a fixed cost component, C_0, as well as a variable component that depends on the number of observations made in each stratum (Cochran, 1977).

$$C = C_0 + \sum_h c_h n_h \tag{2.19}$$

It is worth noting that there are numerous constraints that survey statisticians can impose on the survey design. An overly constrained system, however, may render an infeasible solution. Therefore it is necessary to consider as many constraints as needed to capture the complexity and true nature of the problem, while ensuring that a solution can be obtained.

For illustrative purposes only, we detail how the basic approach to optimization is performed using some of the specific examples above. A simple variance-cost optimization problem is formulated as follows: we would like to determine the set of

stratum sample sizes $\{n_h\}$ that minimize (2.18) subject to the constraint identified in (2.19). Using either an application of the Cauchy-Schwarz Inequality (Cochran, 1977) or Lagrange multipliers (Varberg and Purcell, 1997), one can find that the solution is

$$n_h = n\frac{N_h S_h/\sqrt{c_h}}{\sum_h N_h S_h/\sqrt{c_h}} \tag{2.20}$$

with $n = [\sum N_h S_h \sqrt{c_h}]^{-1}(C - C_0)[\sum (N_h S_h/\sqrt{c_h})]$ when cost is fixed. In other words, the set of $\{n_h\}$ identified in (2.20) will minimize (2.18) subject to (2.19). The implication of allocating sample based on (2.20) is that we sample more from strata that are heterogeneous, i.e., have large S_h values, and less from strata that are expensive, i.e., have high c_h values.

Although not identified as a main component of the optimization problem, the method used to find the solution is another critical element. As demonstrated by the example above, simple problems can usually be solved using Lagrange multipliers or applications of the Cauchy-Schwarz inequality (Cochran, 1977). However, for complex situations more sophisticated techniques are usually needed. Mathematical programming methods are useful for solving these complex problems. We make use of mathematical programming methods (Section 4.4) to develop a set of decision rules under the indirect perspective of incorporating prior information about the sample unit.

2.5.3 General optimality framework

While the basic approach to optimization is helpful when setting up survey design problems for computational exercises, a more general optimality framework is useful for characterizing the effectiveness of the procedures. This is because the optimization problem formulation only addresses one aspect of the effectiveness, e.g., minimization of variance. A general optimality framework helps survey designers understand the relationships among design decisions, survey data quality, and the utility of statistical products (e.g., official estimates of means, totals, or other quantities) from the collected survey data, and the successfulness of the procedures in being responsive. This framework is also helpful when characterizing tradeoffs among potentially competing goals of the design (e.g., customizing the survey to the individual respondent versus balancing burden among all sample units). This general optimality framework may also help the CE program office make decisions regarding their redesigned surveys.

Key components of this framework can be extracted and adapted from optimal design theory and statistical decision theory (Fedorov, 1972; Silvey, 1980; Berger, 1980). To develop this optimality framework, we provide the following notation. First, let $\mathcal{D}$ be the decision, or design, space; D denote the selected design feature (e.g., random mechanism for sampling); and, d be the realization of the specific design feature (e.g., sample). In addition, let Q be the optimality criterion (e.g., mean squared error of a survey statistic, one of the six dimensions of data quality outlined in Brackstone [1999]) and U be a utility function representing a stakeholder's

relative satisfaction with the design.

We define $\mathbf{X}$ to be a vector of observable auxiliary information which we partition into three components, i.e., $\mathbf{X} = (\mathbf{X}_R, \mathbf{X}_B, \mathbf{X}_C)$. We then have $\mathbf{X}_R$ to be a set of resources with which to conduct a survey (e.g., existing survey organization infrastructure, interviewing staff, computer and processing systems); $\mathbf{X}_B$ to be the bounds or constraints (e.g., interview length, burden); and, $\mathbf{X}_C$ to be the cost structure (e.g., per unit interview costs). We also allow for the possibility of other factors that are not directly controllable in real-time and we denote this vector as $\mathbf{Z}$. These may include changes to the underlying survey environment, where a specific example relevant to the CE survey is that of new products (e.g., iPads or Amazon Kindles). Expenditure information about new products is difficult, if not impossible, to obtain after survey design decisions have been made and resources have been allocated. Thus, changes in the underlying survey environment will have an impact on the optimality criteria and subsequent measure of utility for each stakeholder.

Using the above notation, the optimality criterion, Q, can be expressed as the following function.

$$Q = Q(\mathbf{D}, \mathbf{X}, \mathbf{Z}, \gamma) \tag{2.21}$$

Note that in equation (2.21) we also have a vector of parameters, denoted by γ. This vector contains parameters that are unknown and may have an impact on the optimality criterion. For example, there may be a change in the underlying purchase behavior since the model was developed. If one model has been used to develop the decision rules for a responsive split questionnaire, but because of changes

in purchase patterns across sample units, the original model is no longer relevant, then this will have a direct bearing on the measure of optimality. Again, it is essential to consider a range of models and continually conduct sensitivity analyses and diagnostic assessments to guard against these risks.

Finally, we express the stakeholder's utility function, equation (2.22), where β is a vector of parameters representing underlying perceptions of needs of individual stakeholders.

$$U = U(Q, \beta) \tag{2.22}$$

Given this representation of a stakeholder's utility, it is clear that for the same optimality criteria, individual stakeholders' perception of value, or utility, may still vary across stakeholders, due to differences in β. Said differently, while survey designers may make design decisions using one criterion (e.g., mean squared error of a particular statistic), the utility of the survey design may be high for one stakeholder, but quite low for another because that stakeholder has relatively low interest in the particular statistic. For example, some stakeholders' perception of data utility will depend heavily on the success of the survey in being tailored to the individual respondent, while for others tailoring may be much less important than ensuring that burden is balanced across sample units. Thus, having a clear understanding of various stakeholders' utility functions is an important component of the survey design process (including responsive split questionnaires). This understanding of different priorities among stakeholders, together with having a range of diagnostic and evaluation criteria, will help characterize the effectiveness of the methodology

from differing perspectives. This is consistent with the recommendation of Groves and Heeringa (2006).

2.6 Evaluation criteria

As indicated by the discussion in the preceding section, there are many useful criteria to judge the effectiveness of the design; however, the most appropriate criteria may depend on which feature of the design is being evaluated by a stakeholder. In this section, we identify and offer comments about several metrics from traditional survey sampling techniques (Section 2.6.1) and epidemiology (Section 2.6.2) that we use in the evaluation of our proposed methods.

2.6.1 Metrics from traditional survey sampling techniques

The first layer of evaluation for a responsive split questionnaire design should involve an investigation of the loss of information due to the reduced sample size receiving each survey item. A typical measure of the precision gained or lost by using a complex sample design instead of a simple random sample is the design effect, or deff (Lohr, 1999). The design effect is defined as the ratio of the sampling variance reflecting the intricacies of the design, denoted as $V_n(\hat{\theta})$, to the sampling variance that would have been obtained from a simple random sample (without replacement) of the same size of n elements, denoted as $V_{SRS,n}(\hat{\theta})$ (Groves, 1989). It is defined in (2.23).

$$\text{deff}(\hat{\theta}) = \frac{V_n(\hat{\theta})}{V_{SRS,n}(\hat{\theta})} \tag{2.23}$$

The design effect provides a measure of how much different the sample is from a simple random sample (or a sample where data can be treated as independent and identically distributed). A design effect less than unity is interpreted as a gain in precision over simple random sampling, while a design effect greater than unity is interpreted as a loss in precision over simple random sampling. With respect to our research, the design effect would provide an indication as to how much different the responsive split questionnaire is from simply randomly asking a subset of sample members the questions. So, a design effect less than unity would indicate a gain in precision over asking a random subset of sample members the particular question. It is worth pointing out that use of the design effect is also consistent with the empirical assessment of the relative sampling variances for two design modifications in the seminal paper on responsive designs by Groves and Heeringa (2006); therefore, its use to evaluate the loss of information in a responsive split questionnaire seems appropriate.

Using concepts from previous sections and linking the implementation of a responsive split questionnaire to standard two-phase sampling techniques, we demonstrate how the design effect can be used to evaluate features of a responsive split questionnaire. In our proposed implementation of a responsive split questionnaire, the first interview remains as is, and the second interview consists of the tailored set of survey questions. This is similar to the setup of two-phase sampling. Under

a general two-phase sampling design, a sample is selected by an arbitrary sample design during the first phase and information is collected from these units. With the aid of this information, a second phase sample is selected and the key survey variables are observed for every element of the second phase sample (Särndal et al., 1992). To reiterate, in our research, the first phase is the initial interview and the second phase is the (responsive) split questionnaire.

Using standard two-phase sampling techniques, the overall sampling variance of an estimator, e.g., mean, can be decomposed into two components – one component reflecting the variation due to the initial sample selection and the second component reflecting the additional variation incurred due to the subsampling (Särndal et al., 1992). Recall that in Section 2.4.2.4, we showed that the approximate design variance with respect to the split questionnaire is given in equation (2.13). We denote this quantity as $V_{SQ}(\hat{\bar{y}}_{SQ}|S)$. If we take the expected value of this quantity with respect to the original sample selection (i.e., $E_O[V_{SQ}(\hat{\bar{y}}_{SQ}|S)]$), then we have the additional variance component of the overall variance attributable to implementing a responsive split questionnaire. Furthermore, let $V(\hat{\bar{y}}_{SQ})$ denote the overall sampling variance of the mean estimator from a responsive split questionnaire and $V_O[E_{SQ}(\hat{\bar{y}}_{SQ}|S)]$ represent the variance attributable to the first phase, i.e., the original sample selection. Thus, we have the following.

$$V(\hat{\bar{y}}_{SQ}) = V_O[E_{SQ}(\hat{\bar{y}}_{SQ}|S)] + E_O[V_{SQ}(\hat{\bar{y}}_{SQ}|S)] \tag{2.24}$$

Using the subscript n to denote the number of sample units receiving a par-

ticular question under a split questionnaire design, the design effect for $\hat{\bar{y}}_{SQ}$ can be expressed as follows.

$$\text{deff}(\hat{\bar{y}}_{SQ}) = \frac{V_n(\hat{\bar{y}}_{SQ})}{V_{SRS,n}(\hat{\bar{y}})} \tag{2.25}$$

It is worth noting the elimination of the subscript SQ in the denominator of (2.25). This is because the definition of a design effect requires that we treat the quantity in the denominator as if it was obtained via a simple random sample, so we use the estimator for the mean from that sampling design.

One can calculate the design effect in the following way. First, assume that we sampled n_1 units for the first phase using a SRS out of a population of N units and for the second phase n units received the particular question, with S^2 denoting the population element variance for y, then (2.24) can be written as follows.

$$V(\hat{\bar{y}}_{SQ}) = \left(1 - \frac{n_1}{N}\right) \frac{S^2}{n_1} + E_O[V_{SQ,n}(\hat{\bar{y}}_{SQ}|S)] \tag{2.26}$$

Using the sampling variance formula for the mean under SRS, the design effect for $\hat{\bar{y}}_{SQ}$ becomes the following.

$$\text{deff}(\hat{\bar{y}}_{SQ}) = \frac{\left(1 - \frac{n_1}{N}\right) \frac{S^2}{n_1} + E_O[V_{SQ,n}(\hat{\bar{y}}_{SQ}|S)]}{\left(1 - \frac{n}{N}\right) \frac{S^2}{n}} \tag{2.27}$$

When the second phase sample is also determined by SRS, then we have the

following simplification[10] (with $n < n_1 < N$).

$$\begin{aligned}
\text{deff}(\hat{\bar{y}}_{SQ}) &= \frac{\left(1 - \frac{n_1}{N}\right)\frac{S^2}{n_1} + E_O[V_{SQ,n}(\hat{\bar{y}}_{SQ}|S)]}{\left(1 - \frac{n}{N}\right)\frac{S^2}{n}} \\
&= \frac{\left(1 - \frac{n_1}{N}\right)\frac{S^2}{n_1} + \left(1 - \frac{n}{n_1}\right)\frac{S^2}{n}}{\left(1 - \frac{n}{N}\right)\frac{S^2}{n}} \\
&= \frac{\left(1 - \frac{n_1}{N}\right)\frac{1}{n_1} + \left(1 - \frac{n}{n_1}\right)\frac{1}{n}}{\left(1 - \frac{n}{N}\right)\frac{1}{n}} \\
\text{deff}(\hat{\bar{y}}_{SQ}) &= \frac{n_1^2(N-n)}{n_1^2(N-n)} = 1
\end{aligned} \tag{2.28}$$

So, under a design in which both phases are determined by SRS, the design effect for $\hat{\bar{y}}_{SQ}$ is one. This is because this design does not use any Phase 1 information to select the sample in Phase 2. Therefore, nothing is gained or lost by this design. This fact becomes relevant for our research because we know exactly what the design effects should be for split questionnaires that are designed by randomly asking questions in the second phase (i.e., essentially "flipping a coin" to determined whether or not a sample unit is asked a particular question).

While the design effect provides a good measure of the gains or losses in efficiency for the design, it does not, however, fully allow us to make comparisons across designs since different responsive split questionnaires may result in different numbers of units getting asked the particular question. More specifically, the mere fact of asking more units a particular question may artificially deflate the sampling variance. One way to circumvent this problem is to standardize the variance com-

[10]This simplification is achieved by substituting the sample variance formula for the mean under SRS with a sample size of n out of n_1 for $E_O[V_{SQ,n}(\hat{\bar{y}}_{SQ}|S)]$.

ponent attributable to responsive split questionnaire. This facilitates comparisons across designs.

One way to standardize the variance components attributable to different responsive split questionnaires is to express them as functions of a fixed or common sample size. In equation (2.13), we showed that $V_{SQ}(\hat{\bar{y}}_{SQ}|S)$ can be expressed as follows.

$$V_{SQ}(\hat{\bar{y}}_{SQ}|S) \approx (\sum_{i\in S} w_i)^{-2} \left[\sum_{i\in S}\left(\frac{1-p_{ik}}{p_{ik}}\right) w_i^2(y_i^2+\hat{\bar{y}}^2) - 2\hat{\bar{y}}\sum_i \sum_{j>i} y_i w_i w_j \left(\frac{p_{ijk}-p_{ik}p_{jk}}{p_{ik}p_{jk}}\right)\right] \tag{2.29}$$

If we can obtain a value for $V_{SQ}(\hat{\bar{y}}_{SQ}|S)$, then we can use that value as an approximation to S_{SQ}^2/n where S_{SQ}^2 represents the element variance estimated under the split questionnaire design and n is the sample size receiving the question. Since we know how many units received the particular question, then we can solve for S_{SQ}^2, i.e., $S_{SQ}^2 = nV_{SQ}(\hat{\bar{y}}_{SQ}|S)$. Once we have an estimate of S_{SQ}^2 from each responsive split questionnaire design, we can use the values to approximate the added variance component as a result of the split questionnaire for a given sample size n^* as follows.

$$V_{SQ,n^*}(\hat{\bar{y}}_{SQ}|S) = \frac{S_{SQ}^2}{n^*} \tag{2.30}$$

If we have multiple split questionnaire designs, we can make comparisons among them on the basis of their respective $V_{SQ,n^*}(\hat{\bar{y}}_{SQ}|S)$ values because they are expressed in terms of a fixed sample size, n^*. Thus, we circumvent the issue of comparing methods that result in different numbers of sample units receiving each question.

As a final note, we initially intended to use optimal experimental design criteria, e.g., A- and D-optimality, for this part of the evaluation. This is consistent with the evaluation criteria used in the simulation study by Gonzalez and Eltinge (2008). One benefit of using these metrics would have been reducing the dimension of the comparison to a scalar quantity; thus, overall design decisions would have been made on the basis of one value versus a value for each estimate. While these metrics are appropriate for comparing designs that result in the same number of questions being administered, comparisons using the trace and/or determinant are flawed when there are discrepancies among the number of units receiving a question under each method; therefore, we do not use these criteria in our evaluation.

2.6.2 Metrics from epidemiology

To evaluate whether a set of decision rules (i.e., question asking procedure) is successful in terms of its ability to tailor the survey to the individual respondent, we borrow a series of metrics from epidemiology. These metrics are used to describe the effectiveness of a diagnostic test for determining whether a patient has a particular disease.

Recall that an objective of our methods is to identify events (e.g., purchases of particular items) that sample units are more likely to have and, as a consequence, only ask survey questions pertaining to those events while not asking about events that they unlikely had. In Figure 2.4, we depict the possible outcomes of the question asking procedure and the true state of the individual's event history (e.g., purchase

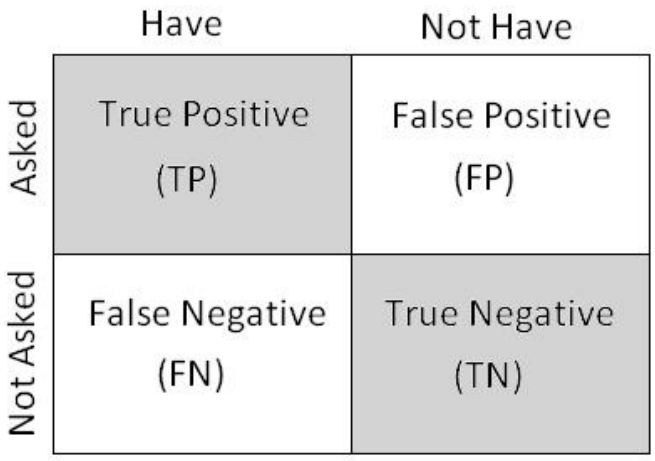

Figure 2.4: Success of decision rule

behavior). The rows correspond to whether or not we ask the question about the event and the columns identify whether or not the sample unit actually had the event. The shaded boxes correspond to the situation when we make the correct decision about asking, denoted as true positive, or not asking, denoted as true negative, the question. The unshaded boxes correspond to the situations when we make an incorrect decision in the question asking procedure. These are false negatives (e.g., we did not ask, but the unit incurred the expense) and false positive (e.g., we asked, but the unit did not incur the expense). False positives and false negatives correspond to Type I and Type II errors, respectively (Gordis, 2000).

The four metrics that are commonly used to judge the efficacy of the testing procedure are: (1) sensitivity; (2) specificity; (3) positive predictive value (PPV); and, (4) negative predictive value (NPV). According to Gordis (2000), these terms are defined as follows. Sensitivity is the proportion of diseased people who are correctly identified as such. Specificity is the proportion of non-diseased people who are correctly identified as negative by the test. In the epidemiological sense,

sensitivity and specificity ask how good the test is at identifying people with and without the disease. In addition, in the clinical setting a different question may be important. Specifically, PPV answers the following question: if the test result is positive in the patient, what is the probability that he has the disease? Finally, NPV answers the question: if the test result is negative, what is the probability that the patient does not have the disease?

In the context of the CE Surveys and a responsive split questionnaire survey, these metrics can be translated as follows. Sensitivity is the proportion of sample members who incurred the expense and were correctly asked about it. Specificity is the proportion of sample members who did not incur the expense and were correctly not asked about it. PPV answers the question: if the procedure recommends asking the survey question to a sample member, what is the probability that he incurred the expense? Finally, NPV answers the question: if the procedure recommends not asking the survey question, what is the probability that the sample member did not incur the expense?

In equations (2.31) – (2.34), we offer formulae for calculating the four metrics[11].

$$\text{Sensitivity} = TP/(TP + FN) \tag{2.31}$$

$$\text{Specificity} = TN/(TN + FP) \tag{2.32}$$

$$PPV = TP/(TP + FP) \tag{2.33}$$

[11] We use unweighted counts of sample units to calculate these metrics.

$$NPV = TN/(TN + FN) \tag{2.34}$$

It is important to highlight some key features of these metrics. First, PPV and NPV have clinical relevance for epidemiologists so these metrics may also have practical significance for survey methodologists. If a diagnostic test for a specific disease was administered to a patient and it was positive, then it would be beneficial to know the probability of actually having the disease, given the positive test result (because the positive test result is the only observable indicator available). When detecting the presence of a disease using a diagnostic test, a positive test result can have significant ramifications for a patient.

For example, a patient may experience added stress, may feel burdened, and/or incur additional costs for follow-up medical evaluations or procedures all as a consequence of the positive test result. Drawing an analogy to the survey setting, the decision to ask a question may (adversely) affect respondent stress, burden, and interviewing costs. The main point, however, is that these two metrics, PPV and NPV, have different implications than sensitivity and specificity; thus, it is essential to keep those distinctions in mind when using these criteria to evaluate the effectiveness of the procedure (in terms of its ability to tailor the survey to the individual respondent).

The second feature is that PPV is affected by two factors: (1) the prevalence of the characteristic in the population tested (equivalently, sampled and interviewed); and, (2) when the prevalence of the characteristic is low, the specificity of the test or procedure used. The association between PPV and prevalence implies that the

results of any procedure must be interpreted in the context of the prevalence of the characteristic in the population being investigated. Said differently, the same procedure can have a very different PPV when it is administered to a high-risk (equivalently, high prevalence) population or to a low-risk (equivalently, low prevalence) population.

To understand how to interpret these metrics and use them for evaluation purposes, consider two hypothetical cases each with 1,000 sample units but with different prevalence rates. In Case 1, the prevalence of the characteristic (e.g., expenditure) is 0.9 (Table 2.5) while in Case 2, the prevalence is 0.25 (Table 2.6). In both cases, we employ a completely random question asking procedure where we essentially "flip a coin" to determine whether or not to ask the question about the expenditure. This implies that the question is asked with probability one-half and not asked with probability one-half. Because the question asking procedure is completely random and does not differentiate between those who incur the expense and those who do not, we would, on average, expect half of each of the samples to get asked the question.

	Have	Not Have	Total
Asked	450	50	500
Not Asked	450	50	500
Total	900	100	1,000

Table 2.5: Case 1: Hypothetical scenario of a completely random question asking procedure when prevalence in the sample is 0.9

For Case 1, we obtain the following calculations: (1) sensitivity is 0.5; (2)

	Have	Not Have	Total
Asked	125	375	500
Not Asked	125	375	500
Total	250	750	1,000

Table 2.6: Case 2: Hypothetical scenario of a completely random question asking procedure when prevalence in the sample is 0.25

specificity is 0.5; (3) PPV is 0.9; and, (4) NPV is 0.1. For Case 2, we obtain the following calculations: (1) sensitivity is 0.5; (2) specificity is 0.5; (3) PPV is 0.25; and, (4) NPV is 0.75. These calculations imply that for a completely random procedure of asking or not asking the question, the sensitivity and specificity are both 0.5 regardless of the prevalence of the expenditure in the sample being investigated. This is because the completely random procedure makes no attempt to differentiate between the sample units who actually incurred the expense and those that did not.

A completely random question asking procedure may serve as an appropriate baseline for comparing against a new set of methods and judging the new methods effectiveness. This is because a completely random procedure does not differentiate between those who have and do not have the expenditure. This non-differentiation can compromise the efficiency of the design. Our hope is that by using information about the characteristics of the sample units in the design, we will better differentiate between those who truly have the expenditure and those that do not thereby improving the efficiency over a completely random question asking procedure.

To understand why attempting to differentiate between the sample members who incurred an expense and those who did not may improve the effectiveness of

a split questionnaire design, consider two additional cases. Each case has similar prevalences as before (0.9 and 0.25, respectively), but now we apply a question asking procedure that better differentiates among those who incur the expense and those who do not (see Tables 2.7 and 2.8). In each case we set the sensitivity and specificity at 0.8. The remaining calculations are as follows. For Case 3, we have a PPV equal to 0.97 and a NPV of 0.31; while for Case 4 the PPV and NPV are 0.57 and 0.92, respectively.

	Have	Not Have	Total
Asked	720	20	740
Not Asked	180	80	260
Total	900	100	1,000

Table 2.7: Case 3: Hypothetical scenario of a differentiating question asking procedure when prevalence in the sample is 0.9

	Have	Not Have	Total
Asked	200	150	350
Not Asked	50	600	650
Total	250	750	1,000

Table 2.8: Case 4: Hypothetical scenario of a differentiating question asking procedure when prevalence in the sample is 0.25

In Tables 2.7 and 2.8, we observe that increases in sensitivity and specificity yield improvements in the PPV and NPV in both cases. We notice, however, that the improvements differ in each case. Increasing the specificity has a greater effect on PPV when the prevalence of the characteristic in the sample being investigated is "low." Specifically, in Case 3, when the prevalence is 0.9, we only achieve about

an 8% improvement in PPV, but in Case 4, when the prevalence is 0.25, we achieve a 128% improvement.

The above hypothetical cases and subsequent calculations provide no guidance on how to select a procedure as the "best" among reasonable alternatives; so given several question asking procedures for the same survey item, one still has to determine which is the preferred procedure to implement. For the case in which one procedure has both higher sensitivity and higher specificity, the choice is easy. In other cases, the distinction is not so clear-cut. This is because one procedure may have a higher sensitivity than another, but performs worse in terms of specificity, and vice-versa. In situations such as these, devising a plan to compare procedures is an essential component of evaluating the success of the responsive design.

Previous research, summarized by Biggerstaff (2000), suggests that the use of positive and negative likelihood ratios, rather than sensitivity and specificity alone, as metrics for diagnostic capabilities has some advantages. Furthermore, these likelihood ratios can be translated into graphical representations that researchers can use to compare how procedures perform relative to each other. Using the same notation as Biggerstaff (2000), we use D to denote the "diseased" population[12], and $\bar{D}$ to denote the "disease-free" population. We let $+$ denote a positive result of the diagnostic test, while $-$ denotes the negative result. Using the same definitions presented earlier in this section, the sensitivity can be expressed as $P[+|D]$ and the specificity can be expressed as $P[-|\bar{D}]$. We also note that the sensitivity can be

[12] "Diseased" population can be interpreted as those individuals who have a particular event (e.g., incur an expense).

referred to as the true-positive rate, while the quantity $1 - P[-|\bar{D}] = P[+|\bar{D}]$ is referred to as the false-positive rate (1-specificity).

Further define $\rho_+ = P[+|D]/P[+|\bar{D}]$ as the likelihood ratio of a positive test and $\rho_- = P[-|D]/P[-|\bar{D}]$ as the likelihood ratio of a negative test for a given diagnostic test[13]. Larger values of ρ_+ suggest greater diagnostic capabilities, while smaller values of ρ_- suggest greater diagnostic capabilities. As we conveyed earlier, PPV and NPV have clinical relevance for epidemiologists and may be of primary interest instead of sensitivity and specificity. Using the notation in the preceding paragraph, we define $P[D|+]$ as the PPV and $P[\bar{D}|-]$ as the NPV. Oftentimes PPV and NPV are not used in practice because they require prior knowledge of the true prevalence of the "disease", denoted as $P[D]$, but when making comparisons between two tests, X and Y, the following equivalences do not require knowledge of $P[D]$.

$$P[D|+_Y] > P[D|+_X] \Longleftrightarrow \rho_+^Y > \rho_+^X \tag{2.35}$$

$$P[\bar{D}|-_Y] > P[\bar{D}|-_X] \Longleftrightarrow \rho_-^Y < \rho_-^X \tag{2.36}$$

In equation (2.35) ρ_+^Y denotes the likelihood of a positive test for test Y and in equation (2.36) ρ_-^Y denotes the likelihood of a negative test for test Y. When condition (2.35) is met, we conclude that test Y outperforms test X for confirming the presence of the disease and when condition (2.36) is met, then Y outperforms X for confirming the absence of the disease. When both conditions are met, then Y is

[13]It is worth noting that Biggerstaff (2000) points out that the term "likelihood" is different than in standard statistical inference.

overall superior to X and when neither condition is met Y is overall inferior to X. The benefit of these equivalences is that they translate the values of sensitivity and specificity for two tests to ρ_+ and ρ_-, which are related to PPV and NPV (the values of clinical significance and of primary interest) through equations (2.35) and (2.36).

Biggerstaff (2000) then relates these quantities to standard receiver operator characteristic (ROC) curve analysis, by plotting the false-positive rate (1-specificity)

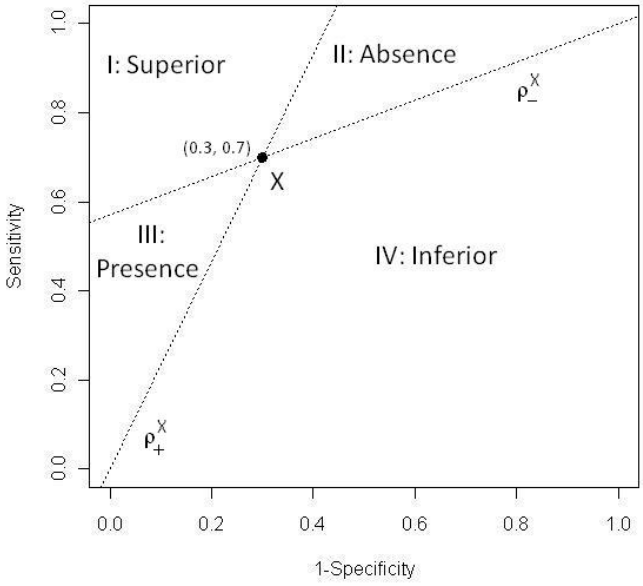

Figure 2.5: Regions of comparison of two diagnostic tests

against the true-positive rate (sensitivity). In Figure 2.5 which we have adapted from Biggerstaff (2000), the false-positive rate and the true-positive rate for a hypothetical test X, with sensitivity and specificity both at 0.7, are plotted against each other. The slope of the dashed line going through point $(1, 1)$ is the negative likelihood ratio, $\rho-$, for test X, while the slope of the dashed line going through point $(0, 0)$ is the positive likelihood ratio, ρ_+, for test X.

Using the regions delineated by the two dashed lines, we have the relationships

between two diagnostic tests, X and Y, in Table 2.9. To fully understand how to

Region	Likelihood ratios	Interpretation
I	$\rho_+^Y > \rho_+^X$ & $\rho_-^Y < \rho_-^X$	Y is superior overall
II	$\rho_+^Y < \rho_+^X$ & $\rho_-^Y < \rho_-^X$	Y is superior for confirming absence of disease
III	$\rho_+^Y > \rho_+^X$ & $\rho_-^Y > \rho_-^X$	Y is superior for confirming presence of disease
IV	$\rho_+^Y < \rho_+^X$ & $\rho_-^Y > \rho_-^X$	Y is inferior overall

Table 2.9: Comparison of two diagnostic tests, Y to X (from Biggerstaff [2000])

interpret these equivalences and the regions delineated on the graph in Figure 2.5, we present two hypothetical cases of comparing two diagnostic tests A to B in Figure 2.6. In Case 1, we see that test A is inferior overall to test B because it falls

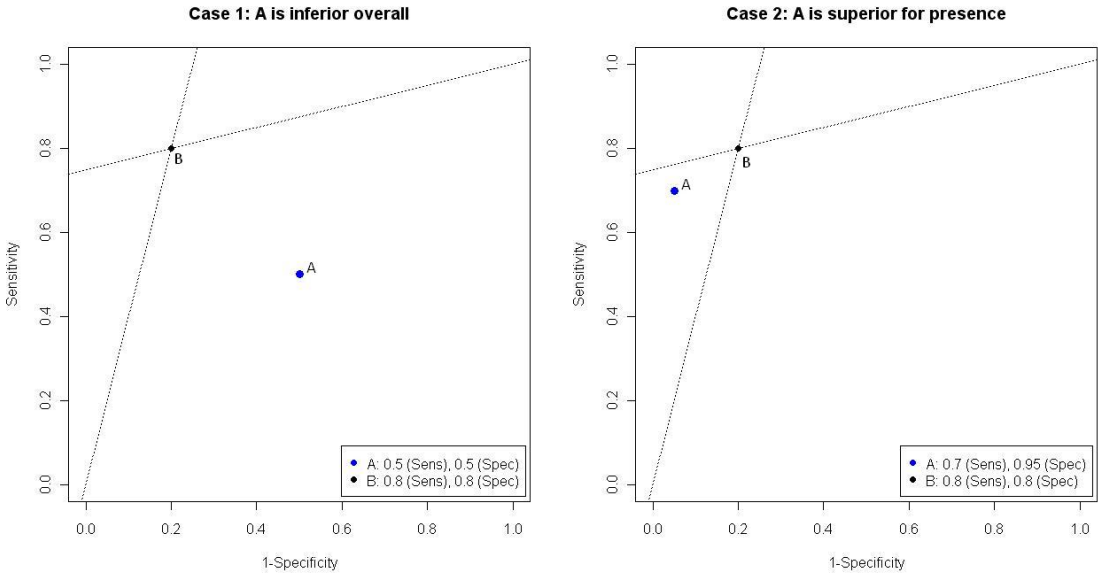

Figure 2.6: Hypothetical cases of comparing diagnostic tests

into Region IV. Test A has both lower sensitivity and specificity values than test B. On the other hand, in Case 2, we see that test A falls into Region III, so test A is superior to test B for confirming the presence of the disease. Thus, we see an explicit tradeoff between a lower sensitivity with a higher specificity.

To recapitulate, in this section, we presented a series of metrics commonly used in epidemiology to describe the effectiveness of a diagnostic test for determining whether a patient has a particular disease. These metrics will enable us to evaluate whether a set of decision rules is successful in terms of its ability to tailor the survey to the individual respondent. We also presented a graphical tool based on functions of these metrics that facilitates simultaneous comparisons among various sets of decision rules, which permits us to select a preferred method of asking questions to individual sample units in a responsive split questionnaire.

Chapter 3

Preliminary analyses

3.1 Overview

In this chapter we detail specific features of the survey of application and present the results of preliminary analyses, including descriptive statistics, analytic statistics on the relationships among expenditures and demographic characteristics across two interviews, and a simulation study extending previous research. These preliminary analyses provide empirical evidence that justify extending the current set of split questionnaire methods and help us identify areas of the research problem that warrant particular focus.

3.2 Exploration of Consumer Expenditure Survey data

3.2.1 Description of the Consumer Expenditure Survey

The Consumer Expenditure Survey Program consists of two national household surveys, the Quarterly Interview Survey (CE) and the Diary Survey[1]. Together these surveys provide information on the spending habits of consumers in the United States by collecting data on their expenditures, income, and household characteristics. The data collected in the surveys provide the basis for revising the cost weights

[1]When we use the acronym CE, we are referring to the Quarterly Interview Survey.

and associated pricing samples for the Consumer Price Index (CPI), one of the nation's leading economic indicators. The surveys are also conducted to meet the need for timely and detailed information on the spending patterns of different types of families (BLS *Handbook of Methods*, 2007). The CE is the focus of the research presented in this dissertation.

The current CE is an ongoing rotating panel survey of U.S. households in which, for each interview, all sample units are generally administered the same survey questionnaire. CE respondents are interviewed five times over a period of 13 months. The first interview is a bounding interview, used to reduce telescoping errors in the second interview. It also collects inventory information on large items, such as vehicles and mortgages for the household[2]. Each of the second through fifth interviews is a bounded interview[3], provided there is no prior interview unit nonresponse. In these interviews, each respondent is generally asked questions on a common set of expenditures. These expenditures are those that respondents can be expected to recall for a period of three months or longer and tend to include relatively large purchases, such as for property, automobiles, and large durable goods, and regularly occurring purchases, such as those for utilities and insurance premiums.

The structure of the current CE questionnaire is such that the survey has 22 main sections with some having explicit subsections, i.e., major expenditure categories are broken down into subcategories. For instance, Section Nine (clothing)

[2]In the inventory process, explicit questions are administered to a sample unit about whether it has refrigerator and other large durable goods inside the housing unit.

[3]In a bounded interview, the interviewer reviews the respondent's responses to the prior interview. This procedure is thought to aid recall and reduce the chance that respondents will report the same event in the current interview.

is divided into four subcategories: (1) clothing for persons age two and older; (2) infants' clothing, watches, jewelry, and hairpieces; (3) clothing services; and, (4) sewing materials. The sections and subsections are detailed in their entirety in Appendix A.1. Sections Two to 20 collect information on various types of expenditure categories, while the remaining sections collect information on the demographic characteristics of the consumer unit or household, as well as credit liability and income. This structure lends itself to exploring the use of split questionnaire methods. This is because each section solicits information on expenditures within a specific category. These expenditures tend to be logically related and are thought to require similar cognitive processes to retrieve the encoded information required by the battery of questions.

There are a few differences among the interviews administered across the survey panel. As previously mentioned, the first interview is primarily used for bounding and inventory purposes, but this initial interview also differs in two additional ways. First, the interview employs a one-month recall period, as opposed to a three-month recall period utilized in the second through fifth interviews. The implication of this is that any mean expenditure estimates produced using first interview data will be average *monthly* expenditures. If expenditure estimates are produced using expenditure data from the second interview, then the estimates will be average *quarterly* expenditures.

The second difference is that survey items pertaining to a few expenditure categories are not administered during the first interview. These categories are identified in the last column of the table presented in Appendix A.2. We only con-

sider responsive split questionnaire methods for items collected in both interviews, but we discuss how our methods can be extended to include items that are collected in the second interview but not the first (see Chapter 5).

3.2.2 CE analysis file creation

We constructed an analysis file from previously collected survey data using the full CE questionnaire. Due to concerns over presenting results from previously unreleased economic data during the time our research was in progress, we chose data collected between January, 2008 and December, 2009 for our analysis file. All data in this time frame were (and are currently) available in the public domain.

The current CE production systems are structured such that information collected from each expenditure section (or sub-section) is contained in a separate data file. Furthermore, a sample unit is only contained in the data file associated with the expenditure section if it reported an expense when that section was administered. We summarized the reported expenditures within a (sub-)section by summing across the expenditure variables for the (sub-)section and then merged the distinct expenditure data files across sample units – to have one data file that contained all the derived summary expenditure variables. We detail the mapping of the CE (sub-)sections to the derived expenditure variables in Appendix A.2.

In addition to expenditure information being contained in distinct data files, demographic and timing information (i.e., the amount of time it took for each section to be administered) on sample units are also contained in separate data files.

Therefore, we merged this additional information onto our final analysis file. Appendix A.3 details the demographic characteristics we use throughout this research. We chose these demographic characteristics for two reasons. First, they are commonly thought to be associated with purchase behavior and second, some of these demographic characteristics are used in the current nonresponse adjustment (BLS *Handbook of Methods*, 2007).

A number of data cleaning procedures were performed prior to conducting the preliminary analyses and developing the responsive split questionnaire methods. First, we restricted our analysis file to only sample units that were respondents in both the first and second interviews. This subset was created because (1) we needed expenditure and other demographic information from the first interview (these data comprise the prior information on the sample unit) and (2) we desired to produce expenditure estimates from information collected during the second interview. Thus, we had to ensure that sample units were not missing due to unit nonresponse. If there was unit nonresponse, then the required expenditure and demographic information would not have been collected. Restricting the data to these sample cases yielded $10,495$ sample units. These $10,495$ sample units became the sample units for which we conducted the preliminary analyses and developed the responsive split questionnaire methods for the second interview.

The next series of data cleaning procedures dealt with extreme observations. When CE data are received from the data collection institution, the U.S. Census Bureau, there are a series of edits that the CE program office performs; however, the current CE processing systems are designed such that expenditure information

collected during the first interview is not subjected to the same extensive editing processes that expenditure information collected during the second through fifth interviews are. This is partly because first interview data are not used in any published or official estimates produced by the CE program. Therefore, we performed some editing procedures to guard against the possibility of extreme observations and other (unplanned) data anomalies affecting our results.

To deal with extreme observations, we examined the 97.5^{th} percentiles of each derived summary expenditure variable, excluding non-reports. "Excluding non-reports" means that if a respondent reported a zero-dollar expense for a particular expenditure category then that respondent was excluded from the calculation of the desired statistic. The 97.5^{th} percentiles for all derived summary expenditure variables, separately for the two interviews, are presented in Appendix B (see Table B.1).

The data cleaning procedure for expenditure information was performed as follows: if a respondent reported an expenditure value as extreme or more extreme than the 97.5^{th} percentile, then that respondent's expenditure value was top-coded to the value of the 97.5^{th} percentile for that expenditure variable. We provide summary statistics for expenditure information for the first and second interviews, before top-coding, in Appendix B (see Tables B.2 and B.3). Since these are only interim results, we do not describe any findings from these statistics. However, we summarize similar statistics for expenditure information for the first and second interviews, after top-coding, in Section 3.2.3.3 since those results are directly relevant to the methods we develop (see Tables 3.4 – 3.7).

3.2.3 Descriptive statistics for CE analysis file

In this section we present descriptive statistics for the CE analysis file. These will be useful when interpreting the results of the preliminary analyses as well as the primary outcomes of the dissertation research. We present a summary of the demographic characteristics (Section 3.2.3.1), timing information (Section 3.2.3.2), and expenditure information (Section 3.2.3.3). We also highlight the implications of these descriptive statistics for our dissertation research (Section 3.2.3.4). As a final note, unless otherwise stated, all descriptive statistics in this section are unweighted. We chose to conduct unweighted analyses because our methods assume a fixed initial data gathering effort (i.e., first interview), then based on these data, we design the responsive split questionnaire for the second interview. Furthermore, the goals of this research essentially deal with questionnaire design and do not, per se, deal with traditional complex sampling issues (e.g., complex sample design).

3.2.3.1 Demographic characteristics

As we stated earlier, certain demographic characteristics are often thought to be associated with expenditures. For instance, a CU containing many persons would likely have a higher amount of total quarterly expenditures than a single person CU. Because of their association with expenditures, demographic characteristics are also used in the CE nonresponse adjustment. Specifically, CU size, household tenure, race, and region are used in the current CE nonresponse adjustment, while urbanicity is used in the calibration adjustment.

In Table 3.1 we present descriptive statistics for our set of six demographic characteristics contained in the analysis file. We observe that sample units in our analysis file tend to live in areas with poverty of less than 20% (85.2%), live in urban areas (80.5%), own (or are buying) their homes (69.1%), are non-black (88.8%), come from 1- or 2-person CUs (60.7%), and a plurality (31.6%) of the sample units are in the South (i.e., regional offices of Atlanta, Charlotte, and Dallas). Although the descriptive statistics alone do not say anything about their relationship to expenditures or purchase behavior, they are an important component to understanding the final results of our methods. In Section 3.3.1.3, we verify that some combination of these demographic characteristics are, in fact, related to purchasing various items, so they should prove useful in developing responsive split questionnaire methods.

3.2.3.2 Timing information

In this section, we summarize the section-level timing information[4] for the first (Table 3.2) and second (Table 3.3) interviews. This summary assists us in evaluating the potential for reducing burden (when burden is measured as interview length, in minutes) by implementing various responsive split questionnaire methods.

On average the first interview takes about 59 minutes to complete while the second interview takes about 62 minutes to complete. Of the expenditure sections administered in the first interview, Section Three (owned living quarters and other owned real estate) and Section Four (utilities and fuels for owned and rented proper-

[4]This timing information is captured as a by-product of the Computer-Assisted Personal Interviewing (CAPI) instrument. These data are not manually entered by the interviewer. They reflect the amount of time the interviewer is entering expenditure information into the section.

Characteristic	Level	Frequency	Percent
Poverty	20% or more	1,551	14.78
	Less than 20%	8,944	85.22
Regional office	Boston	704	6.71
	New York	608	5.79
	Philadelphia	907	8.64
	Detroit	845	8.05
	Chicago	1,001	9.54
	Kansas City	492	4.69
	Seattle	936	8.92
	Charlotte	1,177	11.21
	Atlanta	1,038	9.89
	Dallas	1,097	10.45
	Denver	817	7.78
	Los Angeles	873	8.32
Urbanicity	Urban	8,443	80.45
	Rural	2,052	19.55
Household tenure	Owner	7,255	69.13
	Renter	3,240	30.87
Race	Black	1,167	11.12
	Non-black	9,328	88.88
CU size	1 CU member	2,884	27.48
	2 CU members	3,486	33.22
	3 or 4 CU members	3,012	28.70
	5+ CU members	1,113	10.61
Total CUs		10,495	100.00

Table 3.1: Unweighted descriptive statistics for the demographic characteristics

Section	Mean (minutes)	Std Dev	Min	Q1	Median	Q3	95^{th} Pctl	Max
1	3.61	2.10	0.00	2.50	3.20	4.18	6.63	35.40
2	0.46	1.00	0.00	0.05	0.08	0.63	1.73	24.92
3	5.22	5.30	0.00	0.77	4.17	7.43	14.83	85.67
4	5.21	3.46	0.00	3.07	4.42	6.38	11.28	42.87
5	1.36	1.87	0.00	0.38	0.72	1.58	4.50	31.60
6	2.13	2.16	0.00	0.83	1.52	2.65	6.00	30.85
7	0.49	0.68	0.00	0.17	0.30	0.53	1.47	16.08
8	1.56	1.76	0.00	0.63	1.05	1.85	4.25	32.50
9	2.93	3.01	0.00	1.05	2.07	3.72	8.38	40.65
10	0.34	0.85	0.00	0.08	0.13	0.22	1.53	20.27
11	4.07	3.56	0.00	1.75	3.28	5.38	10.30	85.13
12	2.05	1.85	0.00	0.95	1.58	2.55	5.18	37.68
13	2.66	2.52	0.00	1.12	2.07	3.48	6.97	44.25
14	2.16	2.27	0.00	0.68	1.68	2.87	5.92	38.78
15	1.51	1.85	0.00	0.43	0.97	1.87	4.63	24.05
16	0.74	1.27	0.00	0.18	0.33	0.78	2.67	40.55
17	1.00	1.22	0.00	0.33	0.65	1.23	2.88	20.98
18	0.55	0.86	0.00	0.12	0.27	0.67	1.82	18.75
19	4.05	3.69	0.00	1.95	3.12	4.92	10.32	96.93
20	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
21	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
22	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Back	4.78	5.09	0.15	1.30	2.77	6.33	16.37	55.05
Front	6.62	10.07	0.12	1.18	2.65	7.80	25.43	188.10
Coverage	0.86	1.58	0.00	0.23	0.40	0.85	2.88	26.60
Control	4.50	3.90	0.00	2.27	3.43	5.30	11.33	61.02
Total	58.86	27.03	5.12	40.02	54.20	72.30	107.67	372.60

Table 3.2: Unweighted descriptive statistics for the first interview timing information

ties) take, on average, the longest to administer with each lasting over five minutes. For the second interview, of the expenditure sections administered, Section Four takes, on average, the longest to administer (about six minutes). This is reasonable not only because a majority of the sample units in the analysis file have these expenditures, but also because many sample units use records (e.g., billing statements) to report these expenditures. Consulting records for the correct information to report adds length to the interview.

For the remaining sections in the first interview, the average completion time is under two minutes, while the average completion time for the remaining sections in the second interview is under two and a half minutes. A vast majority of the respondents finish each section in under ten minutes, but there are some instances when the section completion time is greater than 60 minutes. In Table 3.2, we also observe zeros for every entry in each of the rows corresponding to Sections 20 – 22. This is because these sections are not asked in the first interview. As a final note, the expenditure sections consisting of expenditure variables that we consider for our responsive split questionnaire methods (these are Sections Two through 19) constitute about 39 (out of 62) minutes of total interview time.

There are seven additional non-expenditure sections in the current CE survey. They are: (1) general survey information (Section One); (2) credit liability (Section 21); (3) work experience and income (Section 22 – only collected in the second and fifth interviews); (4) back; (5) front; (6) coverage; and, (7) control. These sections collect information pertaining to household demographic characteristics as well as the contact attempts made to the sample unit during the data collection phase.

Section	Mean (minutes)	Std Dev	Min	Q1	Median	Q3	95^{th} Pctl	Max
1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	0.61	2.03	0.00	0.00	0.00	0.70	2.52	44.75
3	2.82	4.05	0.02	0.23	1.72	3.50	9.78	82.43
4	6.07	4.65	0.15	3.17	4.85	7.43	14.75	70.13
5	1.61	2.24	0.07	0.37	0.73	1.97	5.63	32.82
6	2.50	2.68	0.10	0.83	1.70	3.18	7.35	38.62
7	0.58	1.03	0.02	0.17	0.30	0.60	1.87	31.23
8	1.80	2.08	0.10	0.63	1.17	2.17	5.28	33.73
9	3.60	3.77	0.12	1.25	2.47	4.53	10.78	57.97
10	0.47	0.83	0.05	0.18	0.28	0.47	1.35	19.75
11	0.64	1.22	0.00	0.17	0.27	0.52	2.70	23.85
12	3.33	2.84	0.13	1.47	2.62	4.27	8.60	46.43
13	1.75	1.97	0.02	0.62	1.22	2.20	5.17	35.17
14	1.60	1.96	0.05	0.52	1.00	1.97	4.87	40.73
15	2.03	2.57	0.03	0.50	1.27	2.57	6.45	60.95
16	0.84	1.36	0.03	0.18	0.35	0.93	3.18	25.78
17	2.76	2.47	0.10	1.10	2.15	3.65	7.27	31.87
18	2.26	3.71	0.00	0.20	0.58	3.08	9.38	47.75
19	3.97	3.45	0.17	1.82	3.08	4.98	10.02	84.08
20	3.60	2.73	0.17	1.88	3.02	4.55	8.33	40.33
21	1.00	1.40	0.00	0.20	0.58	1.30	3.18	32.75
22	8.38	5.81	0.00	4.33	6.97	10.87	19.68	51.60
Back	3.79	4.66	0.20	1.02	1.85	4.48	14.45	59.60
Front	4.89	7.57	0.13	0.83	1.97	5.62	19.00	120.25
Coverage	0.17	0.74	0.00	0.02	0.03	0.10	0.60	20.42
Control	1.36	2.56	0.08	0.33	0.57	1.20	5.03	39.72
Total	62.43	31.70	6.33	40.48	56.50	77.67	121.72	285.62

Table 3.3: Unweighted descriptive statistics for the second interview timing information

For the purposes of our research, we will not consider responsive split questionnaire methods for these sections since our primary focus is administering questions about expenditure information. At any rate, the non-expenditure sections contain the information that we would likely collect from every sample member.

3.2.3.3 Expenditure information

In this section, we present a series of tables providing summaries of descriptive statistics for the expenditure information collected in the first and second interviews, after top-coding, when non-reports are treated as zeros as well as when zeros are excluded from the calculation (see Tables 3.4, 3.5, 3.6, and 3.7).

Both including and excluding non-reports from the calculations represent quantities of interest for the Consumer Expenditure Survey Program. In Section 2.3.1, we briefly noted that the former has an *unconditional* interpretation while the latter has a conditional interpretation and can be viewed as a domain quantity (where domain is defined by those sample units who purchase the particular item). The *unconditional* mean is defined as the average expense incurred on category k among all sample units whereas the *conditional* mean is defined as the average amount spent on item k for those sample units who incurred the expense. In general, the conditional mean will always be larger than the unconditional mean because the unconditional mean includes many zero-dollar reports.

In all instances, the (unconditional and conditional) mean expenditure estimates, unweighted and weighted, vary across expenditure category and, in general,

the standard deviations tend to be large relative to the corresponding estimate. For first interview expenditure information, after top-coding when non-reports are treated as zeros, the weighted CVs range from about 0.75 to 23.70. The highest CVs correspond to the rarest reported expenditure categories. These are TX4A_4 (modem purchase, Smartphone apps, ringtones) and TX5A_2 (Construction materials for general jobs) with CVs of 23.70 and 21.23, respectively. The lowest CVs correspond to the most prevalent expenditure categories. These are TX4A_1 (residential and mobile telephone services) and TX4D (utilities) with CVs of 0.75 and 0.76, respectively. When non-reports are excluded from the calculation, the weighted CVs are much lower and less variable. They range from 0.37, TX4A_2 (Internet access), to 2.36, TX5B (contractor labor, materials, and tools). These values, however, are still considered relatively high.

Similar trends are observed for the expenditure information collected in the second interview. When non-reports are treated as zeros, TX4A_4 (modem purchase, Smartphone apps, ringtones) has the highest weighted CV, 29.50, while TX4D (utilities) has the lowest, 0.68. When we exclude non-reports, the weighted CVs range from 0.36, TX4A_2 (Internet access), to 1.81, TX16A (educational expenses for tuition, recreational lessons, etc.).

Expenditure	Mean	Std Dev	Weighted Mean	Weighted Std Dev	Weighted CV	Min	Q1	Median	Q3	95th Pctl	Max
TX2	211.45	424.82	199.20	406.65	2.04	0.00	0.00	0.00	242.00	1,130.00	2,250.00
TX3F	68.33	422.11	64.58	404.29	6.26	0.00	0.00	0.00	0.00	0.00	6,000.00
TX3H	33.03	203.72	32.20	200.80	6.24	0.00	0.00	0.00	0.00	150.00	3,200.00
TX4A_1	114.21	84.98	112.29	84.09	0.75	0.00	50.00	100.00	160.00	290.00	346.00
TX4A_2	9.93	17.10	9.74	17.00	1.74	0.00	0.00	0.00	20.00	45.00	70.00
TX4A_3	7.77	22.29	7.50	21.87	2.91	0.00	0.00	0.00	0.00	60.00	130.00
TX4A_4	0.10	2.27	0.09	2.18	23.70	0.00	0.00	0.00	0.00	0.00	110.00
TX4B	2.44	12.45	2.43	12.57	5.17	0.00	0.00	0.00	0.00	15.00	150.00
TX4C	50.00	49.92	49.83	49.53	0.99	0.00	0.00	49.00	84.00	146.00	180.00
TX4D	224.93	171.68	221.78	167.95	0.76	0.00	100.00	196.00	309.00	578.00	750.00
TX5A_1	9.63	116.42	9.08	109.36	12.05	0.00	0.00	0.00	0.00	0.00	3,000.00
TX5A_2	1.36	28.32	1.27	27.03	21.23	0.00	0.00	0.00	0.00	0.00	1,250.00
TX5B	259.11	1,806.62	248.88	1,725.09	6.93	0.00	0.00	0.00	0.00	860.00	50,000.00
TX6A	28.39	194.68	27.85	191.38	6.87	0.00	0.00	0.00	0.00	0.00	2,900.00
TX6B	105.08	301.35	103.08	296.91	2.88	0.00	0.00	0.00	49.00	601.00	2,047.00
TX7A	9.02	51.72	8.85	50.68	5.73	0.00	0.00	0.00	0.00	36.00	700.00
TX8	83.81	320.68	82.27	315.25	3.83	0.00	0.00	0.00	15.00	428.00	2,604.00
TX9A	118.14	202.65	114.32	196.71	1.72	0.00	0.00	22.00	150.00	520.00	1,026.00
TX9B	18.45	73.67	18.12	72.65	4.01	0.00	0.00	0.00	0.00	100.00	750.00
TX9C	1.78	12.51	1.78	12.80	7.21	0.00	0.00	0.00	0.00	0.00	200.00
TX9D	2.08	15.15	2.11	15.20	7.20	0.00	0.00	0.00	0.00	0.00	230.00
TX10C_1	18.67	112.05	17.76	108.79	6.12	0.00	0.00	0.00	0.00	0.00	1,571.00
TX10C_23	42.81	459.21	40.80	447.79	10.98	0.00	0.00	0.00	0.00	0.00	12,000.00
TX11B	650.40	2,156.71	632.17	2,100.70	3.32	0.00	0.00	0.00	0.00	5,000.00	18,000.00
TX12A	91.61	240.43	91.02	238.54	2.62	0.00	0.00	0.00	40.00	600.00	1,500.00
TX12B	13.89	47.01	13.83	47.07	3.40	0.00	0.00	0.00	0.00	90.00	384.00
TX12C_1TO9	0.23	2.82	0.22	2.74	12.29	0.00	0.00	0.00	0.00	0.00	80.00
TX12C_10	194.36	168.36	193.02	166.92	0.86	0.00	70.00	150.00	290.00	560.00	700.00
TX13B	190.54	361.15	187.40	356.66	1.90	0.00	0.00	30.00	215.00	910.00	2,000.00
TX14B	52.33	186.68	52.55	187.28	3.56	0.00	0.00	0.00	0.00	300.00	1,700.00
TX15A	113.72	264.40	113.51	262.76	2.31	0.00	0.00	0.00	100.00	580.00	1,640.00
TX16A	131.16	594.08	125.95	580.37	4.61	0.00	0.00	0.00	0.00	624.00	6,100.00
TX17A	23.52	82.93	22.82	81.91	3.59	0.00	0.00	0.00	4.00	110.00	760.00
TX18A	6.66	92.71	6.73	95.18	14.13	0.00	0.00	0.00	0.00	0.00	3,000.00
TX19A	133.43	278.47	129.78	271.52	2.09	0.00	0.00	25.00	130.00	649.00	1,645.00
TX19B	201.81	421.39	201.48	420.25	2.09	0.00	0.00	1.00	200.00	1,200.00	2,200.00

Table 3.4: Descriptive statistics for the first interview expenditure information after top-coding, non-reports treated as zeros (N=10,495)

Expenditure	Number Reporting	Mean	Std Dev	Weighted Mean	Weighted Std Dev	Weighted CV	Min	Q1	Median	Q3	95th Pctl	Max
TX2	2,921	759.72	481.55	733.60	465.79	0.63	1.00	414.00	660.00	995.00	1,750.00	2,250.00
TX3F	473	1,516.21	1,327.20	1,455.73	1,288.05	0.88	1.00	680.00	1,104.00	1,900.00	4,788.00	6,000.00
TX3H	604	573.87	641.35	559.05	636.80	1.14	1.00	200.00	366.00	600.00	2,000.00	3,200.00
TX4A_1	9,606	124.78	81.06	123.01	80.17	0.65	2.00	60.00	105.00	170.00	300.00	346.00
TX4A_2	3,023	34.49	12.99	34.56	12.95	0.37	1.00	25.00	33.00	40.00	60.00	70.00
TX4A_3	1,364	59.78	26.72	59.59	26.36	0.44	1.00	40.00	54.00	75.00	119.00	130.00
TX4A_4	42	23.98	27.02	23.71	25.87	1.09	2.00	5.00	14.00	30.00	83.00	110.00
TX4B	728	35.13	32.99	35.58	33.66	0.95	1.00	15.00	25.00	45.00	100.00	150.00
TX4C	6,634	79.10	40.50	78.46	40.21	0.51	1.00	50.00	71.50	102.00	160.00	180.00
TX4D	9,619	245.42	164.71	241.26	161.20	0.67	1.00	126.00	210.00	324.00	597.00	750.00
TX5A_1	254	397.76	638.14	377.83	598.81	1.58	1.00	50.00	143.50	450.00	2,000.00	3,000.00
TX5A_2	94	151.52	259.83	142.04	248.07	1.75	3.00	30.00	65.00	150.00	1,000.00	1,250.00
TX5B	1,374	1,979.17	4,641.05	1,857.24	4,384.16	2.36	1.00	150.00	491.50	1,900.00	8,000.00	50,000.00
TX6A	444	670.96	682.40	660.27	671.31	1.02	5.00	150.00	450.00	950.00	2,270.00	2,900.00
TX6B	3,466	318.18	455.19	311.97	449.04	1.44	1.00	50.00	131.50	370.00	1,450.00	2,047.00
TX7A	686	137.99	152.15	131.09	148.41	1.13	1.00	40.00	80.00	175.00	500.00	700.00
TX8	2,948	298.37	549.68	292.34	540.11	1.85	1.00	30.00	80.00	254.00	1,600.00	2,604.00
TX9A	5,719	216.79	232.33	209.97	225.79	1.08	1.00	58.00	135.00	300.00	770.00	1,026.00
TX9B	1,764	109.77	149.26	108.20	147.52	1.36	1.00	26.00	52.00	120.00	400.00	750.00
TX9C	444	41.96	44.92	43.40	46.87	1.08	1.00	12.00	24.00	50.00	150.00	200.00
TX9D	473	46.25	55.30	46.74	55.05	1.18	1.00	10.00	25.00	56.00	200.00	230.00
TX10C_1	391	501.18	308.83	495.36	305.76	0.62	99.00	309.00	400.00	586.00	1,300.00	1,571.00
TX10C_23	158	2,843.53	2,465.78	2,808.07	2,455.65	0.87	142.00	1,200.00	2,000.00	3,500.00	10,000.00	12,000.00
TX11B	1,781	3,832.62	3,901.20	3,697.04	3,804.83	1.03	1.00	1,000.00	2,500.00	5,000.00	12,000.00	18,000.00
TX12A	3,700	259.86	346.80	257.15	343.56	1.34	2.00	35.00	100.00	350.00	1,058.50	1,500.00
TX12B	1,591	91.65	86.33	91.22	86.95	0.95	1.00	32.00	65.00	115.00	300.00	384.00
TX12C_1TO9	130	18.49	17.50	18.34	16.95	0.92	1.00	6.00	12.00	25.00	50.00	80.00
TX12C_10	9,358	217.97	163.23	215.52	162.05	0.75	1.00	100.00	180.00	300.00	600.00	700.00
TX13B	5,533	361.41	430.87	356.00	426.18	1.20	1.00	101.00	200.00	418.00	1,400.00	2,000.00
TX14B	2,081	263.93	346.34	262.07	346.45	1.32	1.00	60.00	145.00	301.00	1,000.00	1,700.00
TX15A	5,052	236.23	341.02	233.48	337.64	1.45	1.00	40.00	100.50	270.00	1,000.00	1,640.00
TX16A	1,898	725.26	1,233.40	707.80	1,216.96	1.72	1.00	100.00	281.00	705.00	3,456.00	6,100.00
TX17A	2,671	92.41	143.75	91.56	143.62	1.57	1.00	25.00	45.00	91.00	384.00	760.00
TX18A	202	346.09	575.05	359.77	597.51	1.66	2.00	50.00	150.00	350.00	1,200.00	3,000.00
TX19A	6,276	223.12	331.16	216.61	322.86	1.49	1.00	40.00	100.00	250.00	907.00	1,645.00
TX19B	5,252	403.28	523.07	401.71	521.24	1.30	1.00	50.00	200.00	530.00	1,650.00	2,200.00

Table 3.5: Descriptive statistics for the first interview expenditure information after top-coding, excluding non-reports (N=10,495)

It is worth pointing out an important relationship between S_k^2, the element variance of characteristic k (e.g., incurring expense k) in the population, and S_{dk}^2, the element variance of the characteristic among the units in domain d. We use the subscript dk to denote the domain defined by members of the population incurring expense k. In other words, S_k^2 is calculated from everyone in the population, i.e., both those who incurred the expense and those that did not, while S_{dk}^2 is calculated by excluding the zero expenditure reports. Furthermore, if we let $\bar{y}_{Udk}$ denote the population mean of characteristic k for members in the domain, P_{dk} be the proportion of the units in the population that are in the domain (i.e., the prevalence of incurring expense k), and $Q_{dk} = 1 - P_{dk}$, then we can rewrite S_k^2 as follows.

$$S_k^2 = P_{dk}(S_{dk}^2 + Q_{dk}\bar{y}_{Udk}^2) \tag{3.1}$$

If we compare the columns containing the standard deviations in the two tables for the second interview (Tables 3.6 and 3.7), then we will observe four instances when $S_k^2 > S_{dk}^2$. These are TX4A_1 (telephone services), TX4C (cable/satellite not reported), TX4D (utilities), and TX12C_10 (average monthly gas expense). It is important to identify under what condition this holds because this may assist in interpreting the results. If we let $\text{relvar}(\bar{y}_{Udk}) = \bar{y}_{Udk}^2/S_{dk}^2$, then we have the

following.

$$
\begin{aligned}
S_k^2 &= P_{dk}(S_{dk}^2 + Q_{dk}\bar{y}_{Udk}^2) \\
\implies \frac{S_k^2}{S_{dk}^2} &= P_{dk}(1 + \frac{Q_{dk}\bar{y}_{Udk}^2}{S_{dk}^2}) >^{?} 1 \\
\implies 1 + \frac{Q_{dk}\bar{y}_{Udk}^2}{S_{dk}^2} &> \frac{1}{P_{dk}} \\
\implies 1 + \frac{Q_{dk}}{\text{relvar}(\bar{y}_{Udk})} &> \frac{1}{P_{dk}} \\
\implies \frac{Q_{dk}}{\text{relvar}(\bar{y}_{Udk})} &> \frac{1 - P_{dk}}{P_{dk}} \\
\implies \frac{Q_{dk}}{\text{relvar}(\bar{y}_{Udk})} &> \frac{Q_{dk}}{P_{dk}} \\
\implies \frac{1}{\text{relvar}(\bar{y}_{Udk})} &> \frac{1}{P_{dk}} \\
\implies \text{relvar}(\bar{y}_{Udk}) &< P_{dk} \qquad (3.2)
\end{aligned}
$$

Thus, we conclude that $S_k^2 > S_{dk}^2$, if $\text{relvar}(\bar{y}_{Udk}) < P_{dk}$. This is exactly what we observe in the second interview for those four expenditure categories.

Expenditure	Mean	Std Dev	Weighted Mean	Weighted Std Dev	Weighted CV	Min	Q1	Median	Q3	95th Pctl	Max
TX2	628.23	1,244.80	590.44	1,190.09	2.02	0.00	0.00	0.00	750.00	3,300.00	6,480.00
TX3F	216.01	1,361.43	206.73	1,314.60	6.36	0.00	0.00	0.00	0.00	0.00	19,869.00
TX3H	104.13	680.24	102.67	673.86	6.56	0.00	0.00	0.00	0.00	375.00	12,000.00
TX4A_1	350.07	252.53	343.79	249.73	0.73	0.00	153.00	300.00	497.00	867.00	1,020.00
TX4A_2	27.55	49.51	27.00	49.11	1.82	0.00	0.00	0.00	57.00	135.00	195.00
TX4A_3	23.84	67.75	23.01	66.65	2.90	0.00	0.00	0.00	0.00	195.00	375.00
TX4A_4	0.11	3.07	0.10	2.95	29.50	0.00	0.00	0.00	0.00	0.00	140.00
TX4B	4.96	22.88	4.88	22.54	4.62	0.00	0.00	0.00	0.00	30.00	220.00
TX4C	61.71	70.06	61.16	68.71	1.12	0.00	0.00	50.00	95.00	199.00	327.00
TX4D	608.26	417.33	603.10	409.18	0.68	0.00	300.00	558.00	844.00	1,420.00	1,781.00
TX5A_1	6.35	87.60	6.14	86.76	14.14	0.00	0.00	0.00	0.00	0.00	2,500.00
TX5A_2	1.42	21.87	1.31	20.78	15.83	0.00	0.00	0.00	0.00	0.00	700.00
TX5B	358.39	1,636.40	347.94	1,597.47	4.59	0.00	0.00	0.00	0.00	1,771.00	16,054.00
TX6A	38.61	228.79	38.51	225.09	5.85	0.00	0.00	0.00	0.00	100.00	2,947.00
TX6B	183.68	428.19	183.59	426.76	2.32	0.00	0.00	0.00	142.00	1,040.00	2,500.00
TX7A	16.33	70.40	16.72	70.64	4.23	0.00	0.00	0.00	0.00	105.00	709.00
TX8	131.34	416.60	129.55	411.95	3.18	0.00	0.00	0.00	50.00	705.00	2,880.00
TX9A	202.98	296.95	197.45	290.52	1.47	0.00	0.00	84.00	285.00	830.00	1,443.00
TX9B	36.16	120.83	35.62	120.12	3.37	0.00	0.00	0.00	0.00	212.00	1,042.00
TX9C	3.90	24.46	3.76	24.03	6.38	0.00	0.00	0.00	0.00	15.00	379.00
TX9D	3.90	25.87	3.99	25.95	6.50	0.00	0.00	0.00	0.00	10.00	374.00
TX10C_1	20.11	116.65	19.17	113.64	5.93	0.00	0.00	0.00	0.00	0.00	1,571.00
TX10C_23	49.92	525.88	47.63	516.01	10.83	0.00	0.00	0.00	0.00	0.00	14,000.00
TX11B	693.33	2,232.33	674.63	2,177.24	3.23	0.00	0.00	0.00	0.00	5,000.00	18,000.00
TX12A	165.83	344.30	163.18	338.60	2.08	0.00	0.00	25.00	150.00	879.00	1,900.00
TX12B	29.83	72.54	30.42	73.72	2.42	0.00	0.00	0.00	20.00	169.00	476.00
TX12C_1TO9	24.46	76.80	22.85	74.03	3.24	0.00	0.00	0.00	5.00	131.00	616.00
TX12C_10	188.48	164.39	186.41	162.61	0.87	0.00	70.00	150.00	250.00	520.00	700.00
TX13B	365.16	471.29	359.11	464.39	1.29	0.00	0.00	213.00	541.00	1,358.00	2,162.00
TX14B	61.65	189.63	61.42	189.40	3.08	0.00	0.00	0.00	10.00	361.00	1,500.00
TX15A	245.49	470.98	246.97	470.99	1.91	0.00	0.00	48.00	267.00	1,210.00	2,520.00
TX16A	229.90	953.93	218.94	921.96	4.21	0.00	0.00	0.00	0.00	1,144.00	8,855.00
TX17A	51.66	148.87	49.87	144.96	2.91	0.00	0.00	0.00	39.00	251.00	1,134.00
TX18A	9.84	99.52	10.04	101.51	10.11	0.00	0.00	0.00	0.00	0.00	2,000.00
TX19A	228.15	436.29	223.08	427.25	1.92	0.00	0.00	51.00	241.00	1,065.00	2,375.00
TX19B	369.52	634.02	369.91	633.30	1.71	0.00	0.00	50.00	500.00	1,800.00	3,006.00

Table 3.6: Descriptive statistics for the second interview expenditure information after top-coding, non-reports treated as zeros (N=10,495)

Expenditure	Number Reporting	Mean	Std Dev	Weighted Mean	Weighted Std Dev	Weighted CV	Min	Q1	Median	Q3	95th Pctl	Max
TX2	2,941	2,241.85	1,382.80	2,164.03	1,336.26	0.62	3.00	1,230.00	1,950.00	2,925.00	5,100.00	6,480.00
TX3F	491	4,617.15	4,396.93	4,447.56	4,279.98	0.96	10.00	1,800.00	3,166.00	5,850.00	15,000.00	19,869.00
TX3H	589	1,855.45	2,236.79	1,804.67	2,215.81	1.23	3.00	600.00	1,200.00	2,040.00	6,000.00	12,000.00
TX4A_1	9,710	378.37	241.29	372.55	238.47	0.64	9.00	180.00	330.00	519.00	885.00	1,020.00
TX4A_2	2,811	102.86	37.50	102.88	37.18	0.36	1.00	75.00	99.00	120.00	180.00	195.00
TX4A_3	1,367	183.06	78.07	183.19	77.58	0.42	23.00	120.00	171.00	237.00	344.00	375.00
TX4A_4	24	48.75	42.58	47.78	43.42	0.91	6.00	15.00	31.50	80.50	135.00	140.00
TX4B	866	60.13	55.03	59.96	54.19	0.90	1.00	20.00	40.00	90.00	180.00	220.00
TX4C	6,839	94.70	66.40	93.19	64.88	0.70	1.00	53.00	77.00	117.00	243.00	327.00
TX4D	9,662	660.70	393.11	653.33	385.43	0.59	15.00	366.00	600.00	876.00	1,452.00	1,781.00
TX5A_1	185	360.41	556.18	354.08	557.85	1.58	2.00	50.00	148.00	380.00	2,000.00	2,500.00
TX5A_2	98	151.63	169.54	151.17	164.58	1.09	3.00	30.00	80.00	200.00	512.00	700.00
TX5B	1,905	1,974.41	3,400.94	1,893.57	3,310.79	1.75	1.00	175.00	550.00	2,000.00	10,000.00	16,054.00
TX6A	634	639.10	695.25	627.67	675.37	1.08	7.00	131.00	400.00	848.00	2,342.00	2,947.00
TX6B	4,609	418.25	565.16	414.68	561.73	1.35	1.00	63.00	189.00	510.00	1,800.00	2,500.00
TX7A	1,027	166.91	159.80	163.66	157.51	0.96	1.00	63.00	107.00	210.00	540.00	709.00
TX8	3,887	354.63	624.07	348.31	616.49	1.77	1.00	38.00	103.00	315.00	1,995.00	2,880.00
TX9A	7,022	303.37	318.33	296.23	312.03	1.05	2.00	83.00	200.00	400.00	1,015.00	1,443.00
TX9B	2,383	159.25	211.45	157.57	211.24	1.34	2.00	32.00	81.00	195.00	619.00	1,042.00
TX9C	714	57.37	75.71	57.63	75.74	1.31	1.00	15.00	30.00	70.00	200.00	379.00
TX9D	672	60.94	83.60	59.95	82.19	1.37	1.00	11.00	28.00	70.50	300.00	374.00
TX10C_1	418	504.88	311.62	501.13	309.82	0.62	75.00	302.00	400.00	600.00	1,300.00	1,571.00
TX10C_23	176	2,977.03	2,796.37	2,926.47	2,817.18	0.96	142.00	1,200.00	2,000.00	3,500.00	10,000.00	14,000.00
TX11B	1,864	3,903.71	3,940.94	3,768.22	3,849.75	1.02	1.00	1,000.00	2,500.00	5,000.00	12,000.00	18,000.00
TX12A	5,741	303.15	418.44	296.01	410.68	1.39	1.00	42.00	120.00	392.00	1,271.00	1,900.00
TX12B	2,928	106.91	103.08	107.28	104.50	0.97	1.00	36.00	75.00	135.00	350.00	476.00
TX12C_1TO9	2,852	89.99	125.74	87.18	123.70	1.42	1.00	17.00	48.00	104.00	360.00	616.00
TX12C_10	9,365	211.22	159.63	208.09	158.13	0.76	1.00	100.00	160.00	300.00	550.00	700.00
TX13B	6,815	562.34	480.79	551.31	474.47	0.86	1.00	230.00	408.00	741.00	1,650.00	2,162.00
TX14B	2,814	229.91	308.94	229.00	308.82	1.35	1.00	44.00	113.50	290.00	900.00	1,500.00
TX15A	6,463	398.64	546.96	398.48	545.47	1.37	1.00	64.00	190.00	477.00	1,675.00	2,520.00
TX16A	2,425	994.95	1,782.69	955.88	1,734.00	1.81	2.00	110.00	318.00	928.00	4,800.00	8,855.00
TX17A	3,502	154.83	224.61	150.81	219.85	1.46	1.00	39.00	74.00	162.00	650.00	1,134.00
TX18A	316	326.77	475.45	333.07	483.98	1.45	2.00	50.00	130.50	330.00	1,500.00	2,000.00
TX19A	7,085	337.96	494.84	329.02	484.13	1.47	1.00	50.00	150.00	400.00	1,445.00	2,375.00
TX19B	6,400	605.95	718.29	605.38	716.82	1.18	1.00	90.00	300.00	880.00	2,200.00	3,006.00

Table 3.7: Descriptive statistics for the second interview expenditure information after top-coding, excluding non-reports (N=10,495)

In Tables 3.5 and 3.7, we also present the number of CUs reporting each expenditure for the first and second interview, respectively. In addition to these values being quite variable, we also observe that there are many sample members in the analysis file with zero-dollar expenditure reports. For the first interview, the number reporting ranges from 42, TX4A_4 (modem purchase, Smartphone apps, ringtones), to 9,606, TX4A_1 (residential and mobile telephone services). These correspond to prevalences of 0.004 and 0.915, respectively. For the second interview, the range is 24, TX4A_4 (modem purchase, Smartphone apps, ringtones), to 9,710, TX4A_1 (residential and mobile telephone services), corresponding to prevalences of 0.002 and 0.925, respectively. We elaborate further on this observation in Section 3.3.1.1, specifically in Table 3.8, when we present various reporting probabilities for each expenditure.

The high CVs and low prevalence have direct bearing on this research. Recall that in Section 2.3.1 we presented two tables (Tables 2.3 and 2.4) summarizing issues related to prevalence, sample size, and CV. The high CV affects the sample size needed to achieve certain precision requirements. The more variable the expenditure is the larger the sample size must be to achieve a specified CV target. The motivation behind targeting likely purchasers of particular items, i.e., implementing a responsive split questionnaire design, is to address the issues associated with the large sample sizes needed to achieve the precision requirements when a standard split questionnaire is implemented.

Finally, even after top-coding extreme values at the 97.5^{th} percentiles, we still observe some relatively high expenditures reported for a few categories. In both

interviews, there are expenditure reports exceeding $\$8,000$. These reports occur in the following categories: mortgages, home loans, contractor labor and materials, car down payments, and educational expenses. It is reasonable to expect high expenses for these categories.

3.2.3.4 Implications of descriptive statistics

The role of this section was to summarize the descriptive statistics for the demographic characteristics, timing information, and expenditure information. While the descriptive statistics, in their own right, do not provide a complete picture of purchase behavior, their presentation is a crucial element in understanding the results from the proposed responsive split questionnaire methods. In addition, these statistics have several implications for the dissertation research. The key insight from these descriptive statistics was that for many expenditure categories, the weighted CVs are quite high. A high CV for a specific characteristic may inhibit the ability to achieve certain precision requirements. The second insight is that the number of sample units reporting certain purchases within a given quarter can be quite low. By developing question asking methods to target likely purchasers, we might address the issues associated with low effective sample sizes that arise when a standard (non-responsive) split questionnaire is utilized.

As a final note, it is beyond the scope of this dissertation to extensively ponder why expenditure data have a high noise-to-signal ratio, i.e., high CV. We believe that this is an intrinsic feature of expenditure data and acknowledge that any redesign

effort will encounter this problem. It may be the case that many expenditure categories are underreported, so this inflates the number of non-reports which in turn decreases the prevalence of the expenditure. The low prevalence directly relates to the high CVs.

3.3 Preliminary studies

In this section we present the results of two preliminary studies. The objective of the first was to explore the relationships among expenditures and demographic characteristics. This information will not only be incorporated into the responsive split questionnaire methods developed in Chapter 4, but will also be used to summarize key findings from our research. The second preliminary study was designed as an initial attempt at demonstrating the utility of extending the current set of split questionnaire methods by addressing some of the limitations of the previous study by Gonzalez and Eltinge (2008). We also use this second preliminary study to establish a baseline for comparison for the methods we develop in Chapter 4.

3.3.1 Preliminary study 1: Understanding data relationships for decision rule development

As mentioned earlier, a key component of a responsive design is the decision rule. It is a function relating information from prior collected data to a design modification. The goal of the decision rule is to select the design modification which leads to an improvement in the error properties of the resulting statistics. In our

case, the design modification is a tailored set of questions administered to each respondent. The intended purpose of this modification is to decrease the length of the interview while reducing burden and improving response quality. Before we can develop the design modification that will acheive this goal, we must understand the relationships in the data. In our research, this entails understanding the relationships among expenditures and demographic characteristics. Therefore, for this first preliminary study, we report on findings from three analyses: (1) calculations of reporting probabilities (i.e., prevalence); (2) calculations of cross-interview bivariate correlations; and, (3) regression analyses examining which demographic characteristics are associated with incurring each expense.

3.3.1.1 Reporting probabilities

The first analysis we report on is the calculation of various reporting probabilities. These provide us with an understanding as to whether incurring an expense in one reference period[5] is related to incurring that same expense in the next, and the basis for classifying (characterizing) that relationship.

The following probabilities were computed using the analysis file and are presented in Table 3.8: (1) $P(Int1)$ is the probability of incurring the expense[6] during the reference period asked about in the first interview; (2) $P(Int2)$ is the probability of incurring the expense during the reference period asked about in the second

[5]Reference period is defined as the time frame for which respondents are asked to report incurred expenses.

[6]We use the phrase "incurring the expense" and "reporting the expense" interchangeably. In doing so, we effectively assume that an incurred expense is reported by the respondent during the interview.

interview; (3) $P(Int1, Int2)$ is the probability of incurring the same expense during both reference periods; (4) $P(Int1^c, Int2)$ is the probability of incurring the expense in the second interview reference period but not in the first; (5) $P(Int1, Int2^c)$ is the probability of incurring the expense in the first interview reference period but not in the second; (6) $P(Int1^c, Int2^c)$ is the probability of not incurring the expense in either reference period; (7) $P(Int2|Int1)$ is the conditional probability of incurring the expense in the second interview reference period, given that the respondent incurred it in the first; and, (8) $P(Int2|Int1^c)$ is the conditional probability of incurring the expense in the second interview reference period, given that the respondent did not incur it in the first.

Expenditure	$P(Int1)$	$P(Int2)$	$P(Int1, Int2)$	$P(Int1^c, Int2)$	$P(Int1, Int2^c)$	$P(Int1^c, Int2^c)$	$P(Int2\|Int1)$	$P(Int2\|Int1^c)$	Rare	Recurrent
TX2	0.278	0.280	0.268	0.012	0.010	0.709	0.963	0.017	No	Yes
TX3F	0.045	0.047	0.041	0.006	0.004	0.949	0.913	0.006	Yes	Yes
TX3H	0.058	0.056	0.049	0.008	0.009	0.935	0.843	0.008	Yes	Yes
TX4A_1	0.915	0.925	0.893	0.032	0.022	0.052	0.976	0.381	No	Yes
TX4A_2	0.288	0.268	0.209	0.059	0.079	0.653	0.726	0.082	No	Yes
TX4A_3	0.130	0.130	0.096	0.034	0.034	0.836	0.737	0.040	No	Yes
TX4A_4	0.004	0.002	0.000	0.002	0.004	0.994	0.095	0.002	Yes	No
TX4B	0.069	0.083	0.036	0.047	0.034	0.884	0.517	0.050	Yes	No
TX4C	0.632	0.652	0.582	0.069	0.050	0.299	0.921	0.188	No	Yes
TX4D	0.917	0.921	0.892	0.029	0.025	0.055	0.973	0.345	No	Yes
TX5A_1	0.024	0.018	0.002	0.016	0.023	0.960	0.067	0.016	Yes	No
TX5A_2	0.009	0.009	0.001	0.008	0.008	0.983	0.148	0.008	Yes	No
TX5B	0.131	0.182	0.056	0.126	0.075	0.743	0.427	0.145	No	No
TX6A	0.042	0.060	0.004	0.056	0.038	0.901	0.099	0.059	Yes	No
TX6B	0.330	0.439	0.206	0.234	0.125	0.436	0.622	0.349	No	No
TX7A	0.065	0.098	0.024	0.074	0.042	0.861	0.363	0.079	Yes	No
TX8	0.281	0.370	0.158	0.212	0.123	0.507	0.563	0.295	No	No
TX9A	0.545	0.669	0.427	0.242	0.118	0.213	0.784	0.532	No	Yes
TX9B	0.168	0.227	0.091	0.136	0.077	0.696	0.544	0.163	No	No
TX9C	0.042	0.068	0.010	0.058	0.032	0.900	0.241	0.060	Yes	No
TX9D	0.045	0.064	0.017	0.047	0.028	0.908	0.383	0.049	Yes	No
TX10C_1	0.037	0.040	0.037	0.003	0.000	0.960	1.000	0.003	Yes	Yes
TX10C_23	0.015	0.017	0.015	0.002	0.000	0.983	1.000	0.002	Yes	Yes
TX11B	0.170	0.178	0.170	0.008	0.000	0.822	1.000	0.010	No	Yes
TX12A	0.353	0.547	0.239	0.308	0.114	0.339	0.677	0.476	No	No
TX12B	0.152	0.279	0.049	0.230	0.103	0.618	0.322	0.271	No	No
TX12C_1TO9	0.012	0.272	0.006	0.266	0.007	0.721	0.446	0.270	Yes	No
TX12C_10	0.892	0.892	0.865	0.027	0.026	0.081	0.971	0.248	No	Yes
TX13B	0.527	0.649	0.446	0.203	0.081	0.270	0.847	0.429	No	Yes
TX14B	0.198	0.268	0.156	0.112	0.042	0.689	0.786	0.140	No	Yes
TX15A	0.481	0.616	0.400	0.216	0.081	0.303	0.831	0.416	No	Yes
TX16A	0.181	0.231	0.123	0.108	0.058	0.711	0.680	0.132	No	No
TX17A	0.255	0.334	0.173	0.161	0.082	0.584	0.678	0.216	No	No
TX18A	0.019	0.030	0.005	0.026	0.015	0.955	0.237	0.026	Yes	No
TX19A	0.598	0.675	0.515	0.161	0.083	0.241	0.860	0.399	No	Yes
TX19B	0.500	0.610	0.424	0.186	0.076	0.314	0.848	0.372	No	Yes

Table 3.8: Unweighted reporting probabilities of expenditures for the first and second interviews

The first two probabilities, (1) and (2), describe the prevalence of incurring the expense during the first and second interview reference periods. The joint probabilities, (3) – (6), give an indication of the extent that a sample unit incurs the expense in both, one, or neither reference period. Finally, the conditional probabilities, (7) and (8), reflect an updated assessment of incurring the expense during the second interview reference period based on the knowledge of whether or not the sample unit incurred the expense during the first interview reference period.

These probabilities can be related to the gaps we identified in Section 2.3.1 and that we address in this dissertation. One gap is that the existing set of split questionnaire methods are ineffective for surveys collecting information on rare events. To assess whether our methods improve the design of split questionnaires for surveys with questions about rare events, we must classify the expenditures on the basis of being rare, i.e., infrequently incurred. Using the probabilities $P(Int1)$ and $P(Int2)$, we identify the rare expenditures. Using a similar cut-off as Cochran (1977), we classify a rare expenditure as any expense for which $P(Int1)$ or $P(Int2)$ is less than or equal to 0.1.

It is worth noting that the medical field uses a more restrictive classification for rare characteristics. Specifically, the *Rare Diseases Act of 2002* defines rare diseases as those which affect populations smaller than 200,000 persons in the United States. This is equivalent to a prevalence of less than 0.001, far lower than the cut-off we used. However, for the purposes of this research, using 0.1 as the cut-off is warranted since characteristics with a prevalence below this value tend to affect the stability of mean estimates, assuming SRS (see Table 2.4). Classifying the expenditures in this

manner yields 14 rare expenditures (see the column labeled "Rare" in Table 3.8).

The other gap we address is that prior information on the sample unit is not fully utilized in the design of a split questionnaire. To address this gap, survey designers may use the sample unit's characteristics to determine which questions to ask. With respect to our research, we can use our knowledge of whether the sample unit incurred the expense in the first interview reference period to determine whether to ask about that expense in the second interview. However, this requires that we have an understanding of whether the expense is recurring across the two reference periods.

A simple definition of a recurrent expenditure is an expense that a sample unit is likely to have either monthly, quarterly, or over some other regularly spaced time interval. However, simply incurring the expense across multiple time intervals does not completely reflect the notion of recurrence. For instance, incurring the expense in a later reference period may be conditional on incurring it during an earlier reference period. In this instance, the expenditure would be the type that the recurrence is dependent on some prior event (e.g., an initial purchase). Therefore, recurrence may not only mean incurring the expense in both reference periods, but also it may mean that given the sample unit incurred the expense, it will incur it again at a later time period.

Using $P(Int1, Int2)$ and $P(Int2|Int1)$, we offer a formal definition of a recurrent expenditure based on the definition of conditional probability. For events A (e.g., $Int1$) and B (e.g., $Int2$), with $P(B) > 0$, the conditional probability of B

given A is defined as follows.

$$P(B|A) = \frac{P(AB)}{P(A)} \tag{3.3}$$

From equation (3.3), we observe that $P(AB) = P(B|A)P(A)$. Basing our definition of a recurrent expenditure solely on high values of $P(AB)$ (e.g., $P[Int1, Int2]$) means that we define recurrence by a high likelihood of having it during both time periods. This represents a stronger, yet incomplete, criterion for recurrence than basing it on on high values of $P(B|A)$ (e.g., $P[Int2|Int1]$). This is because the joint probability requires the product of the conditional and the marginal probabilities to be high whereas the conditional probability requires the ratio of the joint to the marginal to be high (with $P[AB] \leq P[A]$). Thus, we base our definition of a recurrent expenditure on high values for $P(B|A)$. This criterion not only encompasses both notions of recurrence, but it also more appropriately reflects an updated assessment of incurring the expense based on knowledge of whether the sample unit incurred it in the first time period.

We classified expenditures as recurrent if the conditional probability, $P(Int2|Int1)$ was above the arbitrary cut-off of 0.7. From Table 3.8, under the column labeled "Recurrent," we see that we have 18 recurrent expenditures. The expenditures falling into this classification are reasonable since they relate to housing, utilities, clothing, car payments, vehicle operating expenses, insurance (health and other), medical expenses, and cash contributions.

To further our understanding of the relationship between incurring an expense

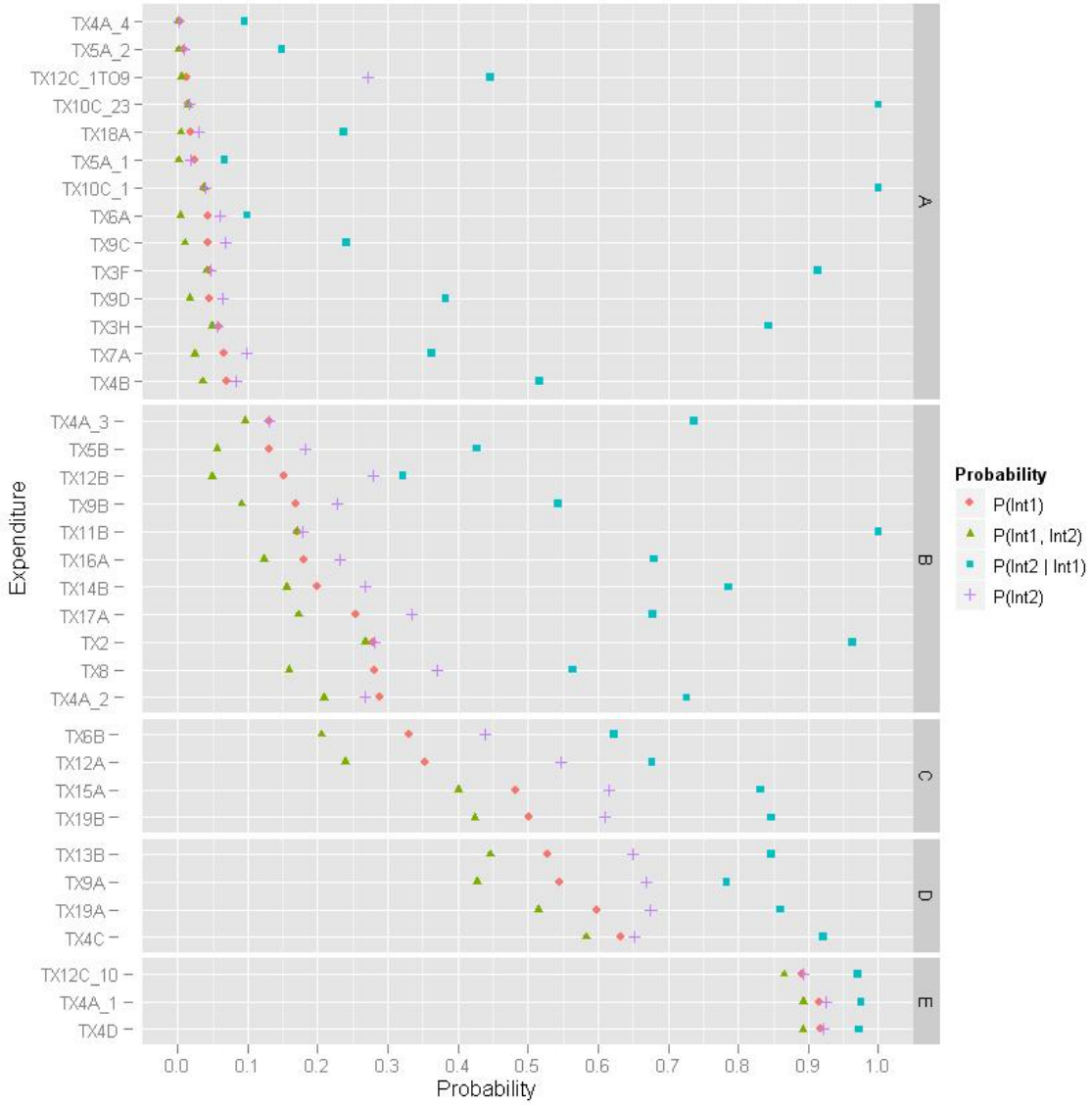

Figure 3.1: Graphical display of unweighted reporting probabilities for the first and second interviews

in one reference period and incurring that same expense in a later reference period, we plot the following probabilities in Figure 3.1: $P(Int1)$, $P(Int2)$, $P(Int1, Int2)$, and $P(Int2|Int1)$. For illustration purposes only, we sorted and classified the expenditures into classes based on their first interview reporting probabilities. We define the following classes: Class A includes the expenditures for which $P(Int1) \in (0, 0.1]$; Class B includes the expenditures for which $P(Int1) \in (0.1, 0.3]$; Class C includes the expenditures for which $P(Int1) \in (0.3, 0.5]$; Class D includes the expenditures for which $P(Int1) \in (0.5, 0.7]$; and, Class E includes the expenditures for which $P(Int1) \in (0.7, 1]$. It is worth noting that Class A includes only rare expenses.

The following trends are discernible from Figure 3.1. In general, $P(Int1)$, $P(Int2)$, and $P(Int1, Int2)$ track each other quite well. This suggests that prevalence of a given expenditure is relatively stable across the two reference periods. Second, even for rare expenses, knowledge of incurring the expense in the first interview reference period is indicative of incurring the expense in the second interview reference period. This is indicated by the relatively high conditional probabilities, $P(Int2|Int1)$. Specifically, for a vast majority of the expenditures, we observe conditional probabilities greater than 0.5. This suggests that prior purchase behavior is indicative of subsequent purchase behavior. This latter finding is encouraging given that we have motivated this research, in part, from the perspective that prior knowledge about incurring a particular expense can inform the decision to ask about that expense in a subsequent phase of data collection.

3.3.1.2 Cross-interview bivariate correlations

Not only is the relationship between incurring the expense in both reference periods relevant to decision rule development, but so is the relationship between the amounts of the expense in both reference periods. Thus, the second analysis we report on is cross-interview bivariate correlations. In our research, these correlations represent the association between the levels of an expenditure of the same type across the two reference periods. This gives an indication of the degree to which the expenditure tended to occur (or not occur) in both reference periods and when it did (or did not) that the amount was about the same.

We acknowledge that it is impossible to summarize the history of purchase behavior into one scalar quantity, but these correlations still provide insight to how the amounts of the expense incurred in both reference periods are related. They provide a quantitative measure of the linear relationship between expenses. A high correlation suggests that large values of the expense in one reference period correspond to large values of the expense in the next reference period and low values of the expense in one reference period correspond to low values of the expense in the next reference period.

In Table 3.9, we present the cross-interview bivariate correlations of expenditures of the same type across the two reference periods[7]. We treated non-reports

[7]For completeness, in Appendix C, we present a series of four tables containing the full set of cross-interview bivariate correlations, including the off-diagonal correlations. The bivariate correlations on the main diagonal represent the correlation between the same expenditure variable whereas the bivariate correlations on the off-diagonal represent the correlation between two different expenditures across the two reference periods. In Tables C.1 – C.4, the darker shading corresponds to the main diagonal bivariate correlations. The lighter shading corresponds to correlations of two expenditures from the same section. Finally, the off-diagonal correlations which are greater than 0.2 are displayed with bold typeface.

as zeros for these calculations. It is worth noting that all of the expenditures with

Expenditure	Correlation	Expenditure	Correlation
TX2	0.929	TX9B	0.219
TX3F	0.882	TX9C	0.144
TX3H	0.568	TX9D	0.385
TX4A_1	0.812	TX10C_1	0.968
TX4A_2	0.622	TX10C_23	0.914
TX4A_3	0.678	TX11B	0.965
TX4A_4	0.016	TX12A	0.143
TX4B	0.386	TX12B	0.018
TX4C	0.683	TX12C_1TO9	0.011
TX4D	0.661	TX12C_10	0.690
TX5A_1	0.024	TX13B	0.266
TX5A_2	0.007	TX14B	0.794
TX5B	0.179	TX15A	0.336
TX6A	0.039	TX16A	0.221
TX6B	0.129	TX17A	0.361
TX7A	0.111	TX18A	0.049
TX8	0.152	TX19A	0.375
TX9A	0.365	TX19B	0.536

Table 3.9: Bivariate correlations for the same expense across the two reference periods

correlations above 0.5 were identified as recurrent expenses using the definition provided in the previous section. Even though the correlation is an incomplete measure of association (e.g., there may be a non-linear association between the two variables that a correlation would not detect), we can use these values to suggest that the more closely related two variables are, the variables can be predicted from each other. This implies that knowledge of the prior reference period expense can help predict the amount of the expense incurred in the next reference period. This has

direct bearing on our research since we proposed the use of prior interview expenditure information in the decision rule regarding whether to administer a particular question to the respondent.

3.3.1.3 Covariates associated with incurring an expense

With the knowledge that expenses are related, in terms of incurring the expense and the amount of the expense, across successive reference periods, the next relevant consideration for decision rule development is whether a sample unit's demographic characteristics can help in the prediction of whether or not that unit will incur the expense. Thus, in this analysis we explore the relationship between certain demographic characteristics and the likelihood of incurring an expense. This analysis will help us identify subgroups of sample members that are more likely to incur an expense during a particular reference period.

To explore this, we fit a series of logistic regression models to the sample units in the analysis file. The first series incorporated only the first interview information whereas the second series incorporated information from both interviews. The first series represents the situation in which we only have information available from the first interview. In essence, we are exploring the relationship between demographic characteristics and incurring an expense within a single reference period. The second series extends this by examining both the relationship among demographic characteristics and incurring the expense, as well as whether the sample unit previously incurred the expense.

For the first series, using only first interview information, we fit the following logistic regression model to every sample unit in the analysis file, i, and every expenditure, k. A listing of the levels of each covariate in equation (3.4) can be found in Table A.2 of Appendix A.3 and the unweighted proportions falling into each level for each characteristic are displayed in Table 3.1 of Section 3.2.3.1.

$$\begin{aligned}\text{logit}(p_{Int1,ik}) = \beta_0 + \sum_j \beta_{1j} \times \text{REGOFF}_{ij} + \beta_2 \times \text{SIZE}_i + \beta_3 \times \text{POVERTY}_i \\ + \beta_4 \times \text{URBAN}_i + \beta_5 \times \text{TENURE}_i + \beta_6 \times \text{RACE}_i \qquad (3.4)\end{aligned}$$

In equation (3.4) $p_{Int1,ik}$ denotes the probability that sample unit i incurred expense k during the reference period asked about in the first interview. In Table 3.10, we display the Type III (variables-added-last) significance of parameters for each of the 36 regression models with the following significance codes: 0 '$***$' 0.001 '$**$' 0.01 '$*$' 0.05 '$\bullet$' 0.1 ' ' 1. The significance codes refer to the joint test that all levels of the factor variable, e.g., REGOFF, are zero. A significance code of '$**$' indicates that the p-value for the Type III significance test of the parameter (or joint test for all levels of the parameter) is between 0.001 and 0.01 while a code of ' ' means the p-value is greater than 0.1.

Given the Type III significance of parameters in Table 3.10, overall, we observe that the set of demographic characteristics we consider in this research, are good predictors of whether or not a sample unit incurred the expense during the first interview reference period. Using only parameter significance as the criterion,

Expenditure	Poverty	Reg Office	Urban	Tenure	Race	CU Size	p-value
TX2		* * *	* * *	* * *	*	* * *	0.412
TX3F	*	* * *	**	* * *		* * *	0.586
TX3H	* * *	* * *		* * *	* * *	* * *	0.176
TX4A_1	* * *	**	*	* * *	*	* * *	0.001
TX4A_2	* * *	* * *	* * *	* * *		* * *	0.121
TX4A_3	* * *	* * *	* * *	* * *		* * *	0.016
TX4A_4		*		**			0.998
TX4B	**	**		* * *		* * *	0.238
TX4C	* * *	* * *		* * *	*		0.593
TX4D	* * *	* * *		* * *		* * *	0.002
TX5A_1		*		* * *			0.794
TX5A_2		* * *		**			0.724
TX5B				* * *		* * *	0.374
TX6A				•		* * *	0.633
TX6B	* * *	* * *			* * *	**	0.002
TX7A	*	* * *		* * *		*	0.982
TX8		* * *		*		* * *	0.533
TX9A	**	* * *	* * *		*	* * *	0.404
TX9B		* * *		**		* * *	0.357
TX9C	**	•	* * *	**			0.819
TX9D		* * *	* * *	* * *	*	* * *	0.672
TX10C_1	* * *	* * *	* * *	* * *	*	* * *	0.910
TX10C_23	•	* * *	**			**	0.691
TX11B	* * *	* * *	*		*	* * *	0.167
TX12A	*	* * *		* * *		* * *	0.607
TX12B	* * *	* * *		* * *	**	* * *	0.380
TX12C_1TO9		*	* * *	*		**	0.751
TX12C_10	* * *	* * *		* * *	* * *	* * *	0.016
TX13B	* * *	* * *	*	* * *		* * *	0.010
TX14B	* * *	* * *		* * *	**		0.778
TX15A	* * *	* * *		* * *	**	**	0.528
TX16A	**	* * *	* * *		* * *	* * *	0.001
TX17A	* * *	* * *	* * *	* * *	* * *		0.790
TX18A		**	*		**		0.840
TX19A	* * *	* * *		* * *	* * *	* * *	0.220
TX19B	* * *	* * *	•	* * *	•	* * *	0.016

Table 3.10: Significance of parameters for the logistic regression models defined by equation (3.4)

regional office, housing tenure, and CU size demonstrate the most significance across the models.

We also use the Hosmer-Lemeshow statistic, with ten groups, to assess the goodness-of-fit of each of the 36 logistic regression models given by equation (3.4). This statistic assesses the discrepancy between the observed event rates and expected event rates in the subgroups determined by the model covariates. Models for which the observed and expected event rates in the subgroups are similar fit well. The Hosmer-Lemeshow statistic is denoted as G^2_{HL} and is given as follows

$$G^2_{HL} = \frac{\sum_{g=1}^{n=10} (O_g - E_g)^2}{N_g \pi_g (1 - \pi_g)} \sim \chi^2_{n-2} \tag{3.5}$$

where O_g is the number of observed events in group g; E_g is the number of expected events in group g; N_g is the number of observations in group g; π_g is the predicted risk for the g^{th} risk decile group; and, n is the number of groups.

The null hypothesis is that the model fits the data well, so observing a p-value greater than 0.05 is the preferred outcome. We report the p-value associated with the Hosmer-Lemeshow statistic for the goodness-of-fit test associated with each logistic regression model in the last column of Table 3.10. We observe that the model fits well for a vast majority of the expenses; however, for eight of the 36 expenses we would reject the null hypothesis. There are no common characteristics among the expenditures for which the model does not fit well. Despite rejecting the null hypothesis for these models, within each, there are significant parameters, so we

believe these models still provide reasonable predictions of whether the sample unit incurred the expense.

For the next series of regression models, we wanted to further our understanding of how incurring an expense in one reference period is related to incurring the expense in the next reference period. Thus, we modeled the probability that the sample unit incurred expense k during the second interview reference period, denoted as $p_{Int2,ik}$, using the same demographic characteristics as equation (3.4), and included an indicator for whether the sample unit incurred the expense in the first interview reference period. So, for every unit i in the analysis file and for each expenditure k, using a combination of first and second interview information, we fit the following model

$$\begin{aligned}\text{logit}(p_{Int2,ik}) = \beta_0 + \sum_j \beta_{1j} \times \text{REGOFF}_{ij} + \beta_2 \times \text{SIZE}_i + \beta_3 \times \text{POVERTY}_i \\ + \beta_4 \times \text{URBAN}_i + \beta_5 \times \text{TENURE}_i + \beta_6 \times \text{RACE}_i + \beta_7 \times I_{ik}\end{aligned} \tag{3.6}$$

where $I_{ik} = 1$ if sample unit i incurred expense k during the reference period inquired about at the first interview and 0 otherwise.

It is worth reiterating the important distinction between equations (3.4) and (3.6). Equation (3.4) is fit only to the first interview data while equation (3.6) is fit to a combination of both first and second interview information. This fact has direct relevance to our research. In most applications of responsive designs, survey designers only have access to the information that will be incorporated into the first

model, (3.4), because this information comes from an initial phase of data collection. In practice, outputs from a model of the form (3.6) are only obtainable if there had been a prior administration of the survey or access to auxiliary data that would have provided the second interview expenditure information. The model represented in equation (3.6) is useful, however, for understanding the interplay among demographic characteristics and incurring an expense in both reference periods. Thus, we can use these models to develop decision rules in two cases: (1) when there is only data available from an initial phase of data collection and (2) when, in addition to initial phase of data collection data, there is access to auxiliary information or a prior administration of the survey. We consider both cases when we develop the methods for responsive split questionnaire panel surveys in Chapter 4.

In Table 3.11, we display the Type III (variables-added-last) significance of parameters for each of the 36 regression models identified by (3.6) with the following significance codes: 0 '$***$' 0.001 '$**$' 0.01 '$*$' 0.05 '$\bullet$' 0.1 ' ' 1. We also use the Hosmer-Lemeshow, given by equation (3.5), to assess the goodness-of-fit for each of the models. Given the Type III significance of parameters in Table 3.10, overall, we observe that the set of demographic characteristics we consider in this research, are good predictors of whether or not a sample unit incurred the expense during the second interview reference period. We also observe that in each of the models, the parameter associated with the indicator of the sample unit incurring the expense in the first interview reference period, e.g. β_7, attains statistical significance (see column labeled "Int1"). This further supports the conclusion that knowledge of incurring the expense in the first interview reference period is indicative of whether

Expenditure	Poverty	Reg Office	Urban	Tenure	Race	CU Size	Int1	p-value
TX2	*	**	***	***			***	0.295
TX3F				***			***	0.941
TX3H		**		***			***	0.642
TX4A_1	**			***		***	***	0.280
TX4A_2	***		*	***	*	***	***	0.001
TX4A_3	***	***	***	***		***	***	0.282
TX4A_4							***	0.299
TX4B	***	***		*	***	***	***	0.295
TX4C	•	***	•		*	*	***	0.002
TX4D	***			***		***	***	0.040
TX5A_1		**		***		**	***	0.804
TX5A_2		*		***			***	0.869
TX5B	**	**		***	•		***	0.672
TX6A			*	***		***	**	0.004
TX6B	***	***	*	***	***	***	***	0.285
TX7A	**	***		***		***	***	0.114
TX8	***	***		***	**	***	***	0.025
TX9A	***	***	**		**	***	***	0.929
TX9B	*	***	**	•		***	***	0.503
TX9C	***	***	***	***	•		***	0.863
TX9D		***		***	***	***	***	0.397
TX10C_1			*				***	0.165
TX10C_23			**			•	***	0.979
TX11B	•	*				**	***	0.683
TX12A	***	***		***	***	***	***	0.004
TX12B	***	***	*	***	**	***	•	0.292
TX12C_10	**	***		***		***	***	0.022
TX12C_1TO9	***	***	***	***	***	***	***	0.003
TX13B	***	***		***		***	***	0.007
TX14B	***	***		***	*		***	0.247
TX15A	***	***		***	**	*	***	0.030
TX16A	***	***	**	•		***	***	0.001
TX17A	***	***	***	***	***	*	***	0.152
TX18A	**	***	**				***	0.780
TX19A	***	***		***	***	***	***	0.311
TX19B	***	***		***			***	0.566

Table 3.11: Significance of parameters for the logistic regression models defined by equation (3.6)

the sample unit will incur the expense in the second interview reference period. As with the previous set of models, defined by equation (3.4), regional office, housing tenure, and CU size demonstrate statistical significance across the models, but for these models, the poverty indicator attains statistical significance in many of the models as well.

In Table 3.11, we report the p-value associated with each logistic regression model in the last column. Despite the null hypothesis being rejected in about a third of the models, we conclude the models still provide better predictions of incurring the expense in the second reference period than would a model that ignores the demographics and prior reference period indicator. One reason for rejecting the null in so many cases of model (3.6) may be due to the fact that the prior reference period indicator of having the expense is a strong predictor of incurring the expense in the second reference period. By including other covariates, we may actually be overfitting the models.

As a final note, since our primary purpose of this analysis was to obtain the best prediction of the propensity to incur the expense, through the identification of subgroups in our analysis file defined by their demographic characteristics, we chose not to eliminate the statistically insignificant parameters from the regression models and reassess the goodness-of-fit for each. If our goal had been to obtain good estimates of the parameters associated with each characteristic in each model, then we would have conducted a more formal and exhaustive model building process. We also believe that keeping all of the models the same represents a more parsimonious and tractable approach to modeling the information. Furthermore, using the same

model for each expense might be easier to implement in real applications of the methods developed in this dissertation.

3.3.1.4 Implications of preliminary study 1

This first preliminary study examined: (1) reporting probabilities; (2) cross-interview correlations; and (3) logistic regression analyses for incurring an expense during each reference period. The primary purpose of this preliminary study was to better understand the relationships among demographic characteristics and expenditures. This information will inform the development of the decision rules for the responsive split questionnaire methods.

Through these analyses, we have gained an understanding of the extent to which sample units incur the same expense across the two reference periods, in doing so, which expenditures are considered recurrent. We also identified which expenditures are considered rare. We examined the extent to which expenditure amounts are linearly related across the two reference periods. We observed that for several expenses, high amounts of the expenditure in the first reference period are indicative of incurring a high expense in the second reference period. We demonstrated that we can obtain a good prediction of the propensity to incur the expense in a reference period by accounting for the demographic characteristics of the sample unit. Finally, we also demonstrated incorporating indicators for incurring the expense during the first interview reference period is predictive of incurring the expense during the second interview reference period.

This preliminary study has one primary implication for the methods we develop in Chapter 4, namely, that prior information, whether in the form of a demographic characteristic, an indicator of incurring the expense in the prior reference period, or the amount of the expense, is indicative of the purchase behavior that a sample unit will exhibit at a later time. These findings support our proposal to incorporate prior information in the decision rules regarding which questions to ask a respondent in a split questionnaire survey.

3.3.2 Preliminary study 2: Extensions of Gonzalez and Eltinge (2008)

The second preliminary study replicated, with a few modifications, the simulation of Gonzalez and Eltinge (2008). Gonzalez and Eltinge (2008) explored adaptive assignments of survey items for the second interview based on probabilities constructed from the expenditure information collected during the first interview. The goal of their subsampling procedure was to oversample the sample units that were likely to be different from the mean expenditure for a specific category in order to capture the variability in that expenditure. In other words, higher probabilities for asking the question about the expenditure were assigned to those units whose deviation (reported expense less the mean across all sample units at the first interview) was large.

The focus of this dissertation is to broaden the scope of the question asking procedure, in part, because the current set of split questionnaire methods are ineffective for surveys collecting information on low prevalence events. For these

surveys, further reducing the sample size by implementing a split questionnaire may present problems when meeting stakeholder needs or producing estimates of desired quantities from these events. So, the primary goal of the methods developed in this dissertation is to oversample those sample units who are likely to incur the expense by the second interview, and not just oversample those that are likely to be different. This should result in a larger effective sample size for the sample units with the expense. By extending the simulation of Gonzalez and Eltinge (2008), we begin to explore ways to accomplish that goal and identify specific issues that need to be addressed when developing the responsive split questionnaire methods.

We modified their simulation study in the following ways. First, we relaxed the constraint of asking only one expenditure question to the sample units. In this preliminary study, each sample unit had the potential to receive multiple items. This was done because it represents a more realistic approach to surveying a population. It is also the situation under which we develop our methods. Second, we expanded the universe of survey items. In their simulation, they restricted their investigation to a shortened version of the questionnaire which consisted of only five sections. In this preliminary study, we considered the full set of 36 expenditure variables; thus, we explored the assignment of questions from the entire CE survey instrument. Third, we incorporated the epidemiological metrics we identified in Section 2.6.2 to evaluate the success of the methods in terms of customizing the survey to the individual respondent. Finally, we investigated the potential for burden reduction. Since we used the analysis file described in Section 3.2.2 to conduct the updated simulation, we had access to the timing information from the original CE instrument.

We computed the average length of interview (in minutes) and the average number of expenditure questions administered to the sample units across the simulation iterations. This enabled us estimate how the split questionnaire methods reduce burden, when burden is operationalized as the length of the survey and/or the number of questions asked to a sample unit.

3.3.2.1 Simulation setup

For this simulation, the questionnaire is split using the design depicted in Figure 1[f]. This implies that there are no restrictions on the number or structure of the questions asked to each sample unit. We explored three simulation conditions and these are summarized in the enumerated list below. Using the same notation as before, we have $y_{Int1,ik}$ denoting the i^{th} sample unit's first interview expense on item k, $\bar{y}_{Int1,k}$ is the average first interview expense for item k across all sample units in the analysis file, and p_{ik} is the probability that the i^{th} sample unit is asked survey item k in the second interview.

1. **One-half**: $p_{ik} = 1/2, \forall k$
2. **Absolute relative deviation (ARD)**: $p_{ik} \propto \left|y_{Int1,ik} - \bar{y}_{Int1,k}\right| / \bar{y}_{Int1,k}$
3. **Four-fifths**: $p_{ik} = 4/5$ if $y_{Int1,ik} > 0$ and $p_{ik} = 1/5$ if $y_{Int1,ik} = 0$

The first condition is labeled "one-half" and the goal of this condition was to examine the effect of reducing the length of the questionnaire by one-half. This serves as a check on the simulation and may be used as a baseline for comparison

since we know exactly what values the design effects and the epidemiological criteria should take under this condition (see Section 2.6). The second condition, ARD, is identical to the preferred method of subsampling (as judged by A- and D-optimality) in the original simulation (Gonzalez and Eltinge, 2008). This condition was designed to oversample the units that were likely to be different from the mean expenditure for the k^{th} item.

The final condition, labeled "four-fifths," was new and designed to be less restrictive than ARD, in that, we were not only interested in asking questions to those sample units that were likely to be different, but rather, we were interested in asking questions to any unit that was likely to incur the expense k during the second interview reference period. Under this condition we classified the sample units into two categories based on whether they incurred the expense during the first interview reference period. If we assume incurring the expense in the first interview reference period is directly related to incurring the expense in the second interview reference period, then the sample units for which $y_{Int1,ik} > 0$ comprise the likely purchaser group. In the first preliminary study, we demonstrated that, for several expenditures, this is not an unrealistic assumption. For sample units who incurred the expense during the first interview reference period, we asked about that expense during the second interview with probability $4/5$[8]. For those who did not incur the expense during the first interview reference period, we asked about it during the second interview with probability $1/5$. This final condition was implemented primarily to demonstrate the utility of extending the current set of split questionnaire

[8]This was an arbitrarily chosen probability designed to oversample likely purchasers.

methods under a responsive design framework by incorporating prior information about the sample unit into the question asking procedure.

3.3.2.2 Computations for simulations

For each condition, we carried out the simulation ($M =$)$1,000$ times and for each iteration, we randomly "asked" sample units questions based on the probabilities under the three conditions using a Bernoulli trial. We then computed various quantities to summarize the simulations. First, if we let S be the set of units in the CE analysis file (equivalently, the set of units for which the split questionnaire is designed) and y_{ik} the i^{th} sample unit's expenditure value for the k^{th} expenditure at the second interview, then we computed the overall simulation mean, $\bar{\theta}_k$, across the M iterations for the k^{th} expenditure; the simulation variance for the k^{th} expenditure mean, denoted as V_k; the simulation standard error, SE_k; the simulation CV, CV_k; the simulation relative bias (measured in percent), RB_k; the simulation relative bias standard error (and associated 95% confidence interval, $RBSE_k$); the simulation root mean squared error, $RMSE_k$; and, the design effect for estimate k, deff_k. These quantities are given in equations (3.7) – (3.14), respectively.

$$\bar{\theta}_k = M^{-1} \sum_{m=1}^{M} \hat{\bar{y}}_{mk} \tag{3.7}$$

where $\hat{\bar{y}}_{mk} = \left(\sum_{i \in S} p_{ik}^{-1} \alpha_{ik}\right)^{-1} \sum_{i \in S} p_{ik}^{-1} \alpha_{ik} y_{ik}$ with $\alpha_{ik} = 1$ if the i^{th} unit is asked the survey item about the k^{th} expenditure (and 0 otherwise); and, $p_{ik} = P(\alpha_{ik} = 1)$.

$$V_k = (M-1)^{-1} \sum_{m=1}^{M} \left(\hat{\bar{y}}_{mk} - \bar{\theta}_k\right)^2 \tag{3.8}$$

$$SE_k = \sqrt{(M-1)^{-1} \sum_{m=1}^{M} \left(\hat{\bar{y}}_{mk} - \bar{\theta}_k\right)^2} \tag{3.9}$$

$$CV_k = SE_k / \bar{\theta}_k \tag{3.10}$$

$$RB_k = 100 \times (\bar{\theta}_k - \bar{y}_k)/\bar{y}_k \tag{3.11}$$

where $\bar{y}_k = N^{-1} \sum_{i \in U} y_{ik}$.

$$RBSE_k = 100 \times SE_k/\bar{y}_k \tag{3.12}$$

with the associated 95% confidence interval for RB_k given as $RB_k \pm 1.96 \times RBSE_k$.

$$RMSE_k = \sqrt{V_k + B_k^2} \tag{3.13}$$

where B_k is the simulation bias, given by $\bar{\theta}_k - \bar{y}_k$, for the k^{th} expenditure.

$$\text{deff}_k = \frac{S_k^2/N + V_k}{S_k^2/n} \tag{3.14}$$

where N is the number of units from the analysis file, n is the average number of times the question was administered in the second interview, and S_k^2 is the element variance from the full analysis file. It is worth noting that the numerator of

equation (3.14) reflects the total sampling variance, so S_k^2/N reflects the variance attributable to the first phase of data collection (i.e., initial sample selection) while V_k reflects the variance attributable to the split questionnaire design.

We also computed three additional summary statistics for each simulation condition. They were: (1) the average number of times the question was administered in the second interview; (2) the average number of questions asked; and, (3) the average interview length. The first may be an additional criteria used to determine success of the procedure. The second and third are useful when assessing burden reduction. Finally, we computed the average sensitivity, specificity, PPV, and NPV for each simulation condition.

3.3.2.3 Results of preliminary study 2

In Table 3.12, we display the summary statistics for the 36 expenditure items that were investigated in the "one-half" simulation condition. The results from this simulation condition are what we would expect. On average, the number of times each question was asked is roughly half the number of units in the analysis file. After adjusting the weights to account for the question-asking procedure (by dividing by probability of asking the question) we obtain design-unbiased estimates of mean quarterly expenditures. TX4A_4 (modem purchase, apps, ringtones) has the largest relative bias at 1.08% and may have the strongest indication for potential bias, but all 95% confidence intervals for the relative bias calculations include zero. It is worth noting, however, that TX4A_4 corresponds to the expenditure with the

lowest second interview prevalence at 0.002. Finally, we observe that all of the design effects are around one. This is consistent with what we would expect since the split questionnaire for the second interview is based on a completely random question asking procedure with no prior information being used in the design so nothing is either gained or lost under this design.

Expenditure	Asked	Mean	Variance	Std Err	Sim CV	Rel Bias	Rel Bias SE	RMSE	Bias LB	Bias UB	deff
TX2	5,250	628.48	136.37	11.68	0.02	0.04	1.86	11.68	-3.60	3.68	0.96
TX3F	5,248	215.96	177.68	13.33	0.06	-0.02	6.17	13.33	-12.12	12.07	1.00
TX3H	5,246	103.82	46.84	6.84	0.07	-0.30	6.57	6.85	-13.18	12.58	2.66
TX4A_1	5,250	350.16	5.81	2.41	0.01	0.03	0.69	2.41	-1.32	1.37	0.98
TX4A_2	5,248	27.56	0.22	0.47	0.02	0.02	1.70	0.47	-3.31	3.36	0.97
TX4A_3	5,246	23.82	0.40	0.64	0.03	-0.08	2.67	0.64	-5.31	5.15	0.96
TX4A_4	5,245	0.11	0.00	0.03	0.27	1.08	27.27	0.03	-52.38	54.53	1.06
TX4B	5,247	4.97	0.05	0.22	0.04	0.18	4.43	0.22	-8.50	8.87	1.00
TX4C	5,249	61.69	0.46	0.68	0.01	-0.03	1.10	0.68	-2.18	2.13	0.99
TX4D	5,247	608.45	16.09	4.01	0.01	0.03	0.66	4.02	-1.26	1.32	0.98
TX5A_1	5,249	6.39	0.77	0.88	0.14	0.65	13.81	0.88	-26.41	27.71	1.03
TX5A_2	5,246	1.41	0.05	0.21	0.15	-0.66	15.09	0.21	-30.23	28.91	1.05
TX5B	5,249	357.85	246.64	15.70	0.04	-0.15	4.38	15.71	-8.74	8.44	0.98
TX6A	5,251	38.66	5.01	2.24	0.06	0.13	5.80	2.24	-11.23	11.50	1.00
TX6B	5,247	183.35	17.52	4.19	0.02	-0.18	2.28	4.20	-4.65	4.28	1.00
TX7A	5,245	16.30	0.49	0.70	0.04	-0.19	4.27	0.70	-8.56	8.19	1.02
TX8	5,247	131.33	14.91	3.86	0.03	-0.01	2.94	3.86	-5.77	5.75	0.95
TX9A	5,248	203.01	8.55	2.92	0.01	0.02	1.44	2.92	-2.81	2.84	1.01
TX9B	5,249	36.20	1.33	1.15	0.03	0.10	3.19	1.15	-6.16	6.35	0.98
TX9C	5,247	3.91	0.06	0.24	0.06	0.18	6.10	0.24	-11.78	12.14	1.03
TX9D	5,249	3.90	0.07	0.26	0.07	0.09	6.56	0.26	-12.77	12.94	1.05
TX10C_1	5,243	20.16	1.33	1.15	0.06	0.23	5.73	1.15	-11.00	11.45	1.01
TX10C_23	5,249	50.01	24.59	4.96	0.10	0.17	9.93	4.96	-19.30	19.64	0.97
TX11B	5,247	693.19	464.64	21.56	0.03	-0.02	3.11	21.56	-6.11	6.07	0.99
TX12A	5,249	165.79	10.09	3.18	0.02	-0.02	1.92	3.18	-3.78	3.73	0.95
TX12B	5,245	29.82	0.48	0.69	0.02	-0.02	2.32	0.69	-4.56	4.53	0.98
TX12C_1TO9	5,249	24.46	0.54	0.73	0.03	0.02	2.99	0.73	-5.85	5.89	0.98
TX12C_10	5,247	188.49	2.45	1.56	0.01	0.00	0.83	1.56	-1.62	1.63	0.98
TX13B	5,250	365.31	21.29	4.61	0.01	0.04	1.26	4.62	-2.44	2.52	1.00
TX14B	5,248	61.69	3.49	1.87	0.03	0.06	3.03	1.87	-5.88	5.99	1.01
TX15A	5,246	245.75	21.69	4.66	0.02	0.10	1.90	4.66	-3.61	3.82	1.01
TX16A	5,246	229.91	85.22	9.23	0.04	0.00	4.02	9.23	-7.87	7.87	0.99
TX17A	5,245	51.72	2.06	1.44	0.03	0.11	2.78	1.44	-5.34	5.56	0.99
TX18A	5,249	9.83	0.98	0.99	0.10	-0.07	10.05	0.99	-19.76	19.62	1.02
TX19A	5,246	228.35	18.43	4.29	0.02	0.09	1.88	4.30	-3.60	3.78	1.01
TX19B	5,250	369.57	34.59	5.88	0.02	0.01	1.59	5.88	-3.11	3.13	0.95

Table 3.12: Simulation summary statistics for one-half condition

In Table 3.13, we present the calculations for each of the four epidemiological criteria and for reference, the prevalence of purchasing the expenditure during the reference period of the second interview. These findings are consistent with the discussion in Section 2.6.2. Specifically, for a completely random question asking procedure in which we essentially "flip a coin" to determine whether or not to ask a particular question, then on average, sensitivity and specificity will be around 0.5, PPV will be equal to the prevalence, and NPV will equal one minus the prevalence (see Tables 2.5 and 2.6).

In Table 3.14, we display simulation summary statistics pertaining to the ARD condition. The results of this condition are somewhat surprising given that it was deemed the best from the previous simulation (Gonzalez and Eltinge, 2008). In keeping with the discussion in Section 2.4.4, this updated simulation demonstrates why we may need to modify an initial subsampling strategy to meet certain objectives or satisfy constraints. Specifically, if we examine the second column of Table 3.14 we see that, on average, each question was only asked about 100 to 120 times. We also observe that the design effects are quite large and range from greater than 1 to over 20. This indicates a substantial loss in precision due to the split questionnaire design. This may be a result of the question asking procedure attempting to only ask those sample units who are very different from the mean about the expenditure. Another contributing factor may be due to the fact that the sampling weights for the "deviant" sampling units became quite variable once they were adjusted for the subsampling procedure.

It appears that we obtain biased estimates of mean quarterly expenditures us-

Expenditure	Sensitivity	Specificity	PPV	NPV	$P(Int2)$
TX2	0.500	0.500	0.280	0.720	0.280
TX3F	0.500	0.500	0.047	0.953	0.047
TX3H	0.499	0.500	0.056	0.944	0.056
TX4A_1	0.500	0.500	0.925	0.075	0.925
TX4A_2	0.500	0.500	0.268	0.732	0.268
TX4A_3	0.499	0.500	0.130	0.870	0.130
TX4A_4	0.501	0.500	0.002	0.998	0.002
TX4B	0.500	0.500	0.083	0.918	0.083
TX4C	0.500	0.500	0.652	0.348	0.652
TX4D	0.500	0.501	0.921	0.080	0.921
TX5A_1	0.500	0.500	0.018	0.982	0.018
TX5A_2	0.499	0.500	0.009	0.991	0.009
TX5B	0.500	0.500	0.181	0.818	0.182
TX6A	0.501	0.500	0.061	0.940	0.060
TX6B	0.500	0.500	0.439	0.561	0.439
TX7A	0.500	0.500	0.098	0.902	0.098
TX8	0.500	0.500	0.370	0.630	0.370
TX9A	0.500	0.500	0.669	0.331	0.669
TX9B	0.500	0.500	0.227	0.773	0.227
TX9C	0.501	0.500	0.068	0.932	0.068
TX9D	0.501	0.500	0.064	0.936	0.064
TX10C_1	0.500	0.500	0.040	0.960	0.040
TX10C_23	0.501	0.500	0.017	0.983	0.017
TX11B	0.500	0.500	0.178	0.822	0.178
TX12A	0.500	0.499	0.547	0.453	0.547
TX12B	0.500	0.500	0.279	0.721	0.279
TX12C_1TO9	0.500	0.500	0.272	0.728	0.272
TX12C_10	0.500	0.500	0.892	0.108	0.892
TX13B	0.500	0.500	0.650	0.351	0.649
TX14B	0.500	0.500	0.268	0.732	0.268
TX15A	0.500	0.500	0.616	0.384	0.616
TX16A	0.500	0.500	0.231	0.769	0.231
TX17A	0.500	0.500	0.334	0.666	0.334
TX18A	0.501	0.500	0.030	0.970	0.030
TX19A	0.500	0.501	0.675	0.325	0.675
TX19B	0.500	0.500	0.610	0.390	0.610

Table 3.13: Epidemiological criteria for one-half condition

ing this sampling strategy (despite adjusting the weights to account for the question-asking procedure) since many of the relative bias calculations are quite large. However, their associated 95% confidence intervals all include zero, so we cannot conclude that the biases are statistically different from zero. We note that the number of times that different items are asked could be increased beyond the observed range of $100 - 200$, displayed in Table 3.14, by multiplying the p_{ik} (under ARD) by some constant greater than 1. We did not explore that option here. It might also be worth considering rescaling the expenditure values, say using a natural log transformation (after accounting for the zero-dollar expenditure reports). The purpose of this transformation would be to stabilize the variance (Meyers, 1990) as the variance of expenditure data might increase as the mean expenditure value increases.

Expenditure	Asked	Mean	Variance	Std Err	Sim CV	Rel Bias	Rel Bias SE	RMSE	Bias LB	Bias UB	deff
TX2	110	673.27	97,226.90	311.81	0.46	7.17	49.63	315.05	-90.11	104.45	6.91
TX3F	111	227.04	35,693.53	188.93	0.83	5.11	87.46	189.25	-166.32	176.53	2.15
TX3H	112	110.76	20,857.70	144.42	1.30	6.36	138.69	144.57	-265.48	278.20	20.59
TX4A_1	101	339.90	13,156.14	114.70	0.34	-2.90	32.76	115.15	-67.12	61.32	20.85
TX4A_2	112	26.69	184.78	13.59	0.51	-3.13	49.34	13.62	-99.84	93.58	8.45
TX4A_3	115	25.50	356.99	18.89	0.74	6.96	79.25	18.97	-148.38	162.30	8.96
TX4A_4	115	0.13	1.02	1.01	7.77	15.00	916.03	1.01	-1,780.43	1,810.43	12.49
TX4B	121	5.08	23.72	4.87	0.96	2.34	98.19	4.87	-190.11	194.80	5.50
TX4C	104	60.97	773.53	27.81	0.46	-1.20	45.07	27.82	-89.54	87.13	16.40
TX4D	100	585.69	41,122.62	202.79	0.35	-3.71	33.34	204.04	-69.05	61.63	23.62
TX5A_1	112	5.03	549.16	23.43	4.66	-20.83	369.04	23.47	-744.15	702.49	8.03
TX5A_2	114	0.91	28.39	5.33	5.86	-35.98	375.22	5.35	-771.41	699.45	6.78
TX5B	109	359.13	248,423.08	498.42	1.39	0.21	139.07	498.42	-272.37	272.79	10.12
TX6A	112	38.82	5,142.98	71.71	1.85	0.55	185.74	71.71	-363.50	364.60	11.02
TX6B	109	174.52	23,640.65	153.76	0.88	-4.99	83.71	154.03	-169.06	159.08	14.06
TX7A	115	16.15	570.28	23.88	1.48	-1.11	146.24	23.88	-287.74	285.51	13.24
TX8	110	127.08	27,767.35	166.64	1.31	-3.24	126.87	166.69	-251.91	245.43	17.61
TX9A	107	197.41	10,993.10	104.85	0.53	-2.75	51.65	105.00	-103.99	98.50	13.35
TX9B	112	36.64	2,035.58	45.12	1.23	1.34	124.77	45.12	-243.21	245.89	15.63
TX9C	116	3.62	65.78	8.11	2.24	-7.13	207.95	8.11	-414.72	400.46	12.77
TX9D	118	3.74	47.40	6.89	1.84	-4.22	176.54	6.89	-350.24	341.80	8.37
TX10C_1	118	22.24	219.15	14.80	0.67	10.58	73.61	14.96	-133.71	154.86	1.91
TX10C_23	117	57.26	2,691.83	51.88	0.91	14.71	103.93	52.40	-189.00	218.42	1.15
TX11B	116	756.91	164,506.74	405.59	0.54	9.17	58.50	410.55	-105.49	123.83	3.84
TX12A	113	159.03	14,299.66	119.58	0.75	-4.10	72.11	119.77	-145.44	137.23	13.64
TX12B	117	29.39	699.57	26.45	0.90	-1.47	88.67	26.45	-175.26	172.32	15.56
TX12C_1TO9	123	22.75	721.05	26.85	1.18	-6.99	109.78	26.91	-222.16	208.18	15.05
TX12C_10	102	186.95	5,553.60	74.52	0.40	-0.81	39.54	74.54	-78.31	76.69	20.97
TX13B	110	353.43	28,433.23	168.62	0.48	-3.21	46.18	169.03	-93.72	87.29	14.09
TX14B	118	65.18	1,410.93	37.56	0.58	5.72	60.93	37.73	-113.70	125.14	4.64
TX15A	111	242.40	24,706.04	157.18	0.65	-1.26	64.03	157.21	-126.75	124.23	12.37
TX16A	115	222.96	96,512.04	310.66	1.39	-3.02	135.13	310.74	-267.88	261.83	12.21
TX17A	114	50.54	1,761.37	41.97	0.83	-2.18	81.24	41.98	-161.41	157.06	9.07
TX18A	117	9.98	981.09	31.32	3.14	1.40	318.32	31.32	-622.50	625.30	11.60
TX19A	109	222.77	29,789.22	172.60	0.78	-2.36	75.65	172.68	-150.63	145.92	17.07
TX19B	112	380.82	46,513.94	215.67	0.57	3.06	58.37	215.97	-111.34	117.45	12.97

Table 3.14: Simulation summary statistics for the ARD condition

It is unlikely that a question asking procedure such as this would be implemented in practice. This is because the procedure asks each question so few times, which is surprising given that we relaxed the restriction of asking only one item per sample unit from the original simulation. On average, we only ask fewer than one question per sample unit at the second interview and the average interview time for questions about expenditures is about 0.3 minutes (see Table 3.18). Furthermore, the basis on which the subsampling procedures for each question were developed are intrinsically flawed for some expenses. Specifically, using interview one expenditure information to identify "deviant" sample units might not be a good approach because it is unlikely that for some expenditure categories, if not a majority of them, a sample unit will have the outlying expense during both subsequent quarters. For example, a sample unit who purchased a large durable good, e.g., a refrigerator, during the reference period inquired about during the first interview is unlikely to have that same "deviant" expense during the reference period inquired about at the second interview. However, based on the question asking strategy that sample unit would be more likely to get asked about those types of purchases at the second interview provided that the refrigerator purchase is outlying with respect to the mean of major household appliance expenses (TX6A).

In Table 3.15, we display the epidemiological evaluation criteria for the ARD method. This method of asking questions performs worse than the "flipping the coin" in terms of sensitivity and PPV. In fact, for all expenditures, we are actually asking the question of substantially fewer sample units with the expenditure than we would by chance. However, one potential bright spot, is that this method of

Expenditure	Sensitivity	Specificity	PPV	NPV	$P(Int2)$
TX2	0.006	0.988	0.155	0.718	0.280
TX3F	0.013	0.990	0.055	0.953	0.047
TX3H	0.007	0.989	0.037	0.944	0.056
TX4A_1	0.005	0.933	0.478	0.070	0.925
TX4A_2	0.008	0.988	0.192	0.731	0.268
TX4A_3	0.010	0.989	0.121	0.870	0.130
TX4A_4	0.014	0.989	0.003	0.998	0.002
TX4B	0.014	0.989	0.098	0.918	0.083
TX4C	0.004	0.979	0.245	0.344	0.652
TX4D	0.005	0.941	0.510	0.075	0.921
TX5A_1	0.002	0.989	0.003	0.982	0.018
TX5A_2	0.007	0.989	0.006	0.991	0.009
TX5B	0.004	0.988	0.073	0.817	0.182
TX6A	0.007	0.989	0.039	0.939	0.060
TX6B	0.005	0.985	0.214	0.558	0.439
TX7A	0.005	0.988	0.043	0.902	0.098
TX8	0.005	0.986	0.185	0.628	0.370
TX9A	0.008	0.985	0.515	0.329	0.669
TX9B	0.003	0.987	0.074	0.771	0.227
TX9C	0.004	0.988	0.023	0.931	0.068
TX9D	0.011	0.989	0.061	0.936	0.064
TX10C_1	0.026	0.989	0.091	0.961	0.040
TX10C_23	0.052	0.990	0.078	0.984	0.017
TX11B	0.006	0.988	0.096	0.821	0.178
TX12A	0.006	0.983	0.283	0.450	0.547
TX12B	0.005	0.986	0.117	0.719	0.279
TX12C_1TO9	0.005	0.986	0.120	0.726	0.272
TX12C_10	0.007	0.967	0.634	0.105	0.892
TX13B	0.003	0.976	0.199	0.346	0.649
TX14B	0.005	0.987	0.123	0.730	0.268
TX15A	0.005	0.981	0.299	0.381	0.616
TX16A	0.007	0.988	0.138	0.768	0.231
TX17A	0.004	0.986	0.115	0.664	0.334
TX18A	0.004	0.989	0.011	0.970	0.030
TX19A	0.005	0.979	0.338	0.321	0.675
TX19B	0.007	0.983	0.393	0.388	0.610

Table 3.15: Epidemiological criteria for ARD condition

asking questions does, in general, have very high specificity. This is reasonable simply because the questions are getting asked so few times and many of the second interview expenditure prevalences are quite low as well. Given the implications of the ARD method summarized in the two preceding tables, it is unlikely that this method would be feasible and practical for implementation.

In Table 3.16, we have the results from the "four-fifths" condition. First, we point out that the average number of times an expenditure question was asked can be expressed as a function of $P(Int1)$, the prevalence of the expenditure during the first interview reference period, since our decision to ask the question was based on that quantity. So, if we let n_k be the number of times that the question about expenditure k was asked, N be the number of units in the analysis file, $\alpha_{ik} = 1$ if sample unit i is asked about expenditure k and zero otherwise, $p_{ik} = P(\alpha_{ik} = 1)$, and $I(x)$ be the indicator function for event x, then the average number of times that question was asked under the split questionnaire is as follows.

$$\begin{aligned}
E_{SQ}(n_k|S) &= E_{SQ}(\sum_{i\in S}\alpha_{ik}|S) = \sum_{i\in S}E_{SQ}(\alpha_{ik}|S) = \sum_{i\in S}p_{ik} \\
&= \sum_{i\in S}\left[\frac{4}{5}I(y_{Int1,ik}>0) + \frac{1}{5}I(y_{Int1,ik}=0)\right] \\
&= \sum_{i\in S}\left[\frac{4}{5}I(y_{Int1,ik}>0)\right] + \sum_{i\in S}\left[\frac{1}{5}I(y_{Int1,ik}=0)\right] \\
&= \frac{4}{5}N\times P(Int1) + \frac{1}{5}N\times P(Int1^c) \\
&= \frac{4}{5}N\times P(Int1) + \frac{1}{5}N\left[1-P(Int1)\right] \\
E_{SQ}(n_k|S) &= \frac{1}{5}N\left[1+3P(Int1)\right] \qquad (3.15)
\end{aligned}$$

Also from Table 3.16, we observe that we obtain design unbiased estimates of mean quarterly expenditures using the simple weighting adjustment (i.e., adjusting the weight by $1/p_{ik}$). Despite containing zero, the 95% confidence intervals for the relative bias of some expenditure categories are quite large. For instance, the relative bias for TX4A_4 (modem purchase, apps, ringtones) was calculated as 3.53% and its associated confidence interval is $(-98.80, 105.85)$. This is consistent with the finding from the one-half simulation condition, namely, this expenditure had the largest relative bias estimate even though its confidence interval contained zero. This expenditure also had the widest confidence interval for the relative bias of the mean expenditure estimate for the ARD condition.

Expenditure	Asked	Mean	Variance	Std Err	Sim CV	Rel Bias	Rel Bias SE	RMSE	Bias LB	Bias UB	deff
TX2	3,849	629.09	158.29	12.58	0.02	0.14	2.00	12.61	-3.79	4.06	0.76
TX3F	2,382	216.44	126.85	11.26	0.05	0.20	5.21	11.27	-10.02	10.42	0.39
TX3H	2,462	104.20	35.60	5.97	0.06	0.07	5.73	5.97	-11.16	11.30	1.01
TX4A_1	7,862	350.00	4.29	2.07	0.01	-0.02	0.59	2.07	-1.18	1.14	1.28
TX4A_2	3,912	27.56	0.40	0.63	0.02	0.03	2.28	0.63	-4.45	4.50	1.01
TX4A_3	2,917	23.83	0.70	0.84	0.04	-0.04	3.50	0.84	-6.90	6.83	0.72
TX4A_4	2,126	0.11	0.00	0.06	0.55	3.53	52.21	0.06	-98.80	105.85	0.20
TX4B	2,536	4.97	0.11	0.33	0.07	0.19	6.57	0.33	-12.69	13.07	0.77
TX4C	6,080	61.71	0.58	0.76	0.01	0.00	1.24	0.76	-2.43	2.42	1.30
TX4D	7,869	608.26	13.43	3.66	0.01	0.00	0.60	3.66	-1.18	1.18	1.36
TX5A_1	2,251	6.33	2.68	1.64	0.26	-0.34	25.80	1.64	-50.91	50.24	1.00
TX5A_2	2,156	1.39	0.18	0.42	0.30	-1.92	29.55	0.42	-59.85	56.01	1.02
TX5B	2,924	358.17	628.36	25.07	0.07	-0.06	6.99	25.07	-13.77	13.65	0.96
TX6A	2,368	38.59	18.72	4.33	0.11	-0.06	11.21	4.33	-22.03	21.90	1.07
TX6B	4,177	183.44	38.31	6.19	0.03	-0.13	3.37	6.19	-6.73	6.47	1.27
TX7A	2,513	16.35	1.40	1.18	0.07	0.13	7.24	1.18	-14.05	14.32	0.95
TX8	3,868	131.35	34.38	5.86	0.05	0.01	4.46	5.86	-8.74	8.76	1.13
TX9A	5,528	203.10	12.91	3.59	0.02	0.06	1.77	3.60	-3.41	3.53	1.34
TX9B	3,156	36.07	3.79	1.95	0.05	-0.26	5.39	1.95	-10.81	10.30	1.12
TX9C	2,365	3.90	0.20	0.44	0.11	0.12	11.39	0.44	-22.21	22.44	1.02
TX9D	2,384	3.90	0.15	0.39	0.10	0.10	10.08	0.39	-19.65	19.85	0.76
TX10C_1	2,332	20.07	0.70	0.84	0.04	-0.18	4.17	0.84	-8.35	7.99	0.34
TX10C_23	2,193	50.06	19.71	4.44	0.09	0.29	8.89	4.44	-17.14	17.72	0.37
TX11B	3,165	694.18	374.24	19.35	0.03	0.12	2.79	19.36	-5.35	5.59	0.54
TX12A	4,318	165.51	26.63	5.16	0.03	-0.19	3.11	5.17	-6.29	5.91	1.38
TX12B	3,052	29.81	1.87	1.37	0.05	-0.08	4.58	1.37	-9.07	8.90	1.38
TX12C_1TO9	2,179	24.38	2.15	1.47	0.06	-0.33	5.99	1.47	-12.07	11.41	1.00
TX12C_10	7,714	188.49	1.80	1.34	0.01	0.00	0.71	1.34	-1.39	1.40	1.25
TX13B	5,418	365.20	37.82	6.15	0.02	0.01	1.68	6.15	-3.29	3.31	1.44
TX14B	3,347	61.66	4.61	2.15	0.04	0.01	3.48	2.15	-6.81	6.84	0.75
TX15A	5,131	245.79	28.25	5.32	0.02	0.12	2.17	5.32	-4.12	4.36	1.14
TX16A	3,238	230.57	190.24	13.79	0.06	0.29	6.00	13.81	-11.47	12.05	0.99
TX17A	3,702	51.65	3.93	1.98	0.04	-0.02	3.84	1.98	-7.55	7.50	1.01
TX18A	2,221	9.85	3.23	1.80	0.18	0.13	18.27	1.80	-35.67	35.94	0.94
TX19A	5,865	228.03	18.01	4.24	0.02	-0.05	1.86	4.25	-3.70	3.59	1.11
TX19B	5,251	369.72	53.75	7.33	0.02	0.05	1.98	7.33	-3.83	3.94	1.20

Table 3.16: Simulation summary statistics for four-fifths condition

When assessing the design effects, this condition indicates gains in precision over a completely random question asking procedure for about 14 of the expenditure categories. This finding suggests that if we stratify the sample into two strata – those that purchased the item and reported it during the first interview and those that did not – and then ask the question about the item with a higher probability to those that purchased it, then we can potentially achieve gains in efficiency over a standard split questionnaire design with the same sample size receiving the particular question. This is an encouraging finding since this method is most similar to methods we develop in the next chapter of this dissertation. Furthermore, this might be an attractive method to utilize given its ease in implementation, i.e., the only information requirement is whether or not the sample unit purchased the item.

In Table 3.17, we display the epidemiological evaluation criteria for the four-fifths condition. About half of the expenditures had sensitivity values above 0.5 and all had specificity values greater than 0.5. All of the PPV calculations were greater than the associated prevalence of the expenditure so that suggests that this method of asking questions performs no worse than "flipping the coin" in correctly asking the unit about the expenditure. This latter finding suggests that under this method, we are detecting the purchase better than we would expect simply by an equal chance of asking the question.

Finally, in Table 3.18, we present a comparison of the simulation conditions. We provide the average number of questions asked, the mean interview length, the percent reduction in interview length from the full instrument, and the average design effect for each condition. This last value gives a sense of the average design

Expenditure	Sensitivity	Specificity	PPV	NPV	$P(Int2)$
TX2	0.774	0.792	0.591	0.900	0.280
TX3F	0.727	0.798	0.150	0.984	0.047
TX3H	0.719	0.794	0.172	0.979	0.056
TX4A_1	0.779	0.620	0.962	0.185	0.925
TX4A_2	0.668	0.735	0.480	0.858	0.268
TX4A_3	0.641	0.776	0.300	0.935	0.130
TX4A_4	0.301	0.798	0.003	0.998	0.002
TX4B	0.460	0.778	0.157	0.941	0.083
TX4C	0.736	0.714	0.828	0.591	0.652
TX4D	0.781	0.614	0.959	0.195	0.921
TX5A_1	0.256	0.786	0.021	0.983	0.018
TX5A_2	0.284	0.795	0.013	0.992	0.009
TX5B	0.385	0.745	0.251	0.845	0.182
TX6A	0.242	0.775	0.065	0.941	0.060
TX6B	0.481	0.667	0.530	0.621	0.439
TX7A	0.345	0.772	0.141	0.916	0.098
TX8	0.456	0.683	0.459	0.681	0.370
TX9A	0.583	0.586	0.740	0.410	0.669
TX9B	0.441	0.741	0.333	0.819	0.227
TX9C	0.290	0.779	0.088	0.938	0.068
TX9D	0.361	0.782	0.102	0.947	0.064
TX10C_1	0.760	0.800	0.136	0.988	0.040
TX10C_23	0.740	0.800	0.059	0.994	0.017
TX11B	0.773	0.800	0.455	0.942	0.178
TX12A	0.462	0.649	0.614	0.500	0.547
TX12B	0.305	0.715	0.292	0.726	0.279
TX12C_1TO9	0.212	0.794	0.278	0.730	0.272
TX12C_10	0.782	0.654	0.949	0.266	0.892
TX13B	0.612	0.662	0.770	0.480	0.649
TX14B	0.549	0.765	0.461	0.822	0.268
TX15A	0.590	0.673	0.743	0.506	0.616
TX16A	0.520	0.755	0.389	0.840	0.231
TX17A	0.510	0.726	0.483	0.748	0.334
TX18A	0.293	0.791	0.042	0.973	0.030
TX19A	0.657	0.646	0.794	0.476	0.675
TX19B	0.617	0.683	0.753	0.533	0.610

Table 3.17: Epidemiological criteria for the four-fifths condition

effect taking into consideration all expenditures and giving them equal importance. We also provide reference values on these metrics from the full analysis file in the last row of the table.

Condition	Asked	Length	% Reduction	Design effect	
				Average	Median
One-half	18.00	19.62	50.00	1.04	1.00
ARD	0.28	0.31	99.21	11.75	12.43
Four-fifths	14.94	16.29	58.49	0.98	1.01
Full analysis file	36.00	39.24	...	...	...

Table 3.18: Comparison of simulation conditions for preliminary study 2

For the "one-half" condition, the calculations for the average number of questions asked and the average interview length conform with theory. Specifically, we reduce the number of questions asked by half, from 36 to 18, and decrease the average interview time from about 40 minutes to 20 minutes. For the "four-fifths" condition, we reduced the interview length by about 58% or a savings of 23 minutes. We also reduced the average number of questions asked from 36 to about 15. It is also worth reiterating the somewhat surprising result associated with the ARD condition, namely, on average, we ask fewer than one expenditure question to each sample member. This occurs despite relaxing the constraint from the original simulation in which we only administered one question to each sample member. This finding may suggest that sample units only tend to be "different" (or have a large deviation) with respect to one expenditure category while for the remaining expenditure categories, they are effectively similar to the average expense incurred.

In the last column of Table 3.18, we have the average design effect under

each method. As mentioned, this takes into consideration all of the expenditure categories and gives them equal importance. It is also important to reiterate that the variance in the denominator of the design effect is for a SRS sample of size equal to the average number of times that a given method samples an expenditure for the second interview. On average, we see that the design effects associated with the mean quarterly expenditures for the "one-half" condition were all around 1. For the "four-fifths" condition, we see that the design effects were, on average, slightly less than 1. For the ARD condition, we did far worse in terms losses in efficiency. Again we conclude that this latter method of asking questions is not feasible for implementation. Given the greater reduction in burden, in terms of both interview minutes and number of questions asked, and a slightly lower design effect associated with the "four-fifths" condition compared to the "one-half", we are optimistic that split questionnaire methods utilizing a sample unit's prior interview information in the design have the potential to improve the split questionnaires' efficiency.

Using the method of graphical display presented in Section 2.6.2 (see Figure 2.5), we compare the three simulation conditions on the basis of the epidemiological criteria in Figures 3.2–3.5. With these figures, we assess the responsive capabilities of the simulation conditions relative to each other. Using the "four-fifths" condition as the reference, with the exception of TX5A_1 (construction materials for specific jobs not started), TX6A (major appliance installation, cost, and rental), TX12B (vehicle license fees), and TX12C_1TO9 (other vehicle fuels), the "four-fifths" condition outperforms the other two conditions. With respect to these four expenditure categories, there appears to be no substantial improvement in the

responsive capabilities over the "one-half" condition.

It is worth noting that none of these four expenditures were classified as recurrent (see Table 3.8) and with the exception of TX12B (vehicle license fees), these expenditures were also classified as rare. Although TX12B is not considered rare, it does have a relatively low $P(Int1)$ value of 0.152 which is on the cusp of the criterion used for classifying an expenditure as rare.

3.3.2.4 Implications of preliminary study 2

In summary, preliminary study 2 highlights some important considerations for responsive split questionnaire designs. We verified that obtaining design unbiased estimates of mean quarterly expenditures is possible in a split questionnaire even when prior information is incorporated into the design. However, there were potential gains and losses in precision. The losses in precision were most substantial for the ARD condition. This latter finding was surprising given that the previous study of Gonzalez and Eltinge (2008) declared this the preferred method of asking questions in the second interview (among those considered). This finding also suggests that using different evaluation criteria can lead to different design decisions so it is important to consider a wide range of strategies and metrics to better understand the tradeoffs under each type of split questionnaire design.

The second point is that depending on the question asking procedure some survey objectives might not be met. It is unlikely that asking a question, on average, 110 times of every 10,000 sample units will yield the necessary levels of precision

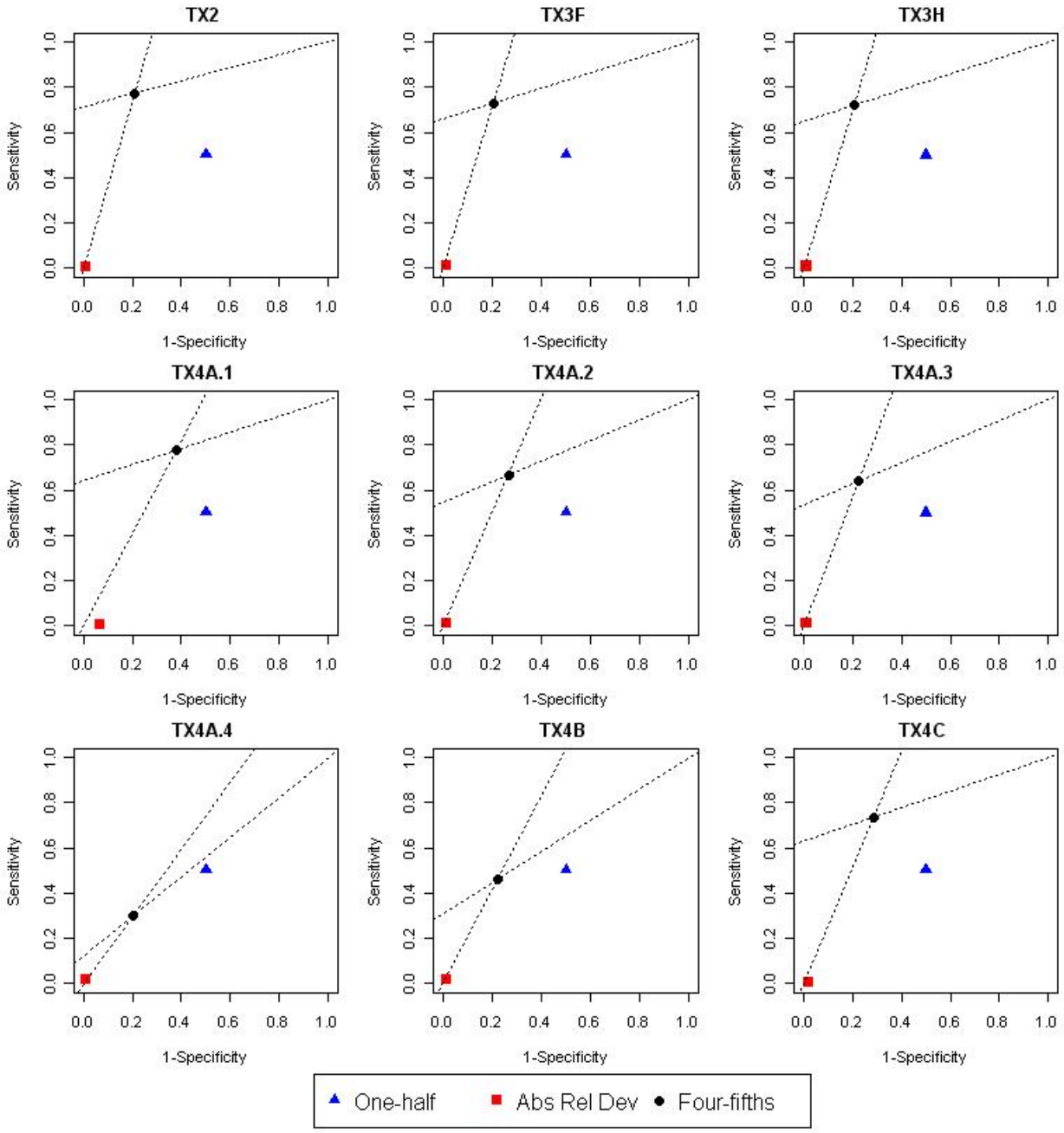

Figure 3.2: Diagnostic test comparisons for preliminary study conditions

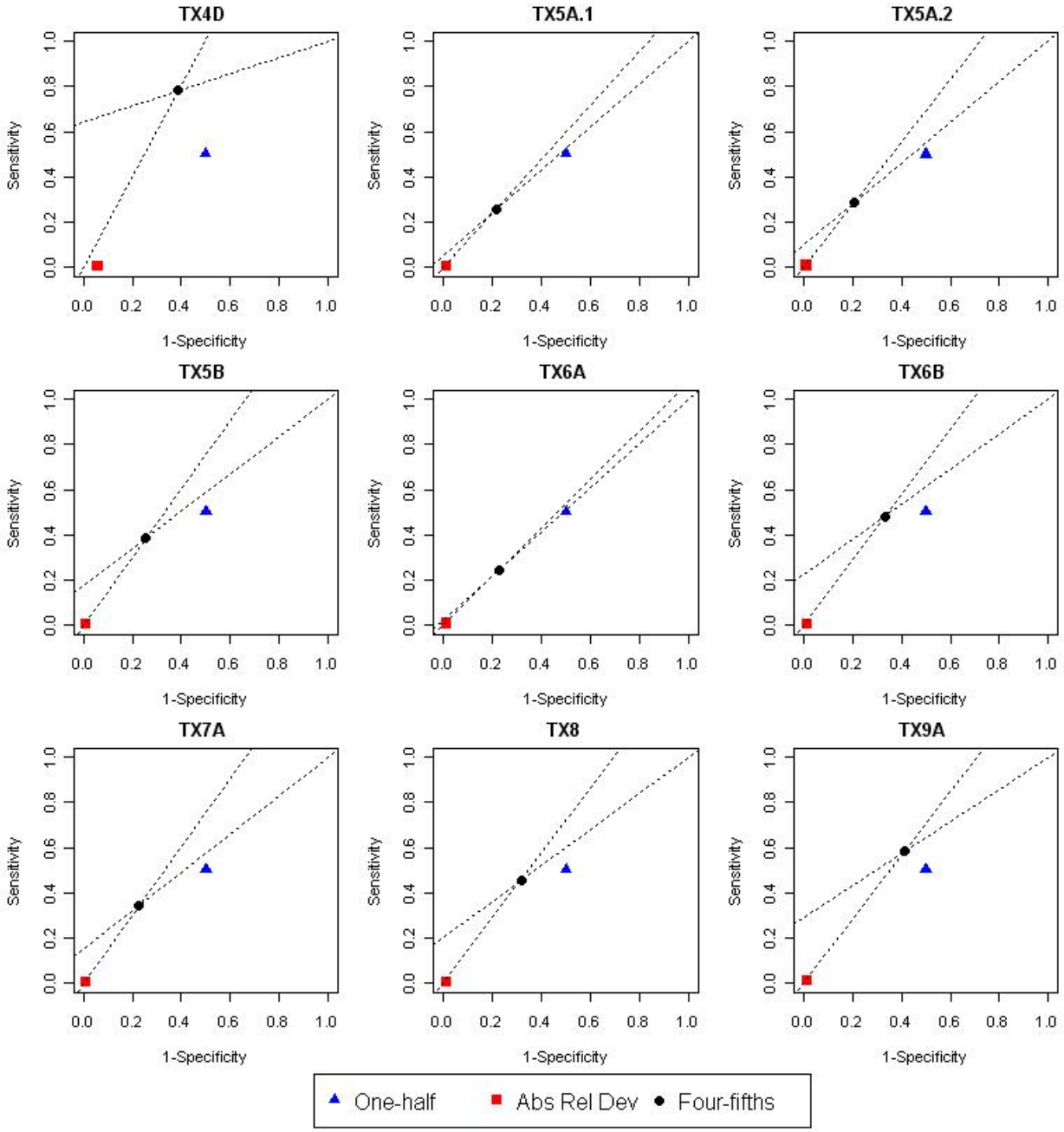

Figure 3.3: Diagnostic test comparisons for preliminary study conditions (2)

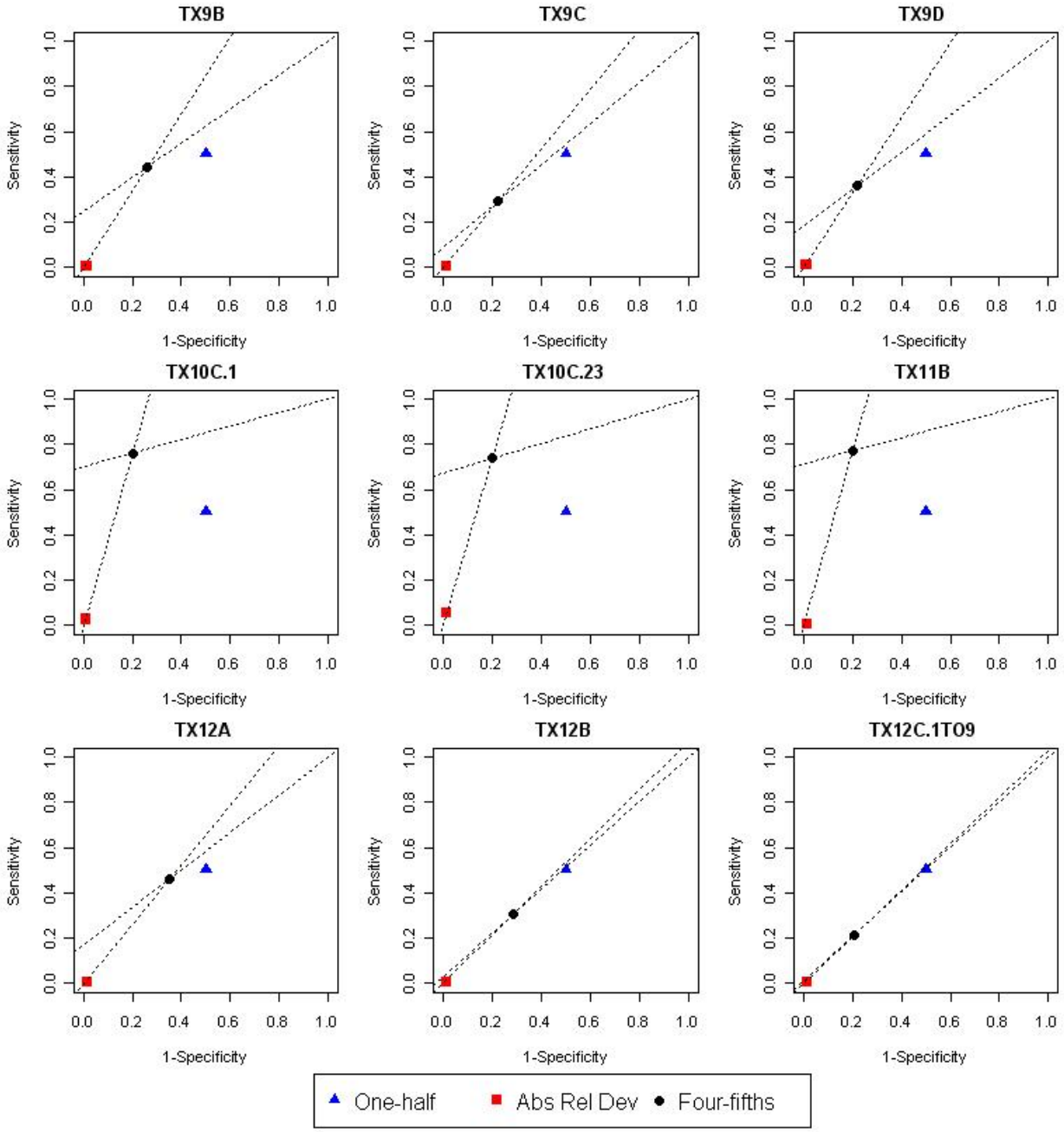

Figure 3.4: Diagnostic test comparisons for preliminary study conditions (3)

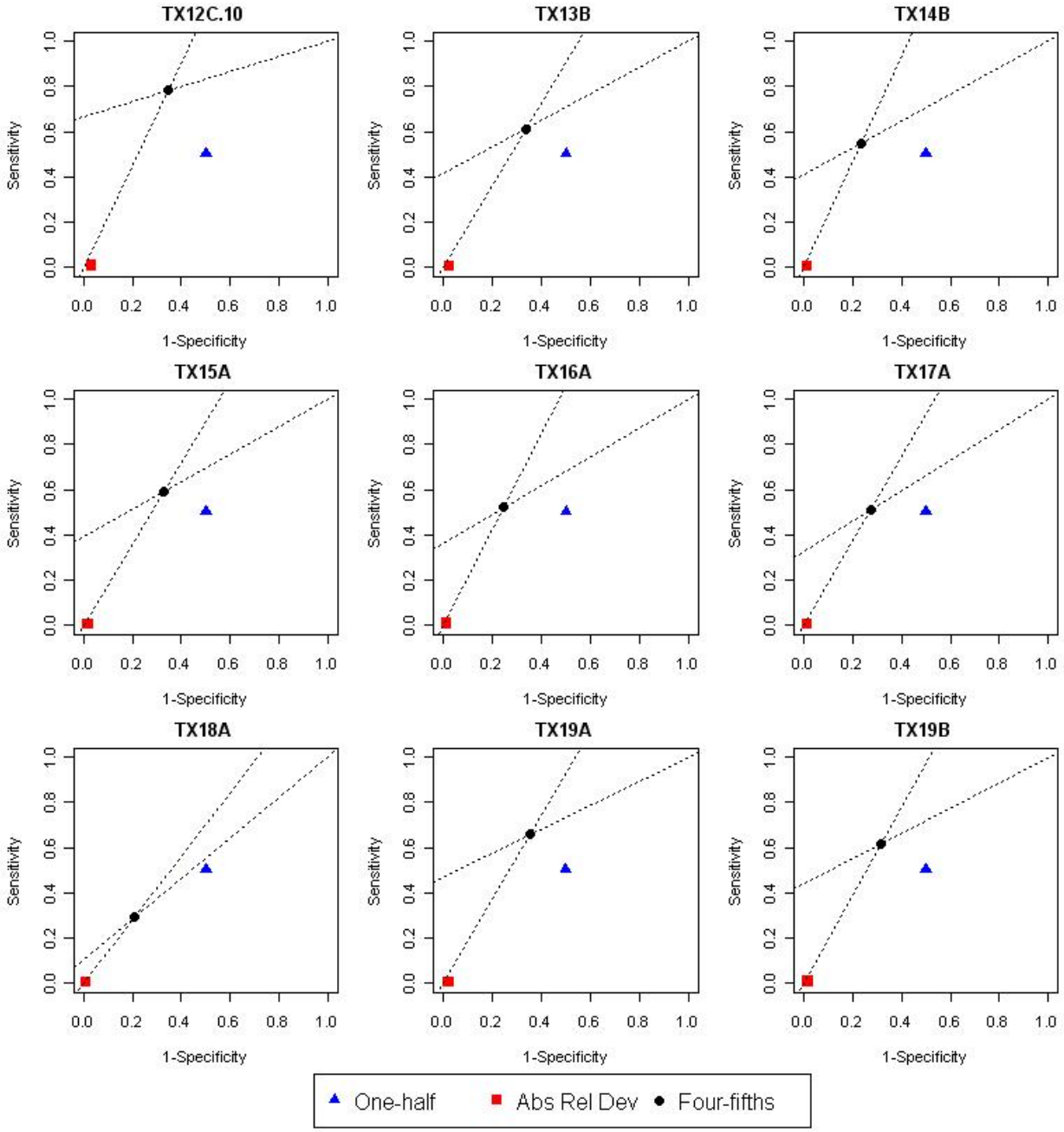

Figure 3.5: Diagnostic test comparisons for preliminary study conditions (4)

for expenditure data. This is because most expenses are incurred with such low frequency and estimates of characteristics of those expenses have a low signal-to-noise ratio. Therefore, careful consideration needs to be given to both the prevalence of the expense and the CVs of estimates about those expenses when determining which responsive split questionnaire design is preferred.

The final point and perhaps the one with most bearing on this research, is as demonstrated in the "four-fifths" simulation condition, the simple strategy of stratifying the sample units into two groups – those who incurred the expense in the first interview reference period and those who did not – and asking questions about the expense of the former group with a higher probability has the potential to yield gains in efficiency over a completely random question asking procedure of the same size. Furthermore, this type of stratification is also more successful at customizing the survey to the individual respondent and can be effective for some rare events. Obviously, this type of strategy might not be as effective for some expenditures, e.g., expenses that are rare and not recurrent, but this simulation study does lend substantial credence to the idea of incorporating a prior information on the sample unit into the decision about whether to ask a particular question to the unit. Armed with this knowledge, we want to improve upon this type of split questionnaire design by developing a more optimal responsive split questionnaire design. This is the primary goal of the next chapter.

Chapter 4

Methods

4.1 Overview

In this chapter, we explore a variety of techniques for incorporating prior information about the sample unit into the decision rules for a responsive split questionnaire in a panel survey. We also employ mathematical programming methods to develop decision rules for a responsive split questionnaire when constraints are specified. In addition, we describe some theoretical properties of each set of decision rules. Finally, we evaluate the performance of the decision rules from each method by simulating hypothetical scenarios of collected data.

4.2 Probability proportional-to-size using first interview information

The first method we use to develop decision rules for a responsive split questionnaire design is probability proportional-to-size (PPS) sampling. PPS sampling represents a classical approach to sampling in which the selection probabilities are chosen to minimize the sampling variances of the desired estimates (Lohr, 1999). Because this classical approach is thought to produce estimates with desirable properties (e.g., low sampling variance), it is a reasonable first method to consider.

The rationale for designing selection probabilities in a standard PPS design

stems from estimating population totals from a sample. Oftentimes, totals in primary sampling units (PSU) are related to the number of units in the PSU. Selection probabilities for the PSU are then related to the relative size of the PSU. Thus, a large PSU has a greater chance of being selected for the sample than a small PSU. Drawing the link to our research, if we were interested in estimating total expenditures for a sample unit (which may, in fact, be an ancillary estimation objective), then we may want to assign higher probabilities of selection for expenditures that comprise a larger proportion of a sample unit's total expenditures.

Assigning higher probabilities of selection following a PPS sampling scheme may still coincide with the main goal of a responsive split questionnaire, e.g., ask sample units about expenses that they are likely to incur. Obviously, if an expense comprises a larger proportion of the total expenses for that CU, then the CU incurs it. However, this method may place greater emphasis on larger, in terms of dollar amount, expenses as opposed to any expense incurred for that category by the CU. Thus, a potential deficiency of this method with respect to our research objective (and a similar deficiency to the ARD condition) may be that it will assign low probabilities of getting asked the question to incurred expenses that comprise a smaller proportion of the total expenditures for that CU when we want the assignment probabilities to be high for any expense likely to be incurred. Nonetheless, this method is worth considering due to its desirable properties for some key survey estimates. As a final note, we refer to this as the **PPS** method.

4.2.1 Statement of the problem

To develop decision rules following a PPS sampling design, we need a measure-of-size (MOS) to form our question assignment probabilities. We can collect the MOS during the initial phase of data collection (e.g., first interview). We propose using the reported expenditure amount from the first interview, denoted as $y_{Int1,ik}$, as the MOS. We believe this is an appropriate MOS given the results from Section 3.3.1.2. In that section, we demonstrated that for some expenditure categories incurring and the amount of the expense were correlated across the two successive reference periods. The implicit assumption of using $y_{Int1,ik}$ as the MOS is that $y_{Int1,ik} \propto y_{Int2,ik}$, where $y_{Int2,ik}$ denotes the i^{th} sample unit's expenditure on item k for the second interview reference period. In other words, for this method to be successful $y_{Int1,ik}$ needs to be a good proxy for the expense incurred in the second interview reference period.

To discuss the sampling properties of this method, we define the following notation.

- $p_{ik} = (\sum_k y_{Int1,ik})^{-1}(y_{Int1,ik})$ where $y_{Int1,ik}$ is the i^{th} sample unit's reported expense for item k in the first interview and p_{ik} is the decision rule for asking about item k to the i^{th} sample unit during the second interview. We note that the summation in the denominator of p_{ik} is over all expenditures for sample unit i only.

- $\alpha_{ik} = 1$ if $i \in S$ receives k and 0 otherwise, where S is the set of sample units for which the split questionnaire is being administered

It is essential to point out that a slight modification was made to the p_{ik} to keep this sampling design *measurable*. If we let π_i be the first-order inclusion probabilities for the i^{th} unit into the sample and π_{jl} be the joint inclusion probabilities of units j and l (for $j \neq l$) into the sample, then a sampling design is said to be *measurable* if the following two conditions are satisfied: (1) $\pi_i > 0$ for every $i \in U$ (where U denotes the "population") and (2) $\pi_{jl} > 0$ for every $j \neq l \in U$. In words, these conditions mean the following: (1) every member of the population has a non-zero probability of being included in the sample and (2) every distinct pair of members in the population has a non-zero probability of being included in the sample. Satisfying the conditions for measurability ensures that survey analysts can calculate valid design-based variance estimates and valid confidence intervals based on the observed survey data (Särndal et al., 1992).

Since $y_{Int1,ik}$ may equal zero for many items and many sample units, we would have the situation that $p_{ik} = 0$ and our responsive split questionnaire would violate the first condition of a measurable sampling design. Thus, we would need to modify the instances in which $p_{ik} = 0$. We do this in the following way. First, if we believe that the probability that the units with $p_{ik} = 0$ will incur expense k is very low (if not zero), then we want p_{ik} to be as small as possible. Thus, for any p_{ik} that was less than 0.005, we set it equal to 0.005 (an arbitrarily chosen low value)[1]. By choosing this low value, we maintain the integrity of the question-asking procedure since we still ask about that expenditure to the sample unit with very low probability

[1]In Table D.1 of Appendix D.1, we present descriptive statistics for the decision rules under the PPS method before and after the modification was imposed.

because we believe it is unlikely to incur the expense. Furthermore, we ensure that the responsive split questionnaire design is measurable.

The sampling properties of the PPS responsive split questionnaire design are as follows. First, if we let $n_{PPS,k}$ be the number of sample units receiving the question about expense k, then the expectation and variance of $n_{PPS,k}$ are given in equations (4.1) and (4.2), respectively, assuming that Poisson sampling (according to p_{ik}) is used to select the items.

$$E_{SQ}(n_{PPS,k}|S) = E_{SQ}(\sum_{i\in S}\alpha_{ik}|S) = \sum_{i\in S}E_{SQ}(\alpha_{ik}|S) = \sum_{i\in S}p_{ik} \tag{4.1}$$

$$V_{SQ}(n_{PPS,k}|S) = V_{SQ}(\sum_{i\in S}\alpha_{ik}|S) = \sum_{i\in S}V_{SQ}(\alpha_{ik}|S) = \sum_{i\in S}p_{ik}(1-p_{ik}) \tag{4.2}$$

Furthermore, from Section 2.4.2.4, we see that the responsive split questionnaire estimator of the full sample mean is approximately design-unbiased. In particular, the expectation and variance of $\hat{\bar{y}}_{SQ,k}$ with respect to the PPS responsive split questionnaire is given in (4.3) and (4.4).

$$\begin{aligned} E_{SQ}(\hat{\bar{y}}_{SQ,k}|S) &\approx E_{SQ}\left[\hat{\bar{y}}_k + (\sum_{i\in S}w_i)^{-1}\left\{\sum_{i\in S}\alpha_{ik}w_iy_{ik}/p_{ik} - \hat{\bar{y}}_k\sum_{i\in S}\alpha_{ik}w_i/p_{ik}\right\}|S\right] \\ &= \hat{\bar{y}}_k \end{aligned} \tag{4.3}$$

$$\begin{aligned} V_{SQ}(\hat{\bar{y}}_{SQ,k}|S) &\approx (\sum_{i\in S}w_i)^{-2}\left[\sum_{i\in S}\left(\frac{1-p_{ik}}{p_{ik}}\right)w_i^2(y_i^2+\hat{\bar{y}}_k^2) - 2\hat{\bar{y}}_k\sum_i\sum_{j>i}y_{ik}w_iw_j\left(\frac{p_{ijk}-p_{ik}p_{jk}}{p_{ik}p_{jk}}\right)\right] \\ &= (\sum_{i\in S}w_i)^{-2}\left[\sum_{i\in S}\left(\frac{1-p_{ik}}{p_{ik}}\right)w_i^2(y_{ik}-\hat{\bar{y}}_k)^2\right] \end{aligned} \tag{4.4}$$

Finally, the expectation of (4.4) with respect to the original sample selection reflects the added contribution of the PPS responsive split questionnaire design to the overall sampling variance.

4.2.2 Simulation setup

Using the same setup and infrastructure developed for Preliminary Study 2, we carried out a simulation with $(M =)1,000$ iterations to evaluate the performance of this method. For each iteration, we randomly "asked" sample units questions based on their respective p_{ik} values described in the previous section. We then computed the quantities given in equations (3.7) – (3.14) to summarize the simulation. We also computed similar summary statistics for the domains defined by the CUs that incurred the expense. We computed the average number of times the question was administered in the second interview, the average number of questions asked, the average interview length, the average time spent answering a question, percent reduction in interview length, and the average and median design effects for the estimated mean quantities. Finally, we computed the average sensitivity, specificity, PPV, and NPV for this method.

4.2.3 Results

In Table 4.1, we display the summary statistics for the PPS responsive split questionnaire method. We observe the following trends from this method. In general, as the mean of the expenditure increases, the average number of times the

question about the expenditure gets asked also increases. This makes intuitive sense since the method assigns a higher probability of asking about the expense to categories that comprise a larger proportion of the total expenses for that CU. For roughly one-third of the expenditure categories, we ask an average of fewer than 100 sample units about the expense and for only two expenditure categories, we ask more than 1,000 sample units about the expense. Given that only two expenditure categories have prevalence rates of less than 0.01 and 14 were classified as rare (i.e., the prevalence of incurring the expense was less than 0.1), we can infer that this method likely does a poor job of asking about the expense given that the CU incurred it. If this was not the case, then fewer categories would be asked to fewer than 100 sample units. Furthermore, more than two categories would be asked to more than 1,000 sample units. We verify this conclusion when we examine the epidemiological criteria for this method in Table 4.3.

Although the magnitudes of some of the relative biases for the mean expenditure estimates appear high, all 95% confidence intervals associated with the relative bias calculations include zero therefore, we cannot conclude that the biases are statistically different from zero. With the exception of TX4A_1 (telephone services), TX4D (utilities), and TX12C_10 (average monthly gas expenditures), all simulation CVs are greater than 0.1 and range upward to 3.86. The expenditures with CVs less than 0.1 correspond to the expenditures with highest prevalence rates (all around or above 90%), while the highest simulation CVs correspond to the expenditures with the lowest prevalence rates. These are TX4A_4 (modem purchases, apps, ringtones) and TX5A_2 (construction materials for general jobs) with CVs of 3.86 and 2.05,

respectively. The large CVs are consistent with the original data (see Table 3.6), as expenditure data seem to have a high noise-to-signal ratio.

Finally, in the last column of Table 4.1, we observe that the design effects for the means range from 0.21 to 5.8 with seven mean expenditure estimates exhibiting design effects of less than 1. This suggests that for those expenses, we have achieved a gain in precision relative to a split questionnaire design in which we essentially "flip a coin" to determine whether to ask a sample unit about the particular expense.

Expenditure	Asked	Mean	Variance	Std Err	Sim CV	Rel Bias	Rel Bias SE	RMSE	Bias LB	Bias UB	deff
TX2	1,216.87	636.66	6,657.66	81.59	0.13	1.34	12.99	82.03	-24.11	26.80	5.34
TX3F	198.61	214.12	3,422.57	58.50	0.27	-0.87	27.08	58.53	-53.96	52.21	0.39
TX3H	138.29	103.75	1,790.19	42.31	0.41	-0.37	40.63	42.31	-80.00	79.27	2.19
TX4A_1	633.98	349.55	305.56	17.48	0.05	-0.15	4.99	17.49	-9.93	9.64	3.10
TX4A_2	86.08	27.67	33.26	5.77	0.21	0.45	20.93	5.77	-40.58	41.47	1.18
TX4A_3	81.35	24.10	45.69	6.76	0.28	1.09	28.35	6.76	-54.48	56.67	0.82
TX4A_4	52.73	0.10	0.16	0.40	3.86	-5.59	364.29	0.40	-719.61	708.42	0.90
TX4B	66.68	5.08	7.34	2.71	0.53	2.39	54.62	2.71	-104.67	109.45	0.94
TX4C	318.76	62.15	43.81	6.62	0.11	0.72	10.73	6.63	-20.30	21.74	2.88
TX4D	1,272.27	608.71	740.43	27.21	0.04	0.07	4.47	27.21	-8.69	8.84	5.53
TX5A_1	75.93	6.41	128.95	11.36	1.77	0.93	178.83	11.36	-349.58	351.43	1.28
TX5A_2	56.61	1.52	9.75	3.12	2.05	7.12	219.88	3.12	-423.85	438.09	1.16
TX5B	385.79	363.34	32,506.21	180.29	0.50	1.38	50.31	180.36	-97.22	99.98	4.72
TX6A	129.26	37.46	859.91	29.32	0.78	-2.97	75.95	29.35	-151.83	145.89	2.14
TX6B	338.62	183.54	2,128.59	46.14	0.25	-0.08	25.12	46.14	-49.31	49.16	3.96
TX7A	80.19	16.00	78.40	8.85	0.55	-2.00	54.22	8.86	-108.28	104.27	1.28
TX8	261.52	130.90	2,079.32	45.60	0.35	-0.34	34.72	45.60	-68.39	67.71	3.16
TX9A	465.39	204.72	734.79	27.11	0.13	0.86	13.35	27.16	-25.32	27.03	3.92
TX9B	110.97	36.25	184.03	13.57	0.37	0.26	37.52	13.57	-73.28	73.79	1.41
TX9C	56.78	4.05	11.34	3.37	0.83	3.93	86.36	3.37	-165.33	173.20	1.08
TX9D	59.73	4.10	9.82	3.13	0.76	5.20	80.35	3.14	-152.28	162.67	0.88
TX10C_1	100.16	20.19	26.92	5.19	0.26	0.42	25.80	5.19	-50.15	50.99	0.21
TX10C_23	113.08	50.02	501.57	22.40	0.45	0.21	44.86	22.40	-87.72	88.14	0.22
TX11B	931.94	707.08	13,880.95	117.82	0.17	1.98	16.99	118.62	-31.32	35.29	2.68
TX12A	361.34	165.62	1,582.25	39.78	0.24	-0.13	23.99	39.78	-47.14	46.89	4.86
TX12B	102.62	29.86	92.70	9.63	0.32	0.11	32.28	9.63	-63.15	63.38	1.82
TX12C_1TO9	53.33	24.40	110.38	10.51	0.43	-0.26	42.95	10.51	-84.45	83.93	1.00
TX12C_10	980.14	189.45	101.56	10.08	0.05	0.52	5.35	10.12	-9.96	10.99	3.78
TX13B	721.44	365.92	1,765.46	42.02	0.11	0.21	11.51	42.02	-22.35	22.76	5.80
TX14B	247.59	63.43	249.55	15.80	0.25	2.89	25.62	15.90	-47.34	53.11	1.74
TX15A	449.96	249.42	1,759.34	41.94	0.17	1.60	17.09	42.13	-31.89	35.09	3.61
TX16A	321.57	228.10	9,465.45	97.29	0.43	-0.78	42.32	97.31	-83.73	82.16	3.38
TX17A	125.19	52.10	265.33	16.29	0.31	0.86	31.53	16.29	-60.94	62.66	1.51
TX18A	70.37	9.70	164.38	12.82	1.32	-1.46	130.30	12.82	-256.84	253.92	1.17
TX19A	475.64	229.48	1,322.21	36.36	0.16	0.58	15.94	36.39	-30.66	31.82	3.35
TX19B	709.10	372.29	2,973.20	54.53	0.15	0.75	14.76	54.60	-28.17	29.67	5.31

Table 4.1: Simulation summary statistics for the PPS method

In Table 4.2, we present the simulation summary statistics for the domain characteristics where the domain is defined by those CUs incurring the expense. We report statistics related to these characteristics primarily because the responsive split questionnaire methods developed in this dissertation aim to identify members of the domain and oversample them during the second interview by asking them about their potential expenses in that category. Regardless of the apparent magnitude of the relative biases, all 95% confidence intervals associated with these calculations include zero, therefore we cannot conclude that any are statistically different from zero. The simulation CVs range from about 0.03 to 1.14. In general, the CVs associated with the domain estimates are lower than the corresponding CV associated with the unconditional mean estimate[2]. This finding is consistent with the full CE analysis file (see Tables 3.6 and 3.7).

Under this method, the average number of times we ask about an expense and the sample unit actually incurred it ranges from 0.11 to about 1,230 with half of the expenditure categories being found in fewer than 100 sample units. Given that only two expenditure categories have $P(Int2)$ values of less than 0.01, TX4A_4 (modem purchases, apps, ringtones) and TX5A_2 (construction materials for general jobs), we interpret this as additional evidence that this method might not be very effective in asking CUs about expenses that they are likely to incur.

[2]Recall that the unconditional mean estimate is computed from every sample unit that gets asked about the expense and not just those who incurred it.

Expenditure	Asked & Have	Mean	Variance	Std Err	Sim CV	Rel Bias	Bias SE	RMSE	Bias LB	Bias UB
TX2	1,137.30	2,252.04	6,826.79	82.62	0.04	0.45	3.69	83.25	-6.77	7.68
TX3F	137.23	4,628.81	507,464.98	712.37	0.15	0.25	15.43	712.46	-29.99	30.49
TX3H	72.97	1,879.32	300,473.45	548.15	0.29	1.29	29.54	548.67	-56.62	59.19
TX4A_1	612.68	377.68	207.61	14.41	0.04	-0.18	3.81	14.43	-7.65	7.28
TX4A_2	37.37	102.63	74.24	8.62	0.08	-0.22	8.38	8.62	-16.64	16.19
TX4A_3	27.84	183.40	464.84	21.56	0.12	0.19	11.78	21.56	-22.90	23.27
TX4A_4	0.11	51.95	2,600.67	51.00	0.98	6.56	104.61	51.10	-198.48	211.59
TX4B	12.57	61.33	532.51	23.08	0.38	2.00	38.38	23.11	-73.22	77.22
TX4C	279.28	95.01	41.94	6.48	0.07	0.33	6.84	6.48	-13.08	13.73
TX4D	1,232.17	661.10	419.59	20.48	0.03	0.06	3.10	20.49	-6.02	6.14
TX5A_1	1.81	369.99	178,251.66	422.20	1.14	2.66	117.14	422.31	-226.95	232.26
TX5A_2	0.74	139.06	21,293.88	145.92	1.05	-8.29	96.24	146.46	-196.92	180.33
TX5B	158.58	1,984.66	764,908.89	874.59	0.44	0.52	44.30	874.65	-86.30	87.34
TX6A	9.94	659.63	179,419.54	423.58	0.64	3.21	66.28	424.08	-126.69	133.12
TX6B	188.67	419.22	8,185.03	90.47	0.22	0.23	21.63	90.48	-42.16	42.63
TX7A	11.73	168.83	4,960.86	70.43	0.42	1.15	42.20	70.46	-81.56	83.86
TX8	133.98	350.17	12,968.22	113.88	0.33	-1.26	32.11	113.97	-64.20	61.68
TX9A	351.68	305.48	1,218.88	34.91	0.11	0.70	11.51	34.98	-21.86	23.25
TX9B	44.25	161.81	2,709.10	52.05	0.32	1.60	32.68	52.11	-62.46	65.67
TX9C	4.59	57.87	1,689.24	41.10	0.71	0.88	71.64	41.10	-139.54	141.29
TX9D	7.35	64.14	1,884.16	43.41	0.68	5.25	71.23	43.52	-134.35	144.86
TX10C_1	49.68	507.33	3,056.37	55.28	0.11	0.49	10.95	55.34	-20.98	21.95
TX10C_23	61.43	2,982.39	281,849.52	530.90	0.18	0.18	17.83	530.92	-34.77	35.13
TX11B	889.04	3,906.28	108,231.23	328.99	0.08	0.07	8.43	329.00	-16.45	16.58
TX12A	221.11	302.71	4,101.28	64.04	0.21	-0.14	21.13	64.04	-41.55	41.26
TX12B	25.95	108.69	698.34	26.43	0.24	1.67	24.72	26.49	-46.78	50.12
TX12C_1TO9	14.56	89.81	1,203.30	34.69	0.39	-0.20	38.55	34.69	-75.76	75.35
TX12C_10	942.22	212.06	62.91	7.93	0.04	0.40	3.76	7.98	-6.96	7.76
TX13B	538.14	561.01	2,517.28	50.17	0.09	-0.24	8.92	50.19	-17.72	17.25
TX14B	181.53	235.76	2,222.30	47.14	0.20	2.54	20.50	47.50	-37.64	42.73
TX15A	365.93	403.53	3,634.81	60.29	0.15	1.23	15.12	60.49	-28.42	30.87
TX16A	188.68	983.83	140,168.54	374.39	0.38	-1.12	37.63	374.56	-74.87	72.64
TX17A	64.94	156.52	1,759.45	41.95	0.27	1.09	27.09	41.98	-52.01	54.19
TX18A	3.47	373.32	148,173.10	384.93	1.03	14.25	117.80	387.74	-216.64	245.13
TX19A	400.01	339.12	2,353.22	48.51	0.14	0.34	14.35	48.52	-27.79	28.48
TX19B	628.62	605.66	5,783.52	76.05	0.13	-0.05	12.55	76.05	-24.65	24.55

Table 4.2: Simulation summary statistics for domains using the PPS method

In Table 4.3, we present the epidemiological criteria to evaluate the method's success in "tailoring" the survey to the individual respondent for the PPS method. We conclude that given a sample unit incurred the expense, this method does a poor job of asking about it. This is inferred from the sensitivity calculations. In particular, this method does worse than the "flipping the coin" method as all the sensitivity calculations are less than 0.5. We posit that this is because the PPS method puts greater emphasis on the amount of the expense and not just whether the expense was incurred. On the other hand, given that a CU did not incur the expense, this method does an excellent job of not asking about it since all the specificity calculations are well above 0.9. This is likely due to the fact that the p_{ik} values are so low. Thus, we are simply not asking about the expenditure very many times.

Another potential bright spot regarding this method is that PPV values are greater than or equal to the second interview prevalence for all but one expenditure category, TX12B (vehicle license fees). This suggests that the PPS method does better than a split questionnaire design in which we essentially "flip a coin" to determine whether to ask the question in detecting the event. In other words, if the PPS method recommends asking about the expenditure, then we will more likely observe an instance of incurring the expense than if we randomly chose to ask about the expense with probability one-half. Furthermore, for 26 expenditure categories, we observe PPV values 20% or greater than the prevalence for that expenditure. This implies that, for these expenditures, we are detecting at least

Expenditure	Sensitivity	Specificity	PPV	NPV	$P(Int2)$
TX2	0.387	0.989	0.935	0.806	0.280
TX3F	0.279	0.994	0.691	0.966	0.047
TX3H	0.124	0.993	0.528	0.950	0.056
TX4A_1	0.063	0.973	0.966	0.077	0.925
TX4A_2	0.013	0.994	0.434	0.734	0.268
TX4A_3	0.020	0.994	0.342	0.871	0.130
TX4A_4	0.005	0.995	0.002	0.998	0.002
TX4B	0.015	0.994	0.189	0.918	0.083
TX4C	0.041	0.989	0.876	0.355	0.652
TX4D	0.128	0.952	0.968	0.086	0.921
TX5A_1	0.010	0.993	0.024	0.982	0.018
TX5A_2	0.008	0.995	0.013	0.991	0.009
TX5B	0.083	0.974	0.411	0.827	0.182
TX6A	0.016	0.988	0.077	0.940	0.060
TX6B	0.041	0.975	0.557	0.565	0.439
TX7A	0.011	0.993	0.146	0.903	0.098
TX8	0.034	0.981	0.512	0.633	0.370
TX9A	0.050	0.967	0.756	0.335	0.669
TX9B	0.019	0.992	0.399	0.775	0.227
TX9C	0.006	0.995	0.081	0.932	0.068
TX9D	0.011	0.995	0.123	0.936	0.064
TX10C_1	0.119	0.995	0.496	0.965	0.040
TX10C_23	0.349	0.995	0.544	0.989	0.017
TX11B	0.477	0.995	0.954	0.898	0.178
TX12A	0.039	0.971	0.612	0.455	0.547
TX12B	0.009	0.990	0.253	0.721	0.279
TX12C_1TO9	0.005	0.995	0.273	0.728	0.272
TX12C_10	0.101	0.966	0.961	0.115	0.892
TX13B	0.079	0.950	0.746	0.358	0.649
TX14B	0.065	0.991	0.733	0.743	0.268
TX15A	0.057	0.979	0.813	0.393	0.616
TX16A	0.078	0.984	0.587	0.780	0.231
TX17A	0.019	0.991	0.519	0.669	0.334
TX18A	0.011	0.993	0.049	0.970	0.030
TX19A	0.056	0.978	0.841	0.333	0.675
TX19B	0.098	0.980	0.887	0.410	0.610

Table 4.3: Epidemiological criteria for the PPS method

20% more instances of purchasing the item than in a completely random design[3]. These 26 expenditures correspond to the shaded cells under the PPV column in Table 4.3. These findings may suggest that $y_{Int1,ik}$ may, in fact, be a good proxy for incurring the expense during the second interview reference period. However, given the sensitivity calculations, the manner in which we used it in a PPS design for a responsive split questionnaire might not be the optimal use of this information.

Finally, in regards to NPV, we observe no NPV values less than one minus the prevalence rate for the expenditure. From the hypothetical cases 1 and 2, presented in Section 2.6.2 (see Tables 2.5 and 2.6), the comparison of a method to a completely random split questionnaire design, with respect to NPV, is one minus the prevalence rate of incurring the expense. Thus, since there are no NPV values less than this quantity, we conclude that we are detecting at least as many instances of not incurring the expense than we would if we "flipped a coin" to determine whether to ask about the expense. However, we do not observe any expenditures that yield NPV values at least 20% greater than one minus the prevalence.

4.3 Logistic regression methods

The next two methods for developing decision rules for a responsive split questionnaire design were directly obtained via a series of logistic regression models. The

[3]For this method as well as the other methods explored in this chapter, we used 20% above the prevalence (one minus the prevalence) as an arbitrary cut-off to discuss how well the procedure is performing relative to a split questionnaire design in which the assignment of questions to respondents is completely random. We arbitrarily chose 20% because we had no substantive guidance on how much higher would be viewed as a significant improvement on the standard design.

first of these methods is akin to the situation in which we only have information from the first phase of the responsive split questionnaire. The second is similar to the situation in which we have auxiliary data that we can use in conjunction with our first phase collected data. For these methods, we consider similar logistic regression models to the ones presented in Section 3.3.1.3. In that section, we demonstrated that those models have sound predictive capabilities of the likelihood that the sample unit will incur the expense during the specified reference period.

4.3.1 Statement of the problem

To implement a responsive split questionnaire using logistic regression methods, we must first estimate the propensity that a unit will incur expense k during the reference period asked about in the second interview. If we only have information from the first phase of data collection available, then we can only estimate the propensity that a unit will incur the expense during the reference period inquired about in the first interview. We can use this estimate as a proxy for the propensity that the unit will incur the expense inquired about during the second interview. If we have an auxiliary data source for which we can model the relationship of incurring the expense across the two successive reference periods and provided that the set of explanatory variables used for modeling that relationship are all collected in the first phase of our responsive split questionnaire, we can obtain a firsthand (i.e., non-proxy) estimate for the propensity that the unit will incur the expense in the second interview reference period.

Regardless of which set of information we have available, to explicitly formulate the general problem, we define the following notation.

- $\text{logit}(p_{ik}) = \mathbf{x}_i^T\beta$ where p_{ik} is the i^{th} sample unit's probability of incurring expense k during a reference period, $\mathbf{x}_i$ is a vector of covariates for the i^{th} sample unit, and β is the set of model parameters;

- $\hat{p}_{ik} = (1 + \exp(\mathbf{x}^T\hat{\beta}))^{-1}(\exp(\mathbf{x}^T\hat{\beta}))$ where $\hat{\beta}$ is the set of estimated regression coefficients and $\hat{p}_{ik}$ is the estimated probability that sample unit i incurs expense k and $\hat{p}_{ik}$ becomes the decision rule with which we ask about expenditure k to the i^{th} sample unit in the second interview; and,

- $\alpha_{ik} = 1$ if $i \in S$ receives k and 0 otherwise where S is the set of sample units for which the split questionnaire is being administered.

Using this notation, we can discuss the sampling properties of the logistic regression responsive split questionnaire methods, which are similar in spirit to those under the PPS method. Specifically, if we let $n_{log,k}$ denote the number of sample units receiving expenditure question k under a logistic regression split questionnaire method, then the expectation and variance of $n_{log,k}$ are given in given equations (4.5) and (4.6), respectively, assuming Poisson sampling (according to $\hat{p}_{ik}$) is used to select items for the second interview.

$$E_{SQ}(n_{log,k}|S) = E_{SQ}(\sum_{i\in S}\alpha_{ik}|S) = \sum_{i\in S}E_{SQ}(\alpha_{ik}|S) = \sum_{i\in S}\hat{p}_{ik} \tag{4.5}$$

$$V_{SQ}(n_{log,k}|S) = V_{SQ}(\sum_{i\in S}\alpha_{ik}|S) = \sum_{i\in S}V_{SQ}(\alpha_{ik}|S) = \sum_{i\in S}\hat{p}_{ik}(1-\hat{p}_{ik}) \tag{4.6}$$

Furthermore, from Section 2.4.2.4, we see that the responsive split questionnaire estimator of the full sample mean is approximately design-unbiased. In particular, the expectation and variance of $\hat{\bar{y}}_{SQ,k}$ with respect to a logistic regression responsive split questionnaire is given in (4.7) and (4.8), respectively.

$$\begin{aligned} E_{SQ}(\hat{\bar{y}}_{SQ,k}|S) &\approx E_{SQ}\left[\hat{\bar{y}}_k + (\sum_{i\in S} w_i)^{-1}\left\{\sum_{i\in S}\alpha_{ik}w_i y_{ik}/\hat{p}_{ik} - \hat{\bar{y}}_k\sum_{i\in S}\alpha_{ik}w_i/\hat{p}_{ik}\right\}|S\right] \\ &= \hat{\bar{y}}_k \end{aligned} \tag{4.7}$$

$$\begin{aligned} V_{SQ}(\hat{\bar{y}}_{SQ,k}|S) &\approx (\sum_{i\in S} w_i)^{-2}\left[\sum_{i\in S}\left(\frac{1-\hat{p}_{ik}}{\hat{p}_{ik}}\right)w_i^2(y_i^2 + \hat{\bar{y}}_k^2) - 2\hat{\bar{y}}_k\sum_i\sum_{j>i} y_{ik}w_i w_j\left(\frac{p_{ijk}-\hat{p}_{ik}\hat{p}_{jk}}{\hat{p}_{ik}\hat{p}_{jk}}\right)\right] \\ &= (\sum_{i\in S} w_i)^{-2}\left[\sum_{i\in S}\left(\frac{1-\hat{p}_{ik}}{\hat{p}_{ik}}\right)w_i^2(y_{ik} - \hat{\bar{y}}_k)^2\right] \end{aligned} \tag{4.8}$$

Finally, the expectation of (4.8) with respect to the original sample selection reflects the added contribution of a logistic regression responsive split questionnaire design to the overall sampling variance.

4.3.2 Logistic regression using first interview information

The first logistic regression method we consider is based on the situation in which we only have information from the first phase (e.g., the first interview) to develop the decision rule. We refer to this as the **Log1** method.

4.3.2.1 Simulation setup

To evaluate the performance of the Log1 method, we conducted a statistical simulation. We first estimated the propensity that the sample unit incurred expense k during the first interview reference period. So, for every sample unit i and expenditure k, using only first interview expenditure information, we fit the following model

$$\begin{aligned}\text{logit}(p_{Int1,ik}) = \beta_{0k} + \sum_j \beta_{1jk} \times \text{REGOFF}_{ij} + \beta_{2k} \times \text{SIZE}_i + \beta_{3k} \times \text{POVERTY}_i \\ + \beta_{4k} \times \text{URBAN}_i + \beta_{5k} \times \text{TENURE}_i + \beta_{6k} \times \text{RACE}_i \qquad (4.9)\end{aligned}$$

where $p_{Int1,ik}$ is the probability that sample unit i incurred expense k during the reference period inquired about in the first interview. This is identical to equation (3.4) in Section 3.3.1.3. Since we fit this model to each expenditure individually, we will have a distinct set of regression parameters for each expenditure. We distinguish each distinct set by the subscript k on the β parameters. From this model, we obtain our estimated value of $p_{Int1,ik}$, denoted as $\hat{p}_{Int1,ik}$, which becomes the proxy for the propensity that the sample unit will incur the expense during the second interview reference period. We then use $\hat{p}_{Int1,ik}$ as the decision rule for asking sample unit i about expenditure k during the second interview[4]. Similar to the PPS method, we made a slight modification to the decision rules in order to keep the design measurable and maintain the integrity of the responsive design. Therefore, for any $\hat{p}_{Int1,ik}$

[4]Following the notation provided in Section 4.3.1, we $\hat{p}_{ik} = \hat{p}_{Int1,ik}$ for the Log1 method.

that was less than or equal to 0.005, we set it equal to 0.005[5].

We then carried out the simulation ($M =$)1,000 times and for each iteration, we randomly "asked" sample units questions based on their respective $\hat{p}_{Int1,ik}$ values. We computed the quantities given in equations (3.7) – (3.14) to summarize the simulation and computed similar summary statistics for the domains defined by the CUs that incurred the expense. We also computed the average number of times the question was administered in the second interview, the average number of questions asked, the average interview length, the average time spent answering a question, percent reduction in interview length, and the average and median design effects. Finally, we computed the average sensitivity, specificity, PPV, and NPV for this method.

4.3.2.2 Results

In Table 4.4, we display the summary statistics for the Log1 responsive split questionnaire method. We observe the following trends from this method. The average number of times an expenditure question is asked tracks the first interview prevalence rate for that expenditure. This is reasonable given that this method assigns questions based on the estimated prevalence of incurring the expense in the first interview reference period. For only one expenditure, TX4A_4 (modem purchase, apps, ringtones), we ask fewer than 100 sample units, on average, about it. This is due to its low first interview prevalence of 0.004.

[5]In Table D.2 of Appendix D.2, we present descriptive statistics for the decision rules under the Log1 method before and after the modification was implemented.

Although the magnitude of some of the relative bias calculations may warrant concern, all of their associated 95% confidence intervals include zero. Thus, we conclude that we are still able to obtain design-unbiased estimates of mean quarterly expenditures under the Log1 method after using the appropriate weighting adjustment to account for the question asking procedure. The simulation CVs for this method range from less 0.01 to 3.40. Interestingly, 20 mean expenditure estimates have CVs of less than 0.1, indicating that we may be able to obtain reasonably precise estimates of mean quarterly expenditures for a majority of our expenditure categories. However, TX4A_4 (modem purchase, apps, ringtones) still poses some concerns in terms of our ability to get a precise estimate of the mean quarterly expense as its simulation CV is 3.40. Again, this is likely a consequence of the low prevalence of this expenditure.

Finally, the design effects under this method range from 0.84 to 10.35. Half of the mean expenditure estimates exhibit design effects of less than 1. This suggests that for these expenses we achieved a gain in precision relative to a split questionnaire design in which we essentially "flip a coin" to determine whether to ask a sample unit about a particular expense. Furthermore, the lowest design effect is associated with TX4A_4 (modem purchase, apps, ringtones). We take this as strong evidence that this type of method may actually improve our ability to obtain precise estimates of mean quarterly expenditures for rare expenses *relative to* completely random split questionnaire design.

The largest design effect is associated with TX2 (rental payment) which may seem counter-intuitive. One explanation for this may be due to the fact that over

90% of the sample units we ask about this expense actually incurred it. We showed in Tables 3.6 and 3.7 that, for this expenditure, the element variance when non-reports are excluded from the calculation is higher than the element variance when non-reports are included in the calculation. Since the denominator of the design effect essentially includes non-reports, we might expect the denominator to be lower than anticipated relative to the numerator.

Expenditure	Asked	Mean	Variance	Std Err	Sim CV	Rel Bias	Rel Bias SE	RMSE	Bias LB	Bias UB	deff
TX2	2,931.97	636.95	5,323.71	72.96	0.11	1.39	11.61	73.48	-21.38	24.15	10.35
TX3F	479.50	218.55	3,735.22	61.12	0.28	1.18	28.29	61.17	-54.28	56.63	1.01
TX3H	615.01	104.16	553.63	23.53	0.23	0.03	22.60	23.53	-44.26	44.32	3.06
TX4A_1	9,606.52	350.09	0.57	0.75	0.00	0.01	0.22	0.75	-0.41	0.43	1.00
TX4A_2	3,022.75	27.54	0.56	0.75	0.03	-0.05	2.73	0.75	-5.39	5.30	0.98
TX4A_3	1,363.41	23.78	2.84	1.68	0.07	-0.25	7.07	1.69	-14.10	13.60	0.97
TX4A_4	63.35	0.10	0.12	0.35	3.40	-5.80	320.16	0.35	-633.31	621.70	0.84
TX4B	727.72	4.96	0.68	0.82	0.17	-0.06	16.57	0.82	-32.53	32.42	1.01
TX4C	6,631.61	61.70	0.29	0.54	0.01	-0.01	0.88	0.54	-1.74	1.71	1.03
TX4D	9,618.20	608.25	1.92	1.39	0.00	0.00	0.23	1.39	-0.45	0.44	1.02
TX5A_1	254.88	6.60	31.01	5.57	0.84	3.90	87.70	5.57	-167.98	175.79	1.05
TX5A_2	254.57	1.42	1.62	1.27	0.90	-0.15	89.61	1.27	-175.79	175.49	0.89
TX5B	1,374.13	358.09	1,449.80	38.08	0.11	-0.08	10.62	38.08	-20.91	20.74	0.87
TX6A	444.53	39.05	118.12	10.87	0.28	1.14	28.15	10.88	-54.03	56.31	1.05
TX6B	3,466.38	183.33	32.26	5.68	0.03	-0.19	3.09	5.69	-6.25	5.87	0.94
TX7A	686.01	16.36	6.16	2.48	0.15	0.20	15.20	2.48	-29.59	29.99	0.92
TX8	2,945.90	131.41	41.35	6.43	0.05	0.05	4.90	6.43	-9.54	9.65	0.98
TX9A	5,722.36	202.90	6.15	2.48	0.01	-0.04	1.22	2.48	-2.44	2.35	0.94
TX9B	1,764.53	36.24	7.42	2.72	0.08	0.22	7.53	2.73	-14.54	14.99	1.06
TX9C	443.02	3.89	1.25	1.12	0.29	-0.23	28.64	1.12	-56.36	55.90	0.97
TX9D	473.52	3.90	1.24	1.11	0.29	-0.10	28.54	1.11	-56.04	55.84	0.92
TX10C_1	390.47	19.81	31.67	5.63	0.28	-1.50	27.98	5.64	-56.34	53.35	0.95
TX10C_23	158.82	52.10	2,143.79	46.30	0.89	4.37	92.75	46.35	-177.42	186.16	1.25
TX11B	1,780.65	688.79	2,384.34	48.83	0.07	-0.65	7.04	49.04	-14.46	13.15	1.02
TX12A	3,701.16	166.17	18.85	4.34	0.03	0.20	2.62	4.36	-4.93	5.34	0.94
TX12B	1,590.91	29.82	2.90	1.70	0.06	-0.05	5.71	1.70	-11.24	11.14	1.03
TX12C_1TO9	131.84	24.59	63.02	7.94	0.32	0.55	32.45	7.94	-63.06	64.16	1.42
TX12C_10	9,358.13	188.47	0.38	0.61	0.00	-0.01	0.33	0.61	-0.64	0.63	1.02
TX13B	5,534.98	364.89	19.18	4.38	0.01	-0.07	1.20	4.39	-2.42	2.28	1.01
TX14B	2,080.72	61.55	13.37	3.66	0.06	-0.17	5.93	3.66	-11.79	11.46	0.97
TX15A	5,052.07	245.47	21.25	4.61	0.02	-0.01	1.88	4.61	-3.69	3.67	0.97
TX16A	1,897.94	230.27	405.78	20.14	0.09	0.16	8.76	20.15	-17.01	17.34	1.03
TX17A	2,673.84	51.68	5.47	2.34	0.05	0.03	4.53	2.34	-8.84	8.90	0.91
TX18A	202.40	9.54	50.54	7.11	0.75	-3.09	72.25	7.12	-144.69	138.51	1.05
TX19A	6,274.41	227.99	11.11	3.33	0.01	-0.07	1.46	3.34	-2.93	2.79	0.96
TX19B	5,252.17	369.37	35.04	5.92	0.02	-0.04	1.60	5.92	-3.18	3.10	0.96

Table 4.4: Simulation summary statistics for the Log1 method

In Table 4.5, we present the simulation summary statistics for the domain characteristics where the domain is defined by those CUs incurring the expense. All of the 95% confidence intervals associated with the relative bias calculations include zero, therefore we cannot conclude that any are statistically different from zero. The simulation CVs range from less than 0.01 to 0.97 with 22 CVs being less than 0.1. This may be an indication that under the Log1 method, we may be able to obtain relatively precise estimates of the domain means for a majority of the expenditure categories.

Furthermore, the average number of times we ask about an expense and the sample unit actually incurred it ranges from 0.19 to about 8,936 with 14 of the expenditure categories being observed in fewer than 100 sample units. This is a slight improvement over the PPS method, but given that only two expenditure categories have second interview prevalence rates of less than 0.01, this indicates that the Log1 method might not be effective in asking about expenses that the sample unit is likely to incur. We explore whether this is the case when we evaluate this method on the basis of the epidemiological criteria.

Expenditure	Asked & Have	Mean	Variance	Std Err	Sim CV	Rel Bias	Bias SE	RMSE	Bias LB	Bias UB
TX2	2,620.35	2,241.66	849.36	29.14	0.01	-0.01	1.30	29.14	-2.56	2.54
TX3F	36.12	4,663.06	707,689.95	841.24	0.18	0.99	18.22	842.49	-34.72	36.71
TX3H	58.05	1,865.21	94,208.74	306.93	0.16	0.53	16.54	307.09	-31.90	32.95
TX4A_1	8,934.88	378.40	0.46	0.68	0.00	0.01	0.18	0.68	-0.35	0.36
TX4A_2	923.86	102.88	1.21	1.10	0.01	0.02	1.07	1.10	-2.08	2.11
TX4A_3	228.39	182.72	28.81	5.37	0.03	-0.18	2.93	5.38	-5.93	5.56
TX4A_4	0.19	51.02	2,433.18	49.33	0.97	4.66	101.18	49.38	-193.66	202.98
TX4B	75.66	60.18	45.78	6.77	0.11	0.08	11.25	6.77	-21.98	22.13
TX4C	4,375.66	94.70	0.39	0.63	0.01	0.00	0.66	0.63	-1.30	1.29
TX4D	8,936.60	660.72	1.62	1.27	0.00	0.00	0.19	1.27	-0.37	0.38
TX5A_1	6.09	364.69	70,050.60	264.67	0.73	1.19	73.44	264.71	-142.75	145.12
TX5A_2	3.02	156.14	14,452.84	120.22	0.77	2.97	79.29	120.30	-152.42	158.37
TX5B	322.18	1,977.26	33,939.40	184.23	0.09	0.14	9.33	184.25	-18.14	18.43
TX6A	29.04	643.03	16,740.10	129.38	0.20	0.61	20.24	129.44	-39.06	40.29
TX6B	1,634.97	417.83	125.24	11.19	0.03	-0.10	2.68	11.20	-5.34	5.14
TX7A	85.77	167.74	315.00	17.75	0.11	0.50	10.63	17.77	-20.34	21.34
TX8	1,138.87	354.65	254.47	15.95	0.04	0.00	4.50	15.95	-8.81	8.82
TX9A	3,929.61	303.35	10.49	3.24	0.01	-0.01	1.07	3.24	-2.10	2.08
TX9B	466.01	159.66	98.79	9.94	0.06	0.26	6.24	9.95	-11.97	12.49
TX9C	33.53	57.69	183.94	13.56	0.24	0.56	23.64	13.57	-45.77	46.90
TX9D	39.77	60.85	213.70	14.62	0.24	-0.15	23.99	14.62	-47.17	46.86
TX10C_1	27.49	502.22	5,421.92	73.63	0.15	-0.53	14.58	73.68	-29.11	28.06
TX10C_23	3.87	3,050.25	3,917,137.46	1,979.18	0.65	2.46	66.48	1,980.53	-127.84	132.76
TX11B	366.45	3,884.93	42,718.81	206.69	0.05	-0.48	5.29	207.54	-10.86	9.90
TX12A	2,112.64	303.66	50.19	7.08	0.02	0.17	2.34	7.10	-4.41	4.75
TX12B	482.28	106.80	19.65	4.43	0.04	-0.11	4.15	4.43	-8.23	8.02
TX12C_1TO9	36.58	90.12	598.49	24.46	0.27	0.14	27.19	24.46	-53.14	53.43
TX12C_10	8,508.91	211.21	0.26	0.51	0.00	0.00	0.24	0.51	-0.48	0.47
TX13B	3,727.21	561.94	29.04	5.39	0.01	-0.07	0.96	5.40	-1.95	1.81
TX14B	618.64	229.89	136.82	11.70	0.05	-0.01	5.09	11.70	-9.98	9.96
TX15A	3,287.24	398.63	44.44	6.67	0.02	0.00	1.67	6.67	-3.28	3.28
TX16A	616.55	996.05	5,863.95	76.58	0.08	0.11	7.70	76.58	-14.98	15.20
TX17A	1,029.66	154.84	34.18	5.85	0.04	0.00	3.78	5.85	-7.40	7.41
TX18A	7.33	316.26	39,661.11	199.15	0.63	-3.21	60.95	199.43	-122.67	116.24
TX19A	4,421.80	337.72	20.11	4.48	0.01	-0.07	1.33	4.49	-2.67	2.53
TX19B	3,296.71	605.73	72.53	8.52	0.01	-0.04	1.41	8.52	-2.79	2.72

Table 4.5: Simulation summary statistics for domains using the Log1 method

In Table 4.6, we present the calculations of the four epidemiological criteria for the Log1 method. We conclude that given a sample unit incurred the expense, this method does a fair job of asking CUs about expenses that they are likely to have incurred. Specifically, about one-third, or 10 out of 36 expenditure categories had sensitivity values higher than 0.5. The Log1 method is fairly effective for TX4A_1 (telephone services), TX4D (utilities), and TX12C_10 (average monthly gas expenditures) as each of these expenditure categories exhibit sensitivity values above 0.9.

In terms of specificity, this method does a good job of correctly not asking the sample units about expenditures that they are not likely to have incurred. Of the 36 expenditure categories that we investigated, 30 had specificity values higher than 0.5. TX4A_1 (telephone services), TX4D (utilities), and TX12C_10 (average monthly gas expenditures) have the lowest specificity values. The specificity values associated with these expenditures were all less than 0.25. Interestingly enough, these expenditures were the same set that we concluded the method was performing well when we investigated the sensitivity values. This seemingly contradictory information may suggest that for these types of expenses, the model used in the Log1 method may be better at detecting incurring the expense rather than not incurring the expense.

The third evaluation criterion we report on is PPV. Overall, we see that the method is effective in predicting which CUs actually incur a given expenditure. In fact, all PPV values were greater than their associated second interview prevalence rates and for 15 out of the 36 expenditure categories, the PPV value was at least 20% higher than the prevalence. These correspond to the shaded cells under the

Expenditure	Sensitivity	Specificity	PPV	NPV	$P(Int2)$
TX2	0.891	0.959	0.894	0.958	0.280
TX3F	0.074	0.956	0.075	0.955	0.047
TX3H	0.099	0.944	0.094	0.946	0.056
TX4A_1	0.920	0.144	0.930	0.128	0.925
TX4A_2	0.329	0.727	0.306	0.747	0.268
TX4A_3	0.167	0.876	0.168	0.875	0.130
TX4A_4	0.008	0.994	0.003	0.998	0.002
TX4B	0.087	0.932	0.104	0.919	0.083
TX4C	0.640	0.383	0.660	0.362	0.652
TX4D	0.925	0.182	0.929	0.173	0.921
TX5A_1	0.033	0.976	0.024	0.983	0.018
TX5A_2	0.031	0.976	0.012	0.991	0.009
TX5B	0.169	0.878	0.234	0.826	0.182
TX6A	0.046	0.958	0.065	0.940	0.060
TX6B	0.355	0.689	0.472	0.577	0.439
TX7A	0.084	0.937	0.125	0.904	0.098
TX8	0.293	0.727	0.387	0.636	0.370
TX9A	0.560	0.484	0.687	0.352	0.669
TX9B	0.196	0.840	0.264	0.780	0.227
TX9C	0.047	0.958	0.076	0.932	0.068
TX9D	0.059	0.956	0.084	0.937	0.064
TX10C_1	0.066	0.964	0.070	0.961	0.040
TX10C_23	0.022	0.985	0.024	0.983	0.017
TX11B	0.197	0.836	0.206	0.828	0.178
TX12A	0.368	0.666	0.571	0.466	0.547
TX12B	0.165	0.853	0.303	0.725	0.279
TX12C_1TO9	0.013	0.988	0.277	0.728	0.272
TX12C_10	0.909	0.248	0.909	0.247	0.892
TX13B	0.547	0.509	0.673	0.377	0.649
TX14B	0.220	0.810	0.297	0.739	0.268
TX15A	0.509	0.562	0.651	0.417	0.616
TX16A	0.254	0.841	0.325	0.790	0.231
TX17A	0.294	0.765	0.385	0.684	0.334
TX18A	0.023	0.981	0.036	0.970	0.030
TX19A	0.624	0.457	0.705	0.369	0.675
TX19B	0.515	0.522	0.628	0.408	0.610

Table 4.6: Epidemiological criteria for Log1 method

PPV column in Table 4.6. This suggests that for these 15 expenditure categroies, we are detecting 20% or more of the instances of incurring the expense than we would by "flipping a coin." An interesting observation is that a majority of these 15 expenditure categories have quite low prevalence rates and may be regarded as rare events. In fact, all but three of the 15 have prevalence rates of less than 0.1. This may indicate that we are improving our ability to detect the rare event relative to just "flipping the coin." So, if we are concerned about missing the event (e.g., the purchase of the item) by not asking questions pertaining to the expense, then this finding provides evidence that under a tailored method of asking questions, we can alleviate that concern to some degree.

The final evaluation criterion we report is NPV. Overall, the Log1 method does a fair job of distinguishing which sample units did not incur the expense. For all expenditures, we did no worse than the completely random split questionnaire design and for four of the 36 expenditure categories the NPV was at least 20% higher than one minus the prevalence. These four correspond to the shaded cells under the NPV column of Table 4.6.

4.3.3 Logistic regression using first interview information in conjunction with auxiliary data

The second logistic regression method arises from the situation when we not only have information collected in the first phase of a responsive split questionnaire but we also have auxiliary information on the relationship between incurring the

expense in two successive reference periods. We refer to this as the **Log2** method.

4.3.3.1 Simulation setup

Before we conducted a simulation to evaluate the performance of the Log2 method, we modeled the relationship of incurring each expense in two successive reference periods from our "auxiliary data" source. Specifically, using the analysis file, we drew $1,000$ simple random samples without replacement of size $1,000$. Using data from each sample, we fit the following model[6] for each expenditure k

$$\begin{aligned}\text{logit}(p_{Int2,ik}) = \beta_{0k} + \beta_{1k} \times \text{REGOFF}_i + \beta_{2k} \times \text{SIZE}_i + \beta_{3k} \times \text{POVERTY}_i \\ + \beta_{4k} \times \text{URBAN}_i + \beta_{5k} \times \text{TENURE}_i + \beta_{6k} \times \text{RACE}_i + \beta_{7k} \times I_{ik}\end{aligned} \tag{4.10}$$

where $p_{Int2,ik}$ denotes the probability that the i^{th} sample unit incurred expense k during the reference period inquired about in the second interview and $I_{ik} = 1$ if sample unit i incurred expense k during the first interview reference period and 0 otherwise. As with the Log1 method, we will have a different set of parameter estimates for each expenditure. It is important to acknowledge that we were constrained by access to additional data so we drew bootstrap samples from our analysis file to estimate the model parameters. We could not use only the cases in the analysis file to model this relationship because it requires information that we are collecting

[6]This model deviates slightly from equation (3.6) presented in Section 3.3.1.3. The difference in this model is that we incorporated REGOFF as a continuous covariate rather than a categorical covariate because were concerned about the small cell sizes resulting from the combinations of explanatory variables in this model.

in the second interview. If we used data directly from the analysis file to model this relationship, then we would likely overstate the (successful) performance of this method.

From each bootstrap sample, we obtain estimates of the eight coefficients. The distinct coefficients are then averaged across the $1,000$ bootstrap samples and these averages became the parameters we used to obtain our decision rules under the Log2 method. In particular, if we let $\hat{\beta}_{jkm}$ be the estimate of the j^{th} parameter of model (4.10) from the m^{th} bootstrap sample for expenditure k and define $\bar{\hat{\beta}}_{jk} = M^{-1}\sum_m \hat{\beta}_{jkm}$ for $j = 0, 1, \ldots, 7$, $k = 1, 2, \ldots, K$, and $m = 1, 2, \ldots, 1,000$, then the model we use to obtain the decision rules for the responsive split questionnaire under the Log2 method becomes the following.

$$\begin{aligned}\text{logit}(\hat{p}_{Int2,ik}) = \bar{\hat{\beta}}_{0k} + \bar{\hat{\beta}}_{1k} \times \text{REGOFF}_i + \bar{\hat{\beta}}_{2k} \times \text{SIZE}_i + \bar{\hat{\beta}}_{3k} \times \text{POVERTY}_i \\ + \bar{\hat{\beta}}_{4k} \times \text{URBAN}_i + \bar{\hat{\beta}}_{5k} \times \text{TENURE}_i + \bar{\hat{\beta}}_{6k} \times \text{RACE}_i + \bar{\hat{\beta}}_{7k} \times I_{ik}\end{aligned} \tag{4.11}$$

The full set of averaged parameter estimates for this model for the 36 expenditure categories is provided in Appendix D.2 (see Table D.3).

Using the information collected in the first phase of the responsive split questionnaire (and contained in our analysis file) and the following relationship

$$\hat{p}_{Int2,ik} = (1 + \exp(\mathbf{x}^T \bar{\hat{\beta}}))^{-1}(\exp(\mathbf{x}^T \bar{\hat{\beta}})) \tag{4.12}$$

we obtain our estimate of $p_{Int2,ik}$, denoted as $\hat{p}_{Int2,ik}$, for each sample unit i and expenditure k. Thus, $\hat{p}_{Int2,ik}$ is the decision rule we use to ask the i^{th} sample unit about expenditure k[7]. As with the other methods, in order to keep the responsive split questionnaire design measurable and maintain the integrity of the responsive design, we set $\hat{p}_{Int2,ik} = 0.005$ if the original estimate was less than this value[8].

After we obtained the $\hat{p}_{Int2,ik}$ values, we carried out a simulation $(M =)1,000$ times and for each iteration, we randomly "asked" sample units questions based on their respective $\hat{p}_{Int2,ik}$ values. We then computed the quantities given in equations (3.7) – (3.14) to summarize the simulation and computed similar summary statistics for the domains defined by the CUs that incurred the expense. We also computed the average number of times the question was administered in the second interview, the average number of questions asked, the average time spent answering a question, the average interview length, percent reduction in interview length, and the average and median design effects. Finally, we computed the average sensitivity, specificity, PPV, and NPV for this method.

There are two final points worth making about the Log2 method before we present the results of the simulation. First, the estimated probability, $\hat{p}_{Int2,ik}$, should be a fairly close approximation to the probability that the sample unit will actually incur the expense in the second interview reference period. This is because we are explicitly modeling the occurrence of this event, as opposed to using a proxy for the event (e.g., incurring the expense in the first interview reference period). Thus,

[7]Under the Log2 method $\hat{p}_{ik} = \hat{p}_{Int2,ik}$.

[8]In Table D.4 of Appendix D.2, we present descriptive statistics for the decision rules under the Log2 method before and after the modification was implemented.

this method should outperform the previous two methods in terms of correctly distinguishing between the units that do and do not incur the expense because no proxy for the event is no used.

The second point is that there are other ways to obtain the estimates of the model parameters for (4.10) when auxiliary data are available. In other words, it is not necessary to draw bootstrap samples from a data source and combine them in some way to estimate the required regression coefficients. An alternative approach would be to fit the model to the auxiliary data source (assuming that the covariates one uses to fit the model are also collected in the initial phase of data collection of the responsive split questionnaire) and obtain the parameter estimates via that data. However, since we did not have access to additional data during the completion of this research, we feel that drawing bootstrap samples was a reasonable approach within the constraints of our research.

4.3.3.2 Results

In Table 4.7, we display the simulation summary statistics for the Log2 method. We observe the following trends for this method. First, the average number of times an expenditure question is asked tracks the second interview prevalence for that expenditure. This is because this method assigns questions based on the estimated prevalence of incurring the expense during the second interview reference period. We ask about two expenditure categories, TX4A_4 (modem purchases, apps, ringtones) and TX5A_2 (construction materials for general jobs), to fewer than 100 sample

units, on average. This is due to the low prevalence rates of these expenses.

Although the magnitudes of some of the relative biases associated with mean quarterly expenditures appear high, all of their 95% confidence intervals include zero. Therefore, we cannot conclude that these biases are statistically different from zero. In addition, some of the 95% confidence intervals for the relative bias calculations are quite wide. For example, TX4A_4 (modem purchases, apps, ringtones), with a relative bias estimate of -10.38%, has a 95% confidence interval of $(-693.07, 672.31)$. As this trend is not unique to this method, we posit that a likely reason for this may be a function of the prevalence of the expenditure and the number of sample units getting asked about the expense. However, despite the width of some of the 95% confidence intervals associated with the relative bias calculations, we conclude that we are still able to obtain design-unbiased estimates of mean quarterly expenditures under the Log2 method after using the appropriate weighting adjustment to account for the responsive split questionnaire design.

The simulation CVs for this method range from less than 0.01 to 3.89, with 21 mean quarterly expenditure estimates having CVs of less than 0.1. This indicates that we can obtain fairly precise estimates of desired quantities under this method. As with the other methods, the least prevalent expenses, TX4A_4 (modem purchases, apps, ringtones), TX5A_1 (construction materials for specific jobs), and TX5A_2 (construction materials for general jobs), exhibit the highest CVs. Their values are 3.89, 1.11, and 1.78, respectively. This may be an indication that we might need to improve the responsive split questionnaire design with the goal of obtaining more precise estimates for the rarest expenditure categories.

Finally, the design effects for this method range from 0.34 to 10.44. In fact, 21 of the expenditure estimates have design effects of less than 1. This indicates substantial gains in precision relative to a split questionnaire design in which we essentially "flip a coin" to determine whether to ask a sample unit about a particular expenditure.

Expenditure	Asked	Mean	Variance	Std Err	Sim CV	Rel Bias	Rel Bias SE	RMSE	Bias LB	Bias UB	deff
TX2	2,966.58	637.40	5,307.85	72.85	0.11	1.46	11.60	73.43	-21.27	24.19	10.44
TX3F	499.96	224.39	5,720.12	75.63	0.34	3.88	35.01	76.09	-64.75	72.51	1.59
TX3H	580.32	105.50	775.39	27.85	0.26	1.31	26.74	27.88	-51.10	53.72	4.02
TX4A_1	9,724.76	350.03	1.82	1.35	0.00	-0.01	0.39	1.35	-0.77	0.74	1.20
TX4A_2	2,804.77	27.57	1.10	1.05	0.04	0.06	3.81	1.05	-7.41	7.52	1.53
TX4A_3	1,356.41	23.85	4.01	2.00	0.08	0.06	8.40	2.00	-16.40	16.52	1.31
TX4A_4	52.46	0.10	0.15	0.38	3.89	-10.38	348.31	0.38	-693.07	672.31	0.82
TX4B	843.34	4.95	0.49	0.70	0.14	-0.12	14.04	0.70	-27.64	27.41	0.86
TX4C	6,840.83	61.69	0.64	0.80	0.01	-0.03	1.29	0.80	-2.57	2.50	1.54
TX4D	9,677.14	608.27	6.29	2.51	0.00	0.00	0.41	2.51	-0.81	0.81	1.27
TX5A_1	134.78	6.53	52.13	7.22	1.11	2.78	113.70	7.22	-220.08	225.64	0.93
TX5A_2	73.60	1.44	6.61	2.57	1.78	1.62	181.12	2.57	-353.38	356.62	1.02
TX5B	1,881.23	357.90	862.82	29.37	0.08	-0.14	8.20	29.38	-16.20	15.93	0.79
TX6A	594.24	38.54	81.73	9.04	0.23	-0.18	23.41	9.04	-46.07	45.71	0.98
TX6B	4,608.87	183.64	20.96	4.58	0.02	-0.02	2.49	4.58	-4.91	4.86	0.97
TX7A	995.06	16.40	4.05	2.01	0.12	0.43	12.33	2.01	-23.73	24.59	0.91
TX8	3,876.02	131.39	24.40	4.94	0.04	0.04	3.76	4.94	-7.33	7.41	0.91
TX9A	7,023.70	202.89	3.85	1.96	0.01	-0.04	0.97	1.97	-1.94	1.85	0.98
TX9B	2,364.46	36.21	4.93	2.22	0.06	0.13	6.14	2.22	-11.90	12.16	1.02
TX9C	686.00	3.87	0.79	0.89	0.23	-0.81	22.76	0.89	-45.41	43.80	0.97
TX9D	632.87	3.94	0.92	0.96	0.24	0.99	24.60	0.96	-47.23	49.21	0.93
TX10C_1	441.87	20.16	17.71	4.21	0.21	0.27	20.93	4.21	-40.74	41.29	0.62
TX10C_23	209.79	49.88	417.58	20.43	0.41	-0.08	40.94	20.43	-80.31	80.16	0.34
TX11B	1,835.34	702.32	10,565.31	102.79	0.15	1.30	14.83	103.18	-27.76	30.35	4.07
TX12A	5,739.50	166.02	8.19	2.86	0.02	0.11	1.73	2.87	-3.27	3.50	0.94
TX12B	2,902.64	29.82	1.29	1.14	0.04	-0.03	3.81	1.14	-7.49	7.43	0.99
TX12C_1TO9	2,827.51	24.36	1.51	1.23	0.05	-0.40	5.02	1.23	-10.24	9.43	0.99
TX12C_10	9,377.28	188.45	1.35	1.16	0.01	-0.02	0.62	1.16	-1.22	1.19	1.36
TX13B	6,816.55	365.06	13.40	3.66	0.01	-0.03	1.00	3.66	-1.99	1.94	1.06
TX14B	2,800.02	61.68	6.01	2.45	0.04	0.04	3.98	2.45	-7.75	7.83	0.73
TX15A	6,476.32	245.64	11.79	3.43	0.01	0.06	1.40	3.44	-2.68	2.80	0.96
TX16A	2,406.53	229.85	354.35	18.82	0.08	-0.02	8.19	18.82	-16.07	16.03	1.17
TX17A	3,490.64	51.63	3.59	1.89	0.04	-0.06	3.67	1.89	-7.25	7.12	0.90
TX18A	256.22	10.16	41.90	6.47	0.64	3.25	65.78	6.48	-125.69	132.18	1.11
TX19A	7,092.00	228.11	8.18	2.86	0.01	-0.02	1.25	2.86	-2.48	2.44	0.98
TX19B	6,408.67	369.61	22.61	4.75	0.01	0.02	1.29	4.76	-2.50	2.55	0.97

Table 4.7: Simulation summary statistics for the Log2 method

In Table 4.8, we present the simulation summary statistics for the domain characteristics where the domain defined by those CUs incurring the expense. The average number of times we ask about an expense and the CU incurred it ranges from 0.11 to about 9,310. However, we observe seven expenditure categories being observed in fewer than 100 sample units. Even though this is a modest improvement over the PPS and Log1 methods, this may be an indication that the general model might not work well for some expenses since there are only two expenditure categories that have prevalence rates this low.

Under the Log2 method, all 95% confidence intervals associated with the relative bias calculations for the domain means include zero, therefore we cannot conclude that any are statistically different from zero. The simulation CVs range from less than 0.01 to 1.06 with 26 expenditure categories exhibiting CVs less than than or equal to 0.1. This indicates that under the Log2 method we can obtain fairly precise estimates of the domain means for about two-thirds of our expenditure categories.

Expenditure	Asked & Have	Mean	Variance	Std Err	Sim CV	Rel Bias	Bias SE	RMSE	Bias LB	Bias UB
TX2	2,777.16	2,241.53	653.55	25.56	0.01	-0.01	1.14	25.57	-2.25	2.22
TX3F	411.92	4,714.20	643,650.17	802.28	0.17	2.10	17.38	808.13	-31.96	36.16
TX3H	424.72	1,872.08	106,964.30	327.05	0.17	0.90	17.63	327.48	-33.65	35.44
TX4A_1	9,307.59	378.40	0.47	0.68	0.00	0.01	0.18	0.68	-0.35	0.36
TX4A_2	1,659.11	102.80	1.74	1.32	0.01	-0.06	1.28	1.32	-2.57	2.45
TX4A_3	767.41	182.51	30.56	5.53	0.03	-0.30	3.02	5.56	-6.22	5.62
TX4A_4	0.11	47.88	2,222.42	47.14	0.98	-1.78	96.70	47.15	-191.32	187.75
TX4B	224.59	60.27	36.15	6.01	0.10	0.24	10.00	6.01	-19.36	19.84
TX4C	5,779.49	94.69	0.31	0.56	0.01	-0.01	0.59	0.56	-1.16	1.14
TX4D	9,255.09	660.63	1.58	1.26	0.00	-0.01	0.19	1.26	-0.38	0.36
TX5A_1	2.84	381.97	163,701.43	404.60	1.06	5.98	112.26	405.17	-214.05	226.01
TX5A_2	0.75	153.33	22,954.79	151.51	0.99	1.12	99.92	151.52	-194.72	196.96
TX5B	506.20	1,973.69	19,823.42	140.80	0.07	-0.04	7.13	140.80	-14.01	13.94
TX6A	40.25	639.89	11,914.97	109.16	0.17	0.12	17.08	109.16	-33.35	33.60
TX6B	2,262.20	418.34	82.37	9.08	0.02	0.02	2.17	9.08	-4.23	4.27
TX7A	164.31	167.26	202.45	14.23	0.09	0.21	8.52	14.23	-16.50	16.92
TX8	1,627.12	354.62	142.83	11.95	0.03	0.00	3.37	11.95	-6.61	6.60
TX9A	4,914.20	303.33	5.82	2.41	0.01	-0.01	0.80	2.41	-1.57	1.55
TX9B	787.57	159.68	65.35	8.08	0.05	0.27	5.08	8.10	-9.68	10.22
TX9C	65.37	57.42	116.37	10.79	0.19	0.08	18.80	10.79	-36.77	36.94
TX9D	96.83	61.35	147.70	12.15	0.20	0.67	19.94	12.16	-38.41	39.76
TX10C_1	391.13	504.53	955.26	30.91	0.06	-0.07	6.12	30.91	-12.07	11.93
TX10C_23	158.10	2,940.10	168,332.36	410.28	0.14	-1.24	13.78	411.94	-28.25	25.77
TX11B	1,781.64	3,895.18	60,192.08	245.34	0.06	-0.22	6.28	245.49	-12.54	12.10
TX12A	3,341.87	303.33	21.82	4.67	0.02	0.06	1.54	4.67	-2.96	3.08
TX12B	858.18	106.92	8.50	2.92	0.03	0.01	2.73	2.92	-5.34	5.35
TX12C_1TO9	837.93	89.71	14.93	3.86	0.04	-0.31	4.29	3.87	-8.73	8.10
TX12C_10	8,921.19	211.20	0.33	0.57	0.00	-0.01	0.27	0.57	-0.54	0.52
TX13B	4,941.40	562.21	16.55	4.07	0.01	-0.02	0.72	4.07	-1.44	1.40
TX14B	1,468.69	229.96	55.01	7.42	0.03	0.02	3.23	7.42	-6.30	6.34
TX15A	4,530.86	398.97	21.39	4.62	0.01	0.08	1.16	4.64	-2.19	2.36
TX16A	1,117.62	994.70	5,031.84	70.94	0.07	-0.03	7.13	70.94	-14.00	13.95
TX17A	1,668.89	154.74	21.61	4.65	0.03	-0.06	3.00	4.65	-5.94	5.83
TX18A	15.91	339.10	31,306.62	176.94	0.52	3.77	54.15	177.37	-102.36	109.90
TX19A	5,406.27	337.97	12.66	3.56	0.01	0.00	1.05	3.56	-2.06	2.07
TX19B	4,535.80	605.98	39.84	6.31	0.01	0.01	1.04	6.31	-2.04	2.05

Table 4.8: Simulation summary statistics for domains using the Log2 method

In Table 4.9, we present the calculations for the four epidemiological criteria for the Log2 method. We conclude that given a sample unit incurred the expense, this method does a good job of asking CUs about expenses that they are likely to have incurred. In particular, over half of the expenditure categories had sensitivity values greater than 0.5, with six exhibiting sensitivity calculations greater than 0.9. These are TX2 (rental payment), TX4A_1 (telephone services), TX4D (utilities and fuels), TX10C_1 (car monthly payment), TX11B (owned car down payment), and TX12C_10 (average monthly gas expense).

In terms of specificity, the method does a good job of correctly not asking sample units about expenses that they are not likely to have incurred. Of the 36 expenditure categories we investigated, only one exhibited a specificity value substantially less than 0.5. This expenditure category was TX9A (clothing) as the specificity value for this expense was 0.393.

The third evaluation criteria we report on is PPV. Overall, we see that this method is effective in predicting which CUs incur a given expense. All PPV values are greater than the second interview prevalence rates suggesting that we do no worse than the "flipping the coin" method. Furthermore, for 20 expenditures, we achieve PPV values 20% or greater than the associated prevalence rates. These 19 correspond to the shaded cells under the PPV column of Table 4.9. This indicates that for these expenses we are detecting 20% more instances of incurring the expense than we would by a completely random split questionnaire design.

The final criteria we report on is NPV. Overall, the method does a good job at distinguishing which sample units did not incur the expense. For all expenditure

Expenditure	Sensitivity	Specificity	PPV	NPV	$P(Int2)$
TX2	0.944	0.975	0.936	0.978	0.280
TX3F	0.839	0.991	0.824	0.992	0.047
TX3H	0.721	0.984	0.732	0.983	0.056
TX4A_1	0.959	0.469	0.957	0.478	0.925
TX4A_2	0.590	0.851	0.592	0.850	0.268
TX4A_3	0.561	0.935	0.566	0.934	0.130
TX4A_4	0.005	0.995	0.002	0.998	0.002
TX4B	0.259	0.936	0.266	0.934	0.083
TX4C	0.845	0.710	0.845	0.710	0.652
TX4D	0.958	0.493	0.956	0.503	0.921
TX5A_1	0.015	0.987	0.021	0.982	0.018
TX5A_2	0.008	0.993	0.010	0.991	0.009
TX5B	0.266	0.840	0.269	0.838	0.182
TX6A	0.063	0.944	0.068	0.940	0.060
TX6B	0.491	0.601	0.491	0.601	0.439
TX7A	0.160	0.912	0.165	0.909	0.098
TX8	0.419	0.660	0.420	0.659	0.370
TX9A	0.700	0.393	0.700	0.393	0.669
TX9B	0.330	0.806	0.333	0.804	0.227
TX9C	0.092	0.937	0.095	0.934	0.068
TX9D	0.144	0.945	0.153	0.942	0.064
TX10C_1	0.936	0.995	0.885	0.997	0.040
TX10C_23	0.898	0.995	0.754	0.998	0.017
TX11B	0.956	0.994	0.971	0.990	0.178
TX12A	0.582	0.496	0.582	0.496	0.547
TX12B	0.293	0.730	0.296	0.727	0.279
TX12C_1TO9	0.294	0.740	0.296	0.737	0.272
TX12C_10	0.953	0.596	0.951	0.603	0.892
TX13B	0.725	0.490	0.725	0.491	0.649
TX14B	0.522	0.827	0.525	0.825	0.268
TX15A	0.701	0.517	0.700	0.519	0.616
TX16A	0.461	0.840	0.464	0.838	0.231
TX17A	0.477	0.739	0.478	0.738	0.334
TX18A	0.050	0.976	0.062	0.971	0.030
TX19A	0.763	0.506	0.762	0.507	0.675
TX19B	0.709	0.543	0.708	0.544	0.610

Table 4.9: Epidemiological criteria for the Log2 method

categories, we do no worse than the completely random split questionnaire design and for 10 out of the 36 expenditures, we observe NPV values that are 20% or greater than one minus the prevalence for that expenditure. The shaded cells under the NPV column in Table 4.9 correpsond to the 10 expenditure categories with NPV values that are at least 20% greater than one minus the prevalence rate.

4.4 Stratification methods

In Section 2.4.3, we identified two perspectives on developing decision rules for a responsive split questionnaire. These perspectives were termed direct and indirect. The focus of this section is on the indirect perspective. Under this perspective, we use the information collected during the first phase of data collection to stratify the sample based on their anticipated likelihood of incurring expense k. This approach is similar to the "four-fifths" method that we explored in Preliminary Study 2. Under that method, we stratified based on the indicator of incurring the expense. We then arbitrarily chose the probability of 4/5 with which we asked about the expenditure in the second interview. In this setup, we will stratify using what we believe is a better indication of incurring the expense and in a manner that is consistent with the survey methodological literature on forming classes for nonresponse adjustments (Little, 1986; Rosenbaum and Rubin, 1983). Furthermore, for these methods, we will also attempt to determine an optimal subsampling probability, as opposed to an arbitrarily chosen one, using mathematical programming methods.

4.4.1 Statement of the problem

To formulate the problem of designing a responsive split questionnaire using stratification methods, let us first assume that we have a simplified version of the problem in which we only have one expenditure item on our survey. Following the standard two-phase sampling for stratification setup as given by Särndal et al. (1992), suppose that in a first phase of data collection a large sample is drawn according to some probability sample design. For the elements selected in this first phase, information is recorded that will allow stratification of these units. Linking these steps to our research, the first phase of data collection is the first interview and the recording of information is the data collection effort of the first interview.

Once we have the information, we stratify the first phase sample into H strata (for simplicity, assume that we have two strata). For now, with respect to our simplified research problem, our mechanism to stratify the first phase sample classifies the units into strata based on their anticipated likelihood of incurring that expense during the second interview reference period. For instance, we may have a "low" likelihood stratum and a "high" likelihood stratum. If our stratification mechanism was effective, then we would expect the "low" stratum to contain most of the units who do not incur expenses in that category. Conversely, members comprising the "high" stratum would be likely purchasers.

Finally, from each stratum h, a sample is drawn according to some probability sampling design. Sampling is carried out independently in each stratum. It is worth noting that when a unit is sampled from the stratum, it is asked about the

expenditure. If we were to set up a standard stratified allocation problem that is consistent with the goals of the methods proposed in this dissertation (i.e., ask sample units about expenses that they are likely to incur), then we would seek an allocation such that a larger proportion of our sample would come from the "high" stratum. This is because this stratum contains the likely purchasers. We would also want the allocation to apportion a small proportion of our overall second interview sample to come from the "low" stratum because those units are thought to not incur the expense.

Before we set up the stratified sample allocation problem that is consistent with optimal allocation methods of Cochran (1977), we make a few comments about relevant stratum-specific quantities under the setup described in the preceding paragraphs. Assuming that we have a powerful stratification mechanism that correctly classifies units, then since most of the members in the "low" stratum do not have the expenditure (i.e., $y_{ik} = 0$), the mean of y_k, the dollar amount of the expense, for that stratum would be close to zero. Furthermore, since everyone is effectively a non-purchaser, the element variance for this stratum would be quite low. In particular, if everyone in the stratum was correctly classified as a non-purchaser, then the element variance would be identically equal to zero (i.e., $S_h = 0$). In the "high" stratum, we would expect $\bar{y}_k > 0$ simply because the members comprising this stratum have $y_{ik} > 0$. Additionally, we would have $S_h > 0$ in the "high" stratum.

Now, to verify whether the optimal allocation methods of Cochran (1977) are consistent with our research goals, we illustrate our simplified version of the problem with one expenditure, TX5B (contractor labor, materials, and tools). Assume that

we have stratified the units in the analysis file into two strata "low" and "high," based on their likelihood of incurring this expense in the second interview reference period[9], as shown in Table 4.10. After we have stratified the units, we also compute the stratum-specific S_h values[10]. Since we have collected information in the first phase of our responsive design, we can estimate these quantities based on the first interview expenditure information. Assume further that we have a linear "cost" function $C = \sum_h c_h n_h$ with a budget of $5,000$ and arbitrary cost parameters of 9 and 1 for the "low" and "high" strata, respectively[11]. In Section 2.5.2, we showed that the set of stratum sample sizes $\{n_h\}$ that minimizes the sampling variance of the mean expenditure estimate subject to the cost constraint will be given as follows

$$n_h = n\frac{N_h S_h/\sqrt{c_h}}{\sum_h N_h S_h/\sqrt{c_h}} \tag{4.13}$$

with $n = [\sum N_h S_h \sqrt{c_h}]^{-1}(C)[\sum (N_h S_h/\sqrt{c_h})]$ when cost is fixed. We display the stratum sample sizes resulting from this optimization problem in the last column of Table 4.10. In our research, the decision rules for asking the i^{th} sample unit about expenditure k become functions of the set of $\{n_h\}$. Specifically, the decision rules for responsive split questionnaires based on stratification methods are n_h/N_h.

What we observe from the allocation results displayed in Table 4.10, is that a higher proportion of our sample comes from the "high" stratum. We note that the

[9]We will provide specific details on how we obtain this stratification classification in the next section.

[10]The subscript k is omitted here since, at this time, we are only detailing the setup for one expenditure, however, the unit standard deviations will depend on the expenditure category.

[11]We note here that "cost" does not necessarily have to mean dollars with which to collect the data. We discuss why this point is relevant in Section 4.4.2.

	N_h	S_h	c_h	n_h
Low	5,288	259.06	9	132.19
High	5,207	2,527.69	1	3,810.25

Table 4.10: Example of stratified sampling allocation for one expenditure, TX5B (contractor labor, materials, and tools)

"high" stratum is internally more variable than the "low" stratum, as indicated by their respective S_h values. Furthermore, we set up this problem so that the "high" stratum is cheaper than the "low" stratum. Finally, we observe that the strata are of roughly equal size, as indicated by the N_h values.

In Section 2.5.2 we noted that the general rules of this type of optimal allocation (per Cochran, 1977) dictate that a higher proportion of the sample will be allocated to (1) larger strata, (2) internally variable strata and, (3) cheaper strata. So provided that our stratification mechanism yields strata with the characteristics given in (1) – (3), we should be able to use standard optimal stratified sampling techniques to determine the decision rules for a responsive split questionnaire when the sample is stratified into groups based on their likelihood of purchasing particular items. We have demonstrated that this setup works well for one expenditure category. Given this evidence, the goal now becomes to extend the problem to more than one expenditure and perhaps additional constraints (imposing additional constraints may help meet diverse stakeholder needs). To accomplish this, a slightly different mathematical formulation of the problem is required; however, the general spirit of the formulation remains the same. We formulate the full problem in the next section.

4.4.2 Mathematical formulation of the full problem: The case of more than one expenditure

To formulate the full problem, recall that our primary goal is to estimate the average expense incurred on item k in the second interview reference period. We can denote this population quantity as $\bar{y}_k$. Because we have stratified the population into H strata based on their anticipated likelihood of incurring expense k during the second interview reference period, we can express $\bar{y}_k$ as a function of the stratum-specific means. This is given as

$$\bar{y}_k = \sum_{h=1}^{H} W_{hk} \bar{y}_{hk} \tag{4.14}$$

where $W_{hk} = N_{hk}/N$ with N_{hk} being the number of population units in stratum hk; $\bar{y}_{hk} = N_{hk}^{-1} \sum_{i=1}^{N_{hk}} y_{hik}$; and, y_{hik} is the expense for item k of the i^{th} unit in stratum hk. It is essential to make an important point about notation here. We use the pair of subscripts hk because expense k determines the H strata for that expense.

The general rules (e.g., estimators and variance formula) associated with a stratified sample design are as follows[12]. The full sample estimator for the mean of expense k is

$$\hat{\bar{y}}_k = \sum_{h=1}^{H} W_{hk} \hat{\bar{y}}_{hk} \tag{4.15}$$

where $\hat{\bar{y}}_{hk}$ is a stratum-specific mean estimator for $\bar{y}_{hk}$.

Since sampling is done independently within each stratum, we can express the

[12]By stratified sample design we mean that a probability sample is selected according to an arbitrary design independently within each stratum.

variance of $\hat{\bar{y}}_k$, with respect to the stratified sampling design, as

$$V_{ST}(\hat{\bar{y}}_k) = \sum_{h=1}^{H} W_{hk}^2 V_{hk}(\hat{\bar{y}}_{hk}) \tag{4.16}$$

where $V_{hk}(\hat{\bar{y}}_{hk})$ is the sampling variance of $\hat{\bar{y}}_{hk}$ according to the probability sampling design in the hk^{th} stratum.

Using the results presented in Section 2.4.2.2 and under the situation of Bernoulli sampling within each stratum with stratum-specific inclusion probabilities denoted as f_{hk}, we can rewrite the quantity given in equation (4.15) as follows.

$$\begin{aligned} \hat{\bar{y}}_k &= \sum_{h=1}^{H} W_{hk} \hat{\bar{y}}_{hk} \\ &= \sum_{h=1}^{H} W_{hk} \left(\sum_{i \in A_{hk}} f_{hk}^{-1} \right)^{-1} \sum_{i \in A_{hk}} f_{hk}^{-1} y_{hik} \end{aligned} \tag{4.17}$$

In equation (4.17), A_{hk} represents the set of sample units selected from the hk^{th} stratum. We note that the formula for $\hat{\bar{y}}_{hk}$ substituted in the equation above also coincides with equation (2.4). Furthermore, under a Bernoulli sampling design within each stratum, we actually have a Poisson sampling design across the strata since the stratum-specific inclusion probabilities vary across the strata.

Additionally, making the same assumption of Bernoulli sampling within each stratum, we can rewrite the quantity given in equation (4.16) using the result pre-

sented in equation (2.8) of Section 2.4.2.2 as follows.

$$
\begin{aligned}
V_{ST}(\hat{\bar{y}}_k) &= \sum_{h=1}^{H} W_{hk}^2 V_{hk}(\hat{\bar{y}}_{hk}) \\
&= \sum_{h=1}^{H} W_{hk}^2 \frac{1}{N_{hk}^2} \sum_{i=1}^{N_{hk}} \left(\frac{1 - f_{hk}}{f_{hk}} \right) (y_{hik} - \bar{y}_{hk})^2 \\
&= \sum_{h=1}^{H} \left(\frac{N_{hk}}{N} \right)^2 \frac{1}{N_{hk}^2} \sum_{i=1}^{N_{hk}} \left(\frac{1 - f_{hk}}{f_{hk}} \right) (y_{hik} - \bar{y}_{hk})^2 \\
V_{ST}(\hat{\bar{y}}_k) &= \frac{1}{N^2} \sum_{h=1}^{H} \sum_{i=1}^{N_{hk}} \left(\frac{1 - f_{hk}}{f_{hk}} \right) (y_{hik} - \bar{y}_{hk})^2
\end{aligned} \tag{4.18}
$$

If we take the special case of $f_{hk} = n_{hk}/N_{hk}$ with n_{hk} denoting the number of units sampled from stratum hk (and subsequently asked about expense k), then we can rewrite $\hat{\bar{y}}_k$ as follows.

$$
\begin{aligned}
\hat{\bar{y}}_k &= \sum_{h=1}^{H} W_{hk} \left(\sum_{i \in A_{hk}} f_{hk}^{-1} \right)^{-1} \sum_{i \in A_{hk}} f_{hk}^{-1} y_{hik} \\
&= \sum_{h=1}^{H} W_{hk} \left(\sum_{i \in A_{hk}} \frac{N_{hk}}{n_{hk}} \right)^{-1} \sum_{i \in A_{hk}} \frac{N_{hk}}{n_{hk}} y_{hik} \\
&= \sum_{h=1}^{H} W_{hk} \frac{1}{N_{hk}} \sum_{i=1}^{n_{hk}} \frac{N_{hk}}{n_{hk}} y_{hik} \\
&= \sum_{h=1}^{H} \frac{N_{hk}}{N} \frac{1}{N_{hk}} \sum_{i=1}^{n_{hk}} \frac{N_{hk}}{n_{hk}} y_{hik} \\
&= \frac{1}{N} \sum_{h=1}^{H} \sum_{i=1}^{n_{hk}} \frac{N_{hk}}{n_{hk}} y_{hik} \\
\hat{\bar{y}}_k &= \frac{1}{N} \sum_{h=1}^{H} \sum_{i=1}^{n_{hk}} y_{hik}/f_{hk}
\end{aligned} \tag{4.19}
$$

In this derivation, we used the following key relationships:

1. there are n_{hk} sample units in the set A_{hk};

2. for simplicity take $w_i = 1$ and note that $p_{ik} = f_{hk}$ for all $i \in A_{hk}$, then we have $\sum_{h=1}^{H}\sum_{i=1}^{n_{hk}} = \sum_{i \in S_k}$ which is the summation over the sample units receiving the k^{th} expenditure question; and,

3. $\sum_{i \in S_k} w_i^* = \sum_{i \in S_k} 1/p_{ik} = \sum_{h=1}^{H}\sum_{i=1}^{n_{hk}} 1/f_{hk} = \sum_{h=1}^{H}\sum_{i=1}^{n_{hk}} \frac{N_{hk}}{n_{hk}} = \sum_{h=1}^{H} N_{hk}\frac{n_{hk}}{n_{hk}} = \sum_{h=1}^{H} N_{hk} = N.$

This proves the equivalence of (4.17) and (2.9); thus, using the stratified sampling estimator for the mean presented in equation (4.17) is consistent with using the estimator given in equation (2.9) presented in Section 2.4.2.2.

With the same special case of $f_{hk} = n_{hk}/N_{hk}$, we can also rewrite $V_{ST}(\hat{\bar{y}}_k)$ as follows.

$$\begin{aligned}
V_{ST}(\hat{\bar{y}}_k) &= \frac{1}{N^2}\sum_{h=1}^{H}\sum_{i=1}^{N_{hk}} \left(\frac{1 - n_{hk}/N_{hk}}{n_{hk}/N_{hk}}\right)(y_{hik} - \bar{y}_{hk})^2 \\
&= \frac{1}{N^2}\sum_{h=1}^{H}\sum_{i=1}^{N_{hk}} \left(\frac{1 - n_{hk}/N_{hk}}{n_{hk}/N_{hk}}\right)\frac{N_{hk}}{N_{hk}}(y_{hik} - \bar{y}_{hk})^2 \\
&= \frac{1}{N^2}\sum_{h=1}^{H} \left(\frac{1 - n_{hk}/N_{hk}}{n_{hk}/N_{hk}}\right)\frac{N_{hk}}{N_{hk}}\sum_{i=1}^{N_{hk}}(y_{hik} - \bar{y}_{hk})^2 \\
&= \frac{1}{N^2}\sum_{h=1}^{H} \left(\frac{1 - n_{hk}/N_{hk}}{n_{hk}/N_{hk}}\right) N_{hk} S_{hk}^2 \\
&= \frac{1}{N^2}\sum_{h=1}^{H} N_{hk}\left(1 - \frac{n_{hk}}{N_{hk}}\right)\frac{N_{hk}}{n_{hk}} S_{hk}^2 \\
&= \sum_{h=1}^{H} \left(\frac{N_{hk}}{N}\right)^2 \left(1 - \frac{n_{hk}}{N_{hk}}\right) S_{hk}^2/n_{hk} \\
V_{ST}(\hat{\bar{y}}_k) &= \sum_{h=1}^{H} W_{hk}^2 \left(\frac{1}{n_{hk}} - \frac{1}{N_{hk}}\right) S_{hk}^2 \qquad (4.20)
\end{aligned}$$

In third line of the derivation above, we defined $S_{hk}^2 = N_{hk}^{-1}\sum_{i=1}^{N_{hk}}(y_{hik} - \bar{y}_{hk})^2$ which is

essentially equivalent to the standard formula of $S_{hk}^2 = (N_{hk} - 1)^{-1} \sum_{i=1}^{N_{hk}} (y_{hik} - \bar{y}_{hk})^2$.

The result in equation (4.20) allows us to set up the standard optimal allocation problem for stratified sampling using the variance represented by equation (4.20) in the objective function. Recall from Section 2.5.2 that it is typical in these problems to minimize the sampling variance subject to various constraints (Valliant and Gentle, 1997). The sampling variance in equation (4.20) represents the variance of our responsive split questionnaire design when stratification methods have been used to stratify the units on the basis of their likelihood of incurring an expense.

So, the full specification of the problem is as follows. Consider the situation in which we have K expenditure categories collected in our survey. Prior to the first phase of data collection, we draw a sample of N units from a population, U, via an arbitrary sample design. In the first phase of data collection, we collect information on the N units that will enable us to stratify them into H groups based on their likelihood of incurring expense k. So, for the first expenditure ($k = 1$), we will have a stratification of the N units. For the second expenditure ($k = 2$), we will have a stratification for the N units and so on until we have stratified the N units on the basis of their likelihood of incurring each of the K expenses. Ultimately, we will have $H \times K$ strata and we can view these $H \times K$ strata as defining a population of $N \times K$ units. We illustrate this stratification setup for the case of two strata per expenditure category in Table 4.11. As with the simplified version of the problem, we compute the S_{hk} values from the reported expenditure information collected

during the first interview.

Expenditure	Stratum	N_{hk}	S_{hk}	n_{hk}
TX2 ($k = 1$)	Low ($h = 1$)	5,258	0.00	$\cdots$
	High ($h = 2$)	5,237	521.27	$\cdots$
TX3F ($k = 2$)	Low ($h = 1$)	5,322	0.00	$\cdots$
	High ($h = 2$)	5,173	593.10	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
TX19B ($k = 36$)	Low ($h = 1$)	5,322	0.00	$\cdots$
	High ($h = 2$)	5,173	593.10	$\cdots$
Total	$H \times K$	$N \times K$		

Table 4.11: Illustration of stratification setup

At this point it is worth mentioning a slight deviation in the formulation of the full problem from the standard stratified population setup. In standard problems, the strata are non-overlapping, meaning that a unit will not be contained in more than one stratum. However, in our problem, each k^{th} set of H strata contains the same units. We formulate the problem this way mostly out of convenience as this allows us to set up and solve one optimization problem as opposed to K separate problems.

Given the set of strata defined by the K expenditure categories, we can determine the optimal number of units to ask each expenditure question from the respective $H \times K$ strata using standard (optimal) stratified sampling techniques. Following the basic approach to optimization outlined in Section 2.5.2, this amounts to finding the set of $\{n_{hk} : h = 1, 2, \ldots, H; k = 1, 2, \ldots K\}$, where n_{hk} is the number of sample units drawn from the hk^{th} stratum (and subsequently asked about expense k), that

minimizes an objective function subject to various constraints. Since we have more than one expenditure that we are interested in, we take our full objective function to be the sum of the variances of the K expenditure means. Summing across key estimands of interest for the objective function is consistent with other allocation problems in survey sampling (Valliant and Gentle; 1997). Specifically, the objective function, denoted as Φ, which we seek to minimize, becomes the following.

$$\Phi = \sum_{k=1}^{K} V_{ST}(\hat{\bar{y}}_k) = \sum_{k=1}^{K} \sum_{h=1}^{H} W_{hk}^2 \left(\frac{1}{n_{hk}} - \frac{1}{N_{hk}} \right) S_{hk}^2 \tag{4.21}$$

We also note that this objective function is nonlinear with respect to the decision variables, the set of $\{n_{hk}\}$.

The possible constraints for the full optimization problem may include the following[13].

1. The sample sizes from each stratum are less than the number of population units in that stratum.

$$n_{hk} \leq N_{hk} \text{ for all } h = 1, 2, \ldots, H \text{ and } k = 1, 2, \ldots, K \tag{4.22}$$

   It is worth noting that as the number of strata and the number of expenditures each gets large, the number of constraints represented by equation (4.22) can increase quite rapidly. For many optimization software packages, there is a

[13]We note that this is not an exhaustive list of the possible constraints one may consider for this type of responsive split questionnaire. We believe this list does, in fact, reflect many of the complexities of the full problem and using any subset of these constraints will yield useful and informative results.

limit on the number of constraints one can impose on the system[14]. Therefore, if access to software packages that can handle the number of constraints is limited, then it might be useful to reduce the dimension of the constraint vector. An equivalent, one-dimensional constraint to those represented by equation (4.22) is to specify the following constraint.

$$\max_{h,k} [n_{hk} - N_{hk}] \leq 0 \tag{4.23}$$

This type of dimension reduction can be made for several types of constraints, but we only illustrate for this case.

2. There is a minimum number of times, denoted as $n_{k,min}$, that the question about expenditure k is asked

$$\sum_h \sum_i \alpha_{hik} = \sum_h \alpha_{h+k} = \sum_h n_{hk} = n_{+k} \geq n_{k,min} \tag{4.24}$$

where $\alpha_{hik} = 1$ if sample unit i, in stratum hk, gets asked question k and 0 otherwise.

3. The total cost (C) across all units and due to all questions (with C_0 being a fixed cost) does not exceed some value

$$C = C_0 + \sum_k \sum_h c_h n_{hk} \tag{4.25}$$

[14]This is the case with Microsoft Excel 2007 Solver as the maximum number of allowable constraints is 100.

where c_h is the cost associated with asking the question to the units in stratum h. The lack of the subscript k on the cost parameters is meant to indicate that stratum-specific costs are independent of the expenditure. Furthermore, as we mentioned in the previous section, the cost function does not have to be used to impose constraints on data collection costs (e.g., dollar amounts). In our formulation of the full problem, we use the "cost" function as a penalty function such that a greater "cost" (i.e., penalty) is incurred by making an incorrect decision to ask the question about expense k. Effectively, we assign a higher "cost" to the lowest likelihood strata and thus a greater penalty will be incurred by asking questions about expenditure k to the unlikely purchaser of expenditure k.

4. The total burden (TB), as measured by the total number of questions asked across all units, does not exceed some value, denoted as b_{max},

$$TB = \sum_k \sum_h \sum_i \alpha_{hik} = \sum_k \alpha_{++k} = \sum_k n_{+k} = n_{++} \leq b_{max} \tag{4.26}$$

5. The total interview minutes (TT) across all units and all questions does not exceed some value, denoted as t_{max},

$$TT = \sum_k \sum_h \sum_i \alpha_{hik} t_k = \sum_k t_k \sum_h \sum_i \alpha_{hik} = \sum_k \alpha_{++k} = \sum_k t_k n_{+k} \leq t_{max} \tag{4.27}$$

where t_k is the time it takes to administer the question about expense k.

6. There are CV targets for certain variables.

$$\left[CV(\hat{\bar{y}}_k)\right]^2 \leq \left[CV_{0k}\right]^2 \tag{4.28}$$

Once a solution to the optimization problem is found, the decision rules for the responsive split questionnaires using stratification methods become functions of the set of $\{n_{hk} : h = 1, 2, \ldots, H; k = 1, 2, \ldots K\}$. In particular, the decision rules are the stratum-specific sampling fractions given as $f_{hk} = n_{hk}/N_{hk}$ for every i in stratum hk i.e., $p_{ik} = f_{hk}$.

4.4.3 Two-Bin stratification

The first responsive split questionnaire using stratification methods that we explore is the situation in which we have two strata per expenditure category. One strata reflects a "low" anticipated likelihood of incurring the expense and one reflects a "high" anticipated likelihood of incurring the expense. In Section 4.3.3.1, we noted that the estimated value of $p_{Int2,ik}$, denoted as $\hat{p}_{Int2,ik}$ for every sample unit and for each expenditure k based on the model presented in equation (4.11) should be a fairly close approximation to the probability that the unit will incur the expense in the second interview reference period since we explicitly model the event. Thus, we can use these same methods to estimate the desired probability and then based on the estimated probability classify the units of the analysis file into one of the two strata. We do this by using the median of $\hat{p}_{Int2,ik}$ for each expenditure k.

So our stratification was performed as follows. If $\hat{p}_{Int2,ik} \leq \text{med}(\hat{p}_{Int2,ik})$ then

the unit was classified into the "Low" stratum for that expenditure because it was thought to have a low propensity of incurring the expense. Conversely, if $\hat{p}_{Int2,ik} > \text{med}(\hat{p}_{Int2,ik})$, then the unit was classified into the "High" stratum for that expenditure because it was thought to have a high propensity of incurring the expense. Thus, the strata information are a combination of data from the first and second interviews. The "Low/High" classification is based on the estimated probability of purchase in the second interview while the stratum standard deviations are computed directly from the first interview data. We refer to this method as the **Two-Bin** stratification method because we have two strata per expenditure category. We have the full listing of stratification classifications for the sample units in the analysis file in Tables D.5 and D.6 in Appendix D.3[15]. As a final point, because there is an intermediate step in devising the decision rules, the responsive split questionnaires using stratification methods conforms with the indirect perspective on developing decision rules.

Once we classified each unit into their respective strata for each expenditure, we assigned cost parameters of 1 and 9 to the "High" and "Low" strata, respectively. We then used Microsoft Excel 2007 Solver to determine the set of $\{n_{hk}\}$ such that equation (4.21) was minimized subject to the constraints listed in equations (4.23), (4.24), and (4.25). We required at least 400 sample units to get asked each question and our "budget" for this problem was 250,000. We also set a minimum number of 30 units sampled from each stratum to be consistent with the

[15]In the last column of these tables we also display the number of units in each stratum that actually incurred the expense in the second interview reference period. This is meant to provide a check on the performance of the stratification mechanism.

modification we made to the decision rules for the previous three methods in order to keep the design measurable. In Microsoft Excel 2007 Solver, we specified the following options: (1) quadratic estimates (recommended for highly nonlinear problems); (2) central differencing (requires more time per iteration but may lead to better solutions); and (3) conjugate searching (useful for large problems). After over 50,000 iterations, Microsoft Excel 2007 Solver found a solution to the optimization problem, and we display the output from the solution in the next section and offer a few comments/observations.

It is worth noting that at this stage we did not consider the full listing of constraints identified in equations (4.23) – (4.28). This is for comparability of the methods explored in Chapter 4 to be as comparable as possible. For example, for the PPS, Log1, and Log2 methods, we did not specify precision targets for key estimates. Thus, we did not include them in this iteration of developing decision rules for a responsive split questionnaire using stratification methods. It is a relatively simple extension to incorporate these into the full problem specification in Microsoft Excel 2007 Solver provided that the number of constraints does not exceed the maximum allowable, 100, in Microsoft Excel 2007 Solver[16].

Finally, we note that some bounds used in the constraints as well as certain parameters (e.g., cost parameters) are adjustable. In our research, we chose the values to illustrate the method, but in general, the values for the constraint bounds and parameters depend, in part, on the nature of the problem. For example, we

[16]If this is the case, then there are other software packages available that will allow for a greater number of constraints.

required at least 400 sample units to get asked each question. We could have chosen any value for this constraint, but as indicated by Table 2.3, 400 is the minimum sample size required for estimating the prevalence of a characteristic with at least a 10% CV (assuming SRS) for characteristics with prevalence of at least 20%. A vast majority of the expenditures we considered in this dissertation have prevalence rates greater than this value. In addition, we specified at least 30 sample units being sampled from each stratum. The chosen value of 30 is consistent with the modification of the p_{ik} in the PPS, Log1, and Log2 methods to keep the design measurable and maintain the integrity of the responsive design aspect our split questionnaire design.

4.4.3.1 Optimization output

In Table 4.12, we display the output from the optimization problem. In particular, we show the stratum-specific sample sizes, n_{hk}, and other relevant quantities, e.g., stratum-specific population sizes, N_{hk} and standard deviations, S_{hk}. As we mentioned in Section 4.4.1, the general rules of optimal allocation (as per Cochran [1977]) dictate that in a given stratum take a larger sample if (1) the stratum is more variable internally, (2) sampling is cheaper in the stratum, and (3) the stratum is larger. To determine whether the optimization results for the Two-Bin stratification method are consistent with this theory, we computed high-level summary statistics for the set of $\{n_{hk}\}$.

First, we sorted the strata by their stratum-specific standard deviations, S_{hk},

Expenditure	Stratum	N_{hk}	S_{hk}	n_{hk}	Expenditure	Stratum	N_{hk}	S_{hk}	n_{hk}
TX2	Low	5,258	0.00	30	**TX9B**	Low	5,250	0.00	30
	High	5,237	521.27	4,238		High	5,245	100.90	3,512
TX3F	Low	5,322	0.00	30	**TX9C**	Low	5,299	0.00	30
	High	5,173	593.10	4,422		High	5,196	17.60	3,482
TX3H	Low	5,353	16.71	30	**TX9D**	Low	5,271	3.11	755
	High	5,142	286.63	1,237		High	5,224	21.06	126
TX4A_1	Low	5,315	75.27	343	**TX10C_1**	Low	5,249	0.00	30
	High	5,180	83.71	669		High	5,246	156.27	3,554
TX4A_2	Low	5,311	0.00	30	**TX10C_23**	Low	5,260	0.00	30
	High	5,184	19.69	3,482		High	5,235	647.38	4,612
TX4A_3	Low	5,252	0.00	30	**TX11B**	Low	5,325	0.00	30
	High	5,243	29.56	3,484		High	5,170	2,925.51	5,170
TX4A_4	Low	5,250	0.32	755	**TX12A**	Low	5,319	89.31	787
	High	5,245	3.19	33		High	5,176	310.03	812
TX4B	Low	5,329	0.00	30	**TX12B**	Low	5,302	35.18	760
	High	5,166	17.40	3,482		High	5,193	55.93	235
TX4C	Low	5,258	40.45	762	**TX12C_10**	Low	5,301	140.37	831
	High	5,237	39.96	208		High	5,194	170.71	463
TX4D	Low	5,252	154.35	845	**TX12C_1TO9**	Low	5,249	0.96	755
	High	5,243	164.91	452		High	5,246	3.86	34
TX5A_1	Low	5,340	162.68	3,751	**TX13B**	Low	5,252	65.38	772
	High	5,155	0.00	30		High	5,243	436.81	1,257
TX5A_2	Low	5,264	12.51	756	**TX14B**	Low	5,255	0.00	30
	High	5,231	38.09	203		High	5,240	253.58	3,671
TX5B	Low	5,288	259.06	961	**TX15A**	Low	5,253	0.00	30
	High	5,207	2,527.69	5,207		High	5,242	337.68	3,813
TX6A	Low	5,299	80.50	781	**TX16A**	Low	5,287	0.00	30
	High	5,196	262.39	692		High	5,208	822.24	5,208
TX6B	Low	5,273	20.56	757	**TX17A**	Low	5,322	1.24	755
	High	5,222	399.98	1,105		High	5,173	113.15	323
TX7A	Low	5,353	1.53	755	**TX18A**	Low	5,349	6.92	755
	High	5,142	72.70	253		High	5,146	131.87	364
TX8	Low	5,294	0.00	30	**TX19A**	Low	5,249	129.78	819
	High	5,201	439.43	4,022		High	5,246	341.71	915
TX9A	Low	5,255	74.20	776	**TX19B**	Low	5,248	23.23	757
	High	5,240	235.27	629		High	5,247	522.97	3,649

Table 4.12: Optimization output for the Two-Bin stratification method

computed the average of n_{hk} separately for the lowest half and highest half of the strata and then compared the two averages. The average of n_{hk} for the lowest half was 648 while the average of n_{hk} for the largest half was $1,976$. We interpret this finding as being consistent with (1) above. Specifically, the largest half of the strata, with respect to S_{hk}, consist of the more internally variable strata. Among these strata, on average, the optimization solution yields a higher allocation of our sample to these strata.

Next, we computed the average of n_{hk} separately for the "Low" and "High" strata. Recall that a higher cost is associated with sampling from the "Low" strata. The average of n_{hk} for the "Low" strata is 540 while the average for the "High" strata is $2,085$. This is consistent with (2) above since a larger sample size is allocated, on average, to the cheaper strata.

Finally, we sorted the strata on the basis of their respective N_{hk} values. It turns out that the set corresponding to the largest strata in terms of N_{hk} coincide exactly with the "Low" strata. This finding is not surprising since the "Low" strata should contain the unlikely purchasers of the particular items. For many expenses, the prevalence of incurring the expense is low. Thus, there should be more unlikely purchasers than likely purchasers and as a consequence, a greater number of units in the "Low" strata. At any rate, the optimization results seem to contradict theory (e.g., take a large sample from larger strata). One explanation for this may be that the rule is being ignored because these strata are expensive. Another explanation may be that the discrepancies among the strata sizes are not large enough to affect the allocation. Nonetheless, based on these high-level summary statistics, we con-

clude that the optimization results generally conform with the theory for optimal allocation.

4.4.3.2 Simulation setup

To evaluate the performance of the Two-Bin stratification method, we carried out a simulation with $(M =)1,000$ iterations. For each iteration, we randomly "asked" sample units questions based on their respective $f_{hk} = n_{hk}/N_{hk}$ values, derived by the appropriate quantities in Table 4.12. We then computed the quantities given in equations (3.7) – (3.14) to summarize the simulations and also computed similar summary statistics for the domains defined by the CUs that incurred the expense. We then computed the average number of times the question was administered in the second interview, the average number of questions asked, the average time spent answering a question, the average interview length, percent reduction in interview length, and the average and median design effects. Finally, we computed the average sensitivity, specificity, PPV, and NPV for this method.

4.4.3.3 Results

In Table 4.13, we display the summary statistics for the Two-Bin stratification method. We observe the following trends for this method. First, the average number of times an expenditure question is asked tends to be correlated with the variance of the mean expenditure. Specifically, the more variable the expenditure is, the more times we ask about it. This is because this method, in part, allocates a higher

proportion of the sample to more variable expenditure categories. The average number of times we ask about each expenditure category ranges from about 790 to 6,170. Recall that for this method we specified the minimum number of times each expenditure was asked to be no less than 400, so we expect to ask about each expenditure to at least 400 sample units.

The average number of times we ask about each expenditure may also provide us with evidence that this method does not do a good job of asking about expenditures that the sample unit is likely to incur. We infer this simply because expenses like TX4A_1 (telephone services) and TX4D (utilities and fuels), which have over 90% prevalence rates, only get asked, on average, slightly more than 1,000 times. One reason for this may be due to how the full problem was speficied. These expenditures have lower S_{hk} values relative to the remaining expenditure categories. As a consequence, less sample is being allocated to these strata.

Although the magnitudes of some of the relative bias calculations may indicate the potential for biased estimates, the confidence intervals include zero. Therefore, we cannot conclude that any are statistically different from zero. So, we are still able to obtain design-unbiased estimates after accounting for the responsive split questionnaire design.

The simulation CVs under this method range from 0.02 to 2.17 with 15 expenditure categories exhibiting CVs of 0.10 or less. As with the previous three methods, the two highest CVs correspond to TX4A_4 (modem purchases, apps, ringtones) and TX5A_1 (construction materials for specific jobs). Their CVs are 2.17 and 1.60, respectively.

Finally, the design effects under this method range from 0.93 to over 50. In fact, only three expenditure estimates result in design effects of less than 1 while 10 have design effects greater than 10. Under traditional situations, this would potentially indicate severe losses in efficiency relative to simple random sampling. Since CVs and design effects both deal with the precision of estimates, one way to address these concerns may be to add constraints for CV targets and resolve the optimization problem. This might rein in the CVs and the design effects for this method.

Expenditure	Asked	Mean	Variance	Std Err	Sim CV	Rel Bias	Rel Bias SE	RMSE	Bias LB	Bias UB	deff
TX2	4,269.15	631.00	3,242.20	56.94	0.09	0.44	9.06	57.01	-17.32	18.21	9.34
TX3F	4,453.32	216.32	828.05	28.78	0.13	0.14	13.32	28.78	-25.97	26.25	2.41
TX3H	1,267.34	103.94	309.36	17.59	0.17	-0.18	16.89	17.59	-33.29	32.92	0.97
TX4A_1	1,011.19	350.16	63.42	7.96	0.02	0.02	2.27	7.96	-4.43	4.48	1.10
TX4A_2	3,512.63	27.63	10.02	3.16	0.11	0.28	11.49	3.17	-22.23	22.80	14.69
TX4A_3	3,512.25	24.08	10.33	3.21	0.13	1.02	13.48	3.22	-25.41	27.44	8.24
TX4A_4	788.26	0.10	0.05	0.21	2.17	-10.39	194.73	0.21	-392.05	371.28	3.91
TX4B	3,512.23	5.00	2.04	1.43	0.29	0.71	28.77	1.43	-55.67	57.09	13.99
TX4C	969.07	61.74	7.64	2.76	0.04	0.05	4.48	2.76	-8.72	8.83	1.60
TX4D	1,296.74	608.37	127.97	11.31	0.02	0.02	1.86	11.31	-3.63	3.66	1.08
TX5A_1	3,780.72	6.55	109.28	10.45	1.60	3.10	164.62	10.46	-319.56	325.76	54.20
TX5A_2	957.37	1.38	0.90	0.95	0.69	-2.57	66.87	0.95	-133.64	128.50	1.90
TX5B	6,167.75	359.45	204.74	14.31	0.04	0.30	3.99	14.35	-7.53	8.12	1.06
TX6A	1,472.97	38.67	31.32	5.60	0.14	0.16	14.50	5.60	-28.25	28.57	1.02
TX6B	1,862.53	183.47	73.71	8.59	0.05	-0.12	4.67	8.59	-9.28	9.05	0.93
TX7A	1,010.04	16.56	7.60	2.76	0.17	1.40	16.88	2.77	-31.68	34.48	1.64
TX8	4,049.38	132.09	852.82	29.20	0.22	0.57	22.23	29.21	-43.01	44.15	20.28
TX9A	1,404.91	202.78	54.66	7.39	0.04	-0.10	3.64	7.40	-7.24	7.04	1.00
TX9B	3,542.50	36.43	74.31	8.62	0.24	0.75	23.84	8.62	-45.97	47.48	18.37
TX9C	3,510.64	4.02	3.96	1.99	0.50	2.96	51.06	1.99	-97.11	103.03	23.60
TX9D	879.64	3.86	2.41	1.55	0.40	-0.98	39.81	1.55	-79.02	77.05	3.25
TX10C_1	3,584.90	20.24	9.70	3.11	0.15	0.64	15.49	3.12	-29.72	31.00	2.90
TX10C_23	4,641.66	50.17	41.10	6.41	0.13	0.50	12.84	6.42	-24.67	25.67	1.13
TX11B	5,200.15	696.55	5,012.65	70.80	0.10	0.46	10.21	70.87	-19.55	20.48	5.73
TX12A	1,599.02	165.55	66.30	8.14	0.05	-0.17	4.91	8.15	-9.79	9.46	1.05
TX12B	994.52	29.87	7.94	2.82	0.09	0.15	9.45	2.82	-18.36	18.66	1.60
TX12C_1TO9	789.69	24.39	58.41	7.64	0.31	-0.27	31.24	7.64	-61.51	60.97	7.89
TX12C_10	1,295.74	188.46	22.13	4.70	0.02	-0.01	2.50	4.70	-4.90	4.88	1.18
TX13B	2,027.67	365.49	81.74	9.04	0.02	0.09	2.48	9.05	-4.76	4.94	0.94
TX14B	3,700.05	62.50	95.72	9.78	0.16	1.37	15.87	9.82	-29.73	32.48	10.20
TX15A	3,842.28	247.81	1,109.86	33.31	0.13	0.95	13.57	33.40	-25.65	27.54	19.59
TX16A	5,238.26	232.72	3,785.66	61.53	0.26	1.23	26.76	61.59	-51.23	53.68	22.29
TX17A	1,078.12	51.25	28.19	5.31	0.10	-0.80	10.28	5.33	-20.95	19.34	1.47
TX18A	1,120.10	9.83	9.06	3.01	0.31	-0.14	30.59	3.01	-60.11	59.82	1.13
TX19A	1,735.56	228.94	91.03	9.54	0.04	0.34	4.18	9.57	-7.85	8.54	1.00
TX19B	4,404.88	369.73	80.26	8.96	0.02	0.06	2.42	8.96	-4.70	4.81	1.30

Table 4.13: Simulation summary statistics for the Two-Bin stratification method

In Table 4.14, we present the simulation summary statistics for the domain characteristics where the domain is defined by those CUs incurring the expense. The average number of times the question is asked and the sample unit incurred it ranges from 1.82 to about 3,370. As with the Log2 method, we observe fewer than 100 instances of incurring the expenditure for seven expenditure categories. We are also observing fewer instances of purchasing for the more prevalent expenditure categories, e.g., TX4A_1 (telephone services) than we we would expect if we asked about the expense to all likely purchasers. Again, this is likely a consequence of asking about these expenditures to so few sample units.

All 95% confidence intervals associated with the relative bias calculations for the domain means include zero, therefore we cannot conclude that any are statistically different from zero. The simulation CVs for this method range from 0.01 to 0.92 with 21 expenditure categories exhibiting CVs less than or equal to 0.1. So, for more than half of our expenditure categories we have evidence of the ability to obtain fairly precise estimates of the domain means.

Expenditure	Asked & Have	Mean	Variance	Std Err	Sim CV	Rel Bias	Bias SE	RMSE	Bias LB	Bias UB
TX2	2,371.49	2,241.53	561.67	23.70	0.01	-0.01	1.06	23.70	-2.09	2.06
TX3F	406.16	4,636.10	66,908.78	258.67	0.06	0.41	5.60	259.36	-10.57	11.39
TX3H	135.95	1,856.56	34,752.11	186.42	0.10	0.06	10.05	186.42	-19.63	19.75
TX4A_1	954.74	378.50	59.51	7.71	0.02	0.03	2.04	7.71	-3.96	4.03
TX4A_2	1,667.34	102.90	12.10	3.48	0.03	0.03	3.38	3.48	-6.59	6.66
TX4A_3	817.97	183.18	75.53	8.69	0.05	0.07	4.75	8.69	-9.24	9.37
TX4A_4	1.82	49.17	1,342.31	36.64	0.75	0.85	75.15	36.64	-146.45	148.15
TX4B	436.19	60.35	98.97	9.95	0.16	0.37	16.54	9.95	-32.05	32.80
TX4C	478.46	94.78	12.10	3.48	0.04	0.08	3.67	3.48	-7.12	7.28
TX4D	1,171.51	660.93	129.84	11.39	0.02	0.03	1.72	11.40	-3.35	3.41
TX5A_1	51.34	337.20	96,391.81	310.47	0.92	-6.44	86.14	311.34	-175.28	162.40
TX5A_2	6.73	146.64	6,622.13	81.38	0.55	-3.29	53.67	81.53	-108.48	101.90
TX5B	1,495.93	1,979.82	5,035.49	70.96	0.04	0.27	3.59	71.17	-6.77	7.32
TX6A	87.86	639.79	4,716.81	68.68	0.11	0.11	10.75	68.68	-20.95	21.17
TX6B	863.99	417.95	287.75	16.96	0.04	-0.07	4.06	16.97	-8.02	7.88
TX7A	77.10	168.09	412.28	20.30	0.12	0.71	12.17	20.34	-23.14	24.55
TX8	1,934.88	355.56	5,170.04	71.90	0.20	0.26	20.28	71.91	-39.48	40.00
TX9A	921.08	303.31	96.07	9.80	0.03	-0.02	3.23	9.80	-6.35	6.31
TX9B	1,136.55	160.03	956.27	30.92	0.19	0.49	19.42	30.93	-37.57	38.55
TX9C	333.33	58.33	385.84	19.64	0.34	1.68	34.24	19.67	-65.43	68.79
TX9D	37.02	60.70	412.69	20.31	0.33	-0.39	33.34	20.32	-65.73	64.94
TX10C_1	275.55	506.00	830.87	28.82	0.06	0.22	5.71	28.85	-10.97	11.41
TX10C_23	152.24	2,985.97	9,739.01	98.69	0.03	0.30	3.31	99.09	-6.20	6.80
TX11B	1,822.24	3,901.24	20,009.35	141.45	0.04	-0.06	3.62	141.48	-7.17	7.04
TX12A	878.79	302.97	181.96	13.49	0.04	-0.06	4.45	13.49	-8.78	8.66
TX12B	248.85	107.18	56.40	7.51	0.07	0.26	7.02	7.52	-13.51	14.02
TX12C_1TO9	167.09	89.99	599.37	24.48	0.27	0.00	27.21	24.48	-53.32	53.32
TX12C_10	1,126.89	211.16	24.09	4.91	0.02	-0.03	2.32	4.91	-4.58	4.53
TX13B	1,413.97	562.30	137.10	11.71	0.02	-0.01	2.08	11.71	-4.09	4.07
TX14B	1,562.30	232.39	747.71	27.34	0.12	1.08	11.89	27.46	-22.23	24.39
TX15A	3,139.50	400.65	2,048.41	45.26	0.11	0.50	11.35	45.30	-21.75	22.76
TX16A	2,056.09	1,004.06	44,362.76	210.62	0.21	0.92	21.17	210.82	-40.58	42.41
TX17A	290.17	153.70	193.50	13.91	0.09	-0.73	8.98	13.96	-18.34	16.88
TX18A	29.60	325.26	7,629.98	87.35	0.27	-0.46	26.73	87.36	-52.86	51.93
TX19A	1,192.38	338.87	165.57	12.87	0.04	0.27	3.81	12.90	-7.19	7.73
TX19B	3,373.87	606.17	152.02	12.33	0.02	0.04	2.03	12.33	-3.95	4.02

Table 4.14: Simulation summary statistics for the domains using the Two-Bin stratification method

In Table 4.15, we present the calculations for the four epidemiological criteria for the Two-Bin stratification method. We conclude that given a sample unit incurred the expense, this method does a fair job of asking CUs about expenses that they are likely to have incurred. This does not contradict our earlier statement in which we claimed that the average number of times we ask about an expense provides evidence that this method does not do a good job of asking about expenses that the sample unit is likely to incur. The key here is that *if* the method can detect incurring the expense, then it will ask about it. About one-third of the expenditure categories had sensitivity values greater than 0.5.

In terms of specificity, the Two-Bin stratification method does a good job of correctly not asking sample units about expenses that they are not likely to have incurred. Of the 36 expenditure categories that we investigated, only one exhibited a specificity value less than 0.5. This expenditure category was TX5B as the specificity value for this expense was 0.456. One explanation for this may be because this method errs on the side of caution by not asking about certain expenses (primarily because the expenditures are of low variability).

The third evaluation criteria we report on is PPV. The findings for this criterion are mixed. On the one hand, for 17 expenditure categories, we achieve PPV values of 20% or greater than the associated prevalence rate for that expenditure. This suggests that for these expenses we are detecting at least 20% more instances of incurring the expense than we would by a completely random split questionnaire design. The expenditures for which the PPV is 20% or greater than the associated prevalence rate correspond to the 17 shaded cells under the PPV column of Ta-

Expenditure	Sensitivity	Specificity	PPV	NPV	$P(Int2)$
TX2	0.806	0.749	0.555	0.909	0.280
TX3F	0.827	0.595	0.091	0.986	0.047
TX3H	0.231	0.886	0.107	0.951	0.056
TX4A_1	0.098	0.928	0.944	0.077	0.925
TX4A_2	0.593	0.760	0.475	0.836	0.268
TX4A_3	0.598	0.705	0.233	0.921	0.130
TX4A_4	0.076	0.925	0.002	0.998	0.002
TX4B	0.504	0.681	0.124	0.938	0.083
TX4C	0.070	0.866	0.494	0.332	0.652
TX4D	0.121	0.850	0.903	0.077	0.921
TX5A_1	0.278	0.638	0.014	0.980	0.018
TX5A_2	0.069	0.909	0.007	0.990	0.009
TX5B	0.785	0.456	0.243	0.905	0.182
TX6A	0.139	0.860	0.060	0.939	0.060
TX6B	0.187	0.830	0.464	0.566	0.439
TX7A	0.075	0.901	0.076	0.900	0.098
TX8	0.498	0.680	0.478	0.697	0.370
TX9A	0.131	0.861	0.656	0.329	0.669
TX9B	0.477	0.703	0.321	0.821	0.227
TX9C	0.467	0.675	0.095	0.945	0.068
TX9D	0.055	0.914	0.042	0.934	0.064
TX10C_1	0.659	0.672	0.077	0.979	0.040
TX10C_23	0.865	0.565	0.033	0.996	0.017
TX11B	0.978	0.609	0.350	0.992	0.178
TX12A	0.153	0.848	0.550	0.453	0.547
TX12B	0.085	0.901	0.250	0.718	0.279
TX12C_1TO9	0.059	0.919	0.212	0.723	0.272
TX12C_10	0.120	0.851	0.870	0.104	0.892
TX13B	0.207	0.833	0.697	0.362	0.649
TX14B	0.555	0.722	0.422	0.816	0.268
TX15A	0.486	0.826	0.817	0.500	0.616
TX16A	0.848	0.606	0.393	0.930	0.231
TX17A	0.083	0.887	0.269	0.659	0.334
TX18A	0.094	0.893	0.026	0.969	0.030
TX19A	0.168	0.841	0.687	0.327	0.675
TX19B	0.527	0.748	0.766	0.503	0.610

Table 4.15: Epidemiological criteria for the Two-Bin stratification method

ble 4.15. On the other hand, we do slightly worse than the "flipping a coin" method for ten of the expenditure categories. This suggests that for these ten, we are better off "flipping a coin" to determine whether to ask the question.

The final criterion is NPV. As with PPV, these findings are mixed. Only five out of the 36 expenditure categories exhibit NPV values 20% or greater than one minus the associated prevalence rate for that expenditure. These are identified by the shaded cells in the NPV column of Table 4.15. This method does worse than the completely random split questionnaire design for 13 expenditure categories.

4.4.4 Five-Bin stratification

The second responsive split questionnaire using stratification methods that we explore is the case in which we have five strata per expenditure category. This is an extension of the Two-Bin method, but here we identify a greater number of strata for each expenditure, in part, to reflect a continuum of purchase behavior, rather than two discrete states of purchase behavior.

For this method, we used the same mechanism as with the Two-Bin method to stratify the units into their respective strata but used a greater number of cut points to delineate the strata bounds. So, our stratification was performed as follows. First, we estimated $p_{Int2,ik}$ for every sample unit and for each expenditure k. We then classified the members of the analysis file into five groups based on the quintiles of the $\hat{p}_{Int2,ik}$ values for each expenditure. In other words, if $\hat{p}_{Int2,ik}$ for sample unit i was less than the first quintile value of the $\hat{p}_{Int2,ik}$, then that unit was classified into

the "Lowest" stratum for that expenditure because he was thought to have a lowest propensity of incurring the expense. If $\hat{p}_{Int2,ik}$ for sample unit i was greater than the fifth quintile value of the $\hat{p}_{Int2,ik}$, then that unit was classified into the "Highest" stratum for that expenditure because he was thought to have a highest propensity of incurring the expense. The strata are named in terms of increasing quintiles. The names are as follows: (1) Lowest; (2) Low; (3) Medium; (4) High; and, (5) Highest. We refer to this as the **Five-Bin** stratification method because we have five strata per expenditure category. We have the full listing of stratification classifications for the sample units in the analysis file in Tables D.7 – D.11 in Appendix D.3[17].

Once we classified each unit into their respective strata for each expenditure, we assigned cost parameters of 1, 3, 5, 7, and 9 to the strata in the following order: (1) Highest; (2) High; (3) Medium; (4) Low; and, (5) Lowest. We note that with the highest anticipated likelihood of incurring the expense, the smallest penalty is assigned for asking the question about the expense. With the lowest likelihood of incurring the expense, the greatest penalty is assigned for asking the question about the expenditure. We then used Microsoft Excel 2007 Solver to determine the set of $\{n_{hk}\}$ such that equation (4.21) was minimized subject to the constraints listed in equations (4.23), (4.24), and (4.25). We required at least 400 sample units to get asked each question and our "budget" for this problem was 162,500. We set the budget constraint lower for this method since we have more units in the cheaper strata. We also set a minimum number of 30 units sampled from each stratum

[17]In the last column of these tables we also display the number of units in each stratum that actually incurred the expense in the second interview reference period. As with the Two-Bin method, this is meant to serve as a check on the performance of the stratification mechanism.

to be consistent with the modification we made to the decision rules for the other methods to keep the split questionnaire design measurable. In Microsoft Excel 2007 Solver, we specified the following options: (1) quadratic estimates (recommended for highly nonlinear problems); (2) central differencing (requires more time per iteration but may lead to better solutions); and (3) conjugate searching (useful for large problems). After over 10,000 iterations, Microsoft Excel Solver found a solution to the optimization problem. We display relevant output from the solution in the next section and offer a few comments.

4.4.4.1 Optimization output

In Tables 4.16 and 4.17, we display the output from the optimization problem for the Five-Bin stratification method. We show stratum-specific sample sizes and other relevant quantities. To determine whether these optimization results are consistent with optimal allocation theory as described in Section 4.4.3.1 (Cochran, 1977), we computed high-level summary statistics for the set of $\{n_{hk}\}$.

We first sorted the strata by the stratum standard deviation, S_{hk}, then computed the average of n_{hk} separately for the lowest half and highest half of the strata. The average of n_{hk} for the lowest half was 51 while the average of n_{hk} for the largest half was 638. This finding is consistent with allocating a larger portion of the sample to strata that are internally variable.

Next, we sorted the strata by their respective cost parameters, c_h, and computed the average of n_{hk} separately for the lowest half and the highest half. The

Expenditure	Stratum	N_h	S_h	n_h	Expenditure	Stratum	N_h	S_h	n_h
	Lowest	2102	0.00	30		Lowest	2103	97.92	212
	Low	2155	0.00	30		Low	2128	135.24	337
TX2	Medium	2049	0.00	30	**TX4D**	Medium	2087	161.70	471
	High	2107	401.07	1,789		High	2087	166.41	467
	Highest	2082	492.07	2,082		Highest	2090	169.49	591
	Lowest	2107	0.00	30		Lowest	2132	71.59	161
	Low	2099	0.00	30		Low	2079	109.32	241
TX3F	Medium	2127	16.46	118	**TX5A_1**	Medium	2152	220.31	536
	High	2113	62.55	228		High	2146	0.00	30
	Highest	2049	900.77	2,049		Highest	1986	0.00	30
	Lowest	2108	0.00	30		Lowest	2107	10.05	33
	Low	2099	26.67	32		Low	2092	9.54	57
TX3H	Medium	2097	0.00	30	**TX5A_2**	Medium	2099	61.61	274
	High	2175	0.00	30		High	2235	0.00	30
	Highest	2016	438.02	1,964		Highest	1962	0.00	30
	Lowest	2102	60.74	80		Lowest	2138	0.00	30
	Low	2057	73.95	108		Low	2110	409.11	907
TX4A_1	Medium	2215	74.04	110	**TX5B**	Medium	2122	0.00	30
	High	2080	80.98	109		High	2010	0.00	30
	Highest	2041	87.22	281		Highest	2115	3,848.30	2,115
	Lowest	2134	0.00	30		Lowest	2104	26.30	78
	Low	2119	0.00	30		Low	2096	121.30	265
TX4A_2	Medium	2057	0.00	30	**TX6A**	Medium	2025	104.64	249
	High	2087	18.78	385		High	2241	132.07	369
	Highest	2098	13.09	63		Highest	2029	376.54	1,258
	Lowest	2099	0.00	30		Lowest	2116	0.00	30
	Low	2129	0.00	30		Low	2119	32.01	96
TX4A_3	Medium	2087	0.00	30	**TX6B**	Medium	2112	56.13	157
	High	2111	0.00	30		High	2054	365.49	1,045
	Highest	2069	35.69	364		Highest	2094	473.70	1,759
	Lowest	2113	0.22	47		Lowest	2109	0.00	30
	Low	2098	0.00	30		Low	2108	0.44	41
TX4A_4	Medium	2097	0.89	56	**TX7A**	Medium	2229	2.37	55
	High	2159	2.91	52		High	1956	1.80	69
	Highest	2028	4.08	227		Highest	2093	108.53	320
	Lowest	2180	0.00	30		Lowest	2099	0.00	30
	Low	2076	0.00	30		Low	2121	0.00	30
TX4B	Medium	2064	0.00	30	**TX8**	Medium	2112	36.46	115
	High	2090	0.00	30		High	2079	324.67	920
	Highest	2085	25.70	359		Highest	2084	576.58	2,084
	Lowest	2101	0.00	30		Lowest	2103	0.00	30
	Low	2099	30.55	139		Low	2110	45.19	123
TX4C	Medium	2060	42.72	272	**TX9A**	Medium	2103	184.74	435
	High	2166	40.91	124		High	2095	224.79	601
	Highest	2069	37.95	590		Highest	2084	252.89	778

Table 4.16: Optimization output for the Five-Bin stratification method

Expenditure	Stratum	N_h	S_h	n_h	Expenditure	Stratum	N_h	S_h	n_h
	Lowest	2116	0.00	30		Lowest	2105	0.07	114
	Low	2085	0.00	30		Low	2103	0.15	169
TX9B	Medium	2113	0.00	30	**TX12C_1TO9**	Medium	2214	1.48	169
	High	2084	0.00	30		High	2099	1.25	199
	Highest	2097	142.65	543		Highest	1974	6.10	386
	Lowest	2080	0.00	30		Lowest	2105	0.00	30
	Low	2150	0.00	30		Low	2105	0.00	30
TX9C	Medium	2076	2.93	267	**TX13B**	Medium	2097	255.17	625
	High	2169	1.50	47		High	2092	477.05	1,100
	Highest	2020	27.14	181		Highest	2096	439.94	1,328
	Lowest	2146	1.41	46		Lowest	2111	0.00	30
	Low	2119	4.69	62		Low	2181	0.00	30
TX9D	Medium	2130	0.26	56	**TX14B**	Medium	2011	0.00	30
	High	2092	1.35	62		High	2057	0.00	30
	Highest	2008	32.92	242		Highest	2135	344.43	1,027
	Lowest	2088	0.00	30		Lowest	2099	0.00	30
	Low	2112	0.00	30		Low	2123	0.00	30
TX10C_1	Medium	2180	0.00	30	**TX15A**	Medium	2089	193.08	471
	High	2094	0.00	30		High	2105	348.58	712
	Highest	2021	240.06	792		Highest	2079	354.02	763
	Lowest	2123	0.00	30		Lowest	2136	0.00	30
	Low	2092	0.00	30		Low	2068	0.00	30
TX10C_23	Medium	2103	0.00	30	**TX16A**	Medium	2093	0.00	30
	High	2084	0.00	30		High	2091	282.28	565
	Highest	2093	1,010.38	2,093		Highest	2107	1,171.18	2,107
	Lowest	2115	0.00	30		Lowest	2153	0.00	30
	Low	2074	0.00	30		Low	2047	0.02	47
TX11B	Medium	2138	0.00	30	**TX17A**	Medium	2123	4.05	42
	High	2142	0.00	30		High	2074	62.61	254
	Highest	2026	3,865.26	2,026		Highest	2098	152.40	190
	Lowest	2100	57.17	308		Lowest	2140	0.00	30
	Low	2098	94.53	437		Low	2153	10.91	30
TX12A	Medium	2101	139.60	784	**TX18A**	Medium	2024	0.00	267
	High	2097	250.32	1,803		High	2099	3.31	42
	Highest	2099	372.12	1,145		Highest	2079	205.89	316
	Lowest	2100	22.97	192		Lowest	2127	0.00	30
	Low	2108	39.39	279		Low	2071	30.39	51
TX12B	Medium	2184	39.20	268	**TX19A**	Medium	2112	279.07	662
	High	2053	46.99	278		High	2116	330.84	893
	Highest	2050	69.79	331		Highest	2069	368.66	1,115
	Lowest	2118	110.26	46		Lowest	2102	0.00	30
	Low	2151	136.66	61		Low	2098	0.00	30
TX12C_10	Medium	2026	148.86	69	**TX19B**	Medium	2098	345.16	921
	High	2131	161.66	74		High	2102	522.80	1,830
	Highest	2069	178.16	218		Highest	2095	558.53	2,095

Table 4.17: Optimization output for the Five-Bin stratification method (2)

average of n_{hk} for the lowest half of the strata was 549 while the average for the highest half was 140. This is consistent with allocating a larger proportion of the sample to cheaper strata.

Finally, we sorted the strata on the basis of their N_{hk} values. The average n_{hk} of the lowest half was 457 while the average of the highest half was 232. While this may seem to contradict theory (e.g., taking a larger sample from larger strata), it turns out that the average cost of the lowest half of the strata, with respect to N_{hk}, was 3.71 while the average cost of the highest half of the strata, with respect to N_{hk}, was 6.28. As with the Two-Bin stratification method optimization output, this rule is likely being offset because the largest strata are so costly relative to the smallest strata. Nonetheless, based on these high-level summary statistics, we conclude that the optimization results generally conform with the theory for optimal allocation.

4.4.4.2 Simulation setup

To evaluate the performance of the Five-Bin stratification method, we carried out a simulation with $(M =)1,000$ iterations. For each iteration, we randomly "asked" sample units questions in the second interview based on the $f_{hk} = n_{hk}/N_{hk}$ values derived from the appropriate quantities in Tables 4.16 and 4.17. We then computed the quantities given in equations (3.7) – (3.14) to summarize the simulations and we also computed similar summary statistics for the domains defined by the CUs that incurred the expense. We then computed the average number of times the question was administered in the second interview, the average number of

questions asked, the average time spent answering a question, the average interview length, percent reduction in interview length, and the average and median design effects. Finally, we computed the average sensitivity, specificity, PPV, and NPV for this method.

4.4.4.3 Results

In Table 4.18, we display the simulation summary statistics for the Five-Bin stratification method. We observe the following trends for this method. The average number of times an expenditure question is asked coincides very closely with the allocation results presented in Tables 4.16 and 4.17. In fact, these averages range from 413 to about 4,905. In general, there are more expenditure categories being asked to fewer than 1,000 sample units when compared to the Two-Bin stratification method. We suspect that this is due to the fact that there are more strata for which the minimum number of sample units, 30, is being allocated to.

The simulation CVs for key estimates under this method range from 0.01 to 1.69 with the three least prevalent expenditure categories exhibiting the highest CVs. For this method, slightly more than half of the mean expenditure estimates have CVs of 0.10 or less. This suggests that under this method we may be able to obtain fairly precise estimates of desired quantities.

As with the other four methods, despite evidence of potentially biased estimates as indicated by the magnitude of the relative bias calculations, all the 95% confidence intervals associated with the relative bias calculations include zero.

Therefore, we cannot conclude that the bias is significantly different from zero. Finally, the design effects under this method range from 0.56 to 8.22. Only four expenditure categories have mean estimates with design effects that are strictly less than one. These are TX3F (mortgage/lump sum home equity loan), TX9D (sewing materials), TX10C_1 (car monthly payment), and TX12B (vehicle license fees) as their design effects are 0.56, 0.87, 0.56, and 0.91, respectively. This may indicate the potential for gains in precision over a split questionnaire design in which we essentially "flip a coin" to determine whether or not to ask a question. As with the Two-Bin method, incorporating constraints on the precision of key estimtes (e.g., through CV targets) and then solving the optimization problem, may help ameliorate this aspect of the method.

Expenditure	Asked	Mean	Variance	Std Err	Sim CV	Rel Bias	Rel Bias SE	RMSE	Bias LB	Bias UB	deff
TX2	3,960.65	634.24	1,721.78	41.49	0.07	0.96	6.60	41.93	-11.99	13.90	4.78
TX3F	2,454.32	216.89	248.82	15.77	0.07	0.41	7.30	15.80	-13.91	14.72	0.56
TX3H	2,086.25	104.34	382.24	19.55	0.19	0.20	18.78	19.55	-36.60	37.00	1.92
TX4A_1	686.57	351.02	102.11	10.10	0.03	0.27	2.89	10.15	-5.39	5.93	1.16
TX4A_2	538.02	27.72	8.45	2.91	0.10	0.60	10.55	2.91	-20.08	21.29	1.91
TX4A_3	485.54	23.89	9.35	3.06	0.13	0.19	12.83	3.06	-24.95	25.34	1.04
TX4A_4	413.66	0.11	0.03	0.18	1.69	-3.82	162.47	0.18	-322.26	314.63	1.44
TX4B	478.80	4.93	1.26	1.12	0.23	-0.68	22.64	1.12	-45.05	43.70	1.20
TX4C	1,152.89	61.81	8.15	2.86	0.05	0.17	4.63	2.86	-8.90	9.23	2.03
TX4D	2,078.06	608.29	78.89	8.88	0.01	0.01	1.46	8.88	-2.86	2.87	1.14
TX5A_1	997.15	6.35	39.28	6.27	0.99	0.04	98.70	6.27	-193.42	193.49	5.20
TX5A_2	422.92	1.39	2.02	1.42	1.02	-2.42	99.97	1.42	-198.37	193.53	1.82
TX5B	3,111.83	355.73	6,819.99	82.58	0.23	-0.74	23.04	82.63	-45.91	44.42	8.22
TX6A	2,219.41	38.49	27.84	5.28	0.14	-0.32	13.67	5.28	-27.10	26.46	1.39
TX6B	3,088.26	183.02	197.44	14.05	0.08	-0.36	7.65	14.07	-15.35	14.63	3.62
TX7A	514.25	16.37	11.78	3.43	0.21	0.23	21.02	3.43	-40.97	41.42	1.27
TX8	3,178.86	131.49	305.53	17.48	0.13	0.11	13.31	17.48	-25.97	26.20	5.90
TX9A	1,968.54	203.13	92.45	9.62	0.05	0.07	4.74	9.62	-9.21	9.36	2.25
TX9B	662.43	36.59	58.75	7.66	0.21	1.20	21.20	7.68	-40.35	42.74	2.73
TX9C	556.01	3.90	1.67	1.29	0.33	-0.08	33.12	1.29	-64.98	64.83	1.60
TX9D	468.62	3.90	1.17	1.08	0.28	0.07	27.78	1.08	-54.38	54.53	0.87
TX10C_1	910.05	20.23	7.14	2.67	0.13	0.58	13.28	2.67	-25.45	26.62	0.56
TX10C_23	2,213.11	50.24	153.48	12.39	0.25	0.63	24.82	12.39	-48.01	49.27	1.44
TX11B	2,146.36	694.12	4,313.54	65.68	0.09	0.11	9.47	65.68	-18.45	18.68	2.06
TX12A	4,475.04	165.65	18.93	4.35	0.03	-0.11	2.62	4.35	-5.25	5.04	1.14
TX12B	1,347.74	29.91	3.04	1.74	0.06	0.28	5.84	1.74	-11.17	11.73	0.91
TX12C_1TO9	1,036.92	24.42	5.32	2.31	0.09	-0.15	9.43	2.31	-18.63	18.33	1.03
TX12C_10	467.56	188.59	62.42	7.90	0.04	0.06	4.19	7.90	-8.16	8.27	1.12
TX13B	3,113.11	365.42	463.24	21.52	0.06	0.07	5.89	21.52	-11.48	11.62	6.79
TX14B	1,148.21	61.99	71.13	8.43	0.14	0.56	13.68	8.44	-26.25	27.37	2.38
TX15A	2,006.84	245.80	390.35	19.76	0.08	0.13	8.05	19.76	-15.65	15.90	3.72
TX16A	2,762.64	229.45	1,847.99	42.99	0.19	-0.20	18.70	42.99	-36.85	36.45	5.87
TX17A	561.38	51.60	44.95	6.70	0.13	-0.12	12.98	6.70	-25.56	25.32	1.19
TX18A	684.37	10.04	25.74	5.07	0.51	2.00	51.56	5.08	-99.06	103.06	1.84
TX19A	2,752.44	228.95	207.71	14.41	0.06	0.35	6.32	14.43	-12.03	12.73	3.27
TX19B	4,905.00	370.98	594.85	24.39	0.07	0.40	6.60	24.43	-12.54	13.33	7.73

Table 4.18: Simulation summary statistics for the Five-Bin stratification method

In Table 4.19, we present the simulation summary statistics for the domain characteristics where the domain is defined by those CUs incurring the expense. The average number of times the expenditure question is asked and the sample unit incurred it ranges from 1.15 to about 3,985. Similar to the Log2 and the Two-Bin stratification methods, we observe fewer than 100 instances of incurring the expenditure for eight of the expenditure categories. Even for the most prevalent expenditure categories, e.g., TX4A_1 (telephone services) and TX4D (utilities and fuels), we are observing far fewer instances of incurring the expense than we would expect if this method truly customized the set of expenditure questions a sample unit receives to their expense pattern.

All 95% confidence intervals associated with the relative bias calculations for the domain expenditure means include zero, therefore we cannot conclude that any are statistically different from zero. The simulation CVs for this method range from 0.01 to 0.83 with 20 expenditure categories exhibiting CVs less than or equal to 0.1. So, for more than half of our expenditure categories, we can obtain fairly precise estimates of the domain means. In addition, the Five-Bin method compared with the four other methods considered in this chapter produces the tightest range of CVs for the domain mean estimates. We interpret this as evidence that this type of method has the potential to produce the most precise estimates of desired characteristics of the methods considered.

Expenditure	Asked & Have	Mean	Variance	Std Err	Sim CV	Rel Bias	Bias SE	RMSE	Bias LB	Bias UB
TX2	2,791.48	2,242.56	206.26	14.36	0.01	0.03	0.64	14.38	-1.22	1.29
TX3F	450.19	4,617.08	30,025.31	173.28	0.04	0.00	3.75	173.28	-7.36	7.35
TX3H	517.26	1,863.33	60,448.57	245.86	0.13	0.42	13.25	245.99	-25.55	26.40
TX4A_1	650.67	379.49	90.00	9.49	0.02	0.29	2.51	9.55	-4.62	5.21
TX4A_2	202.37	102.84	13.98	3.74	0.04	-0.02	3.63	3.74	-7.14	7.10
TX4A_3	192.06	182.99	81.88	9.05	0.05	-0.04	4.94	9.05	-9.72	9.65
TX4A_4	1.15	49.26	1,656.08	40.69	0.83	1.04	83.48	40.70	-162.57	164.66
TX4B	92.37	59.74	84.63	9.20	0.15	-0.64	15.30	9.21	-30.63	29.34
TX4C	960.14	94.68	7.85	2.80	0.03	-0.02	2.96	2.80	-5.82	5.78
TX4D	1,970.99	660.83	67.65	8.23	0.01	0.02	1.24	8.23	-2.42	2.46
TX5A_1	17.00	338.32	64,733.25	254.43	0.75	-6.13	70.59	255.38	-144.49	132.24
TX5A_2	5.33	146.07	9,903.73	99.52	0.68	-3.67	65.63	99.67	-132.31	124.97
TX5B	868.32	1,964.68	165,750.35	407.12	0.21	-0.49	20.62	407.24	-40.91	39.92
TX6A	171.19	638.72	4,444.96	66.67	0.10	-0.06	10.43	66.67	-20.51	20.39
TX6B	1,829.45	416.78	718.37	26.80	0.06	-0.35	6.41	26.84	-12.91	12.21
TX7A	80.04	167.28	596.65	24.43	0.15	0.22	14.63	24.43	-28.46	28.91
TX8	1,693.29	354.72	1,907.59	43.68	0.12	0.03	12.32	43.68	-24.11	24.17
TX9A	1,514.12	303.31	134.63	11.60	0.04	-0.02	3.82	11.60	-7.52	7.48
TX9B	287.26	160.33	835.14	28.90	0.18	0.68	18.15	28.92	-34.89	36.25
TX9C	44.25	56.97	239.94	15.49	0.27	-0.70	27.00	15.50	-53.62	52.22
TX9D	46.29	61.14	198.47	14.09	0.23	0.33	23.12	14.09	-44.98	45.64
TX10C_1	155.47	506.30	931.50	30.52	0.06	0.28	6.05	30.55	-11.57	12.13
TX10C_23	168.13	2,951.11	141,256.38	375.84	0.13	-0.87	12.62	376.73	-25.62	23.87
TX11B	1,788.05	3,902.21	29,445.88	171.60	0.04	-0.04	4.40	171.60	-8.65	8.58
TX12A	2,717.44	303.04	52.64	7.26	0.02	-0.04	2.39	7.26	-4.73	4.65
TX12B	390.45	107.20	23.04	4.80	0.04	0.27	4.49	4.81	-8.53	9.07
TX12C_1TO9	318.82	89.50	51.08	7.15	0.08	-0.55	7.94	7.16	-16.12	15.02
TX12C_10	437.77	211.28	62.93	7.93	0.04	0.03	3.76	7.93	-7.33	7.39
TX13B	2,553.19	563.03	582.38	24.13	0.04	0.12	4.29	24.14	-8.29	8.53
TX14B	807.16	230.53	688.02	26.23	0.11	0.27	11.41	26.24	-22.09	22.63
TX15A	1,575.71	399.49	735.19	27.11	0.07	0.21	6.80	27.13	-13.12	13.55
TX16A	1,524.64	991.24	25,895.12	160.92	0.16	-0.37	16.17	160.96	-32.07	31.33
TX17A	258.21	154.35	277.50	16.66	0.11	-0.31	10.76	16.67	-21.40	20.78
TX18A	25.01	336.57	18,046.54	134.34	0.40	3.00	41.11	134.69	-77.58	83.57
TX19A	2,354.53	338.65	329.97	18.17	0.05	0.20	5.37	18.18	-10.33	10.74
TX19B	3,985.59	607.29	947.07	30.77	0.05	0.22	5.08	30.80	-9.73	10.18

Table 4.19: Simulation summary statistics for the domains using the Five-Bin stratification method

In Table 4.20, we present the calculations for the four epidemiological criteria for the Five-Bin stratification method. We conclude that given a sample unit incurred the expense, this method does a poor job of asking CUs about expenses that they are likely to have incurred. Only seven of the expenditure categories had sensitivity values greater than 0.5. In terms of specificity, the Five-Bin stratification method does a good job of correctly not asking sample units about expenses that they are not likely to have incurred. All 36 expenditure categories had specificity values exceeding 0.5.

The third evaluation criteria we report on is PPV. Of the 36 expenditure categories we investigated, 27 exhibited PPV values of 20% or greater than the associated prevalence. These are identified by the shaded cells in the PPV column of Table 4.20. This suggests that for these expenses we are detecting at least 20% more instances of incurring the expense than we would by a completely random split questionnaire design. For only one expenditure category, TX5A_1 (construction materials for specific jobs not yet started), we do slightly worse than the completely random split questionnaire design.

The final criteria we report on is NPV. Under no instances do we perform worse than the completely random split questionnaire design and for five of the 36 expenditure categories, we achieve NPV values of at least 20% or greater than one minus the associated prevalence rate. The expenditures corresponding to this situation are identified by the shaded cells in the NPV column of Table 4.20.

Expenditure	Sensitivity	Specificity	PPV	NPV	$P(Int2)$
TX2	0.949	0.845	0.705	0.977	0.280
TX3F	0.917	0.800	0.183	0.995	0.047
TX3H	0.878	0.842	0.248	0.991	0.056
TX4A_1	0.067	0.954	0.948	0.076	0.925
TX4A_2	0.072	0.956	0.376	0.738	0.268
TX4A_3	0.140	0.968	0.396	0.883	0.130
TX4A_4	0.048	0.961	0.003	0.998	0.002
TX4B	0.107	0.960	0.193	0.923	0.083
TX4C	0.140	0.947	0.833	0.371	0.652
TX4D	0.204	0.871	0.948	0.086	0.921
TX5A_1	0.092	0.905	0.017	0.982	0.018
TX5A_2	0.054	0.960	0.013	0.991	0.009
TX5B	0.456	0.739	0.279	0.860	0.182
TX6A	0.270	0.792	0.077	0.944	0.060
TX6B	0.397	0.786	0.592	0.625	0.439
TX7A	0.078	0.954	0.156	0.905	0.098
TX8	0.436	0.775	0.533	0.700	0.370
TX9A	0.216	0.869	0.769	0.354	0.669
TX9B	0.121	0.954	0.434	0.787	0.227
TX9C	0.062	0.948	0.080	0.933	0.068
TX9D	0.069	0.957	0.099	0.938	0.064
TX10C_1	0.372	0.925	0.171	0.973	0.040
TX10C_23	0.955	0.802	0.076	0.999	0.017
TX11B	0.959	0.958	0.833	0.991	0.178
TX12A	0.473	0.630	0.607	0.498	0.547
TX12B	0.133	0.873	0.290	0.723	0.279
TX12C_1TO9	0.112	0.906	0.307	0.732	0.272
TX12C_10	0.047	0.974	0.936	0.110	0.892
TX13B	0.375	0.848	0.820	0.423	0.649
TX14B	0.287	0.956	0.703	0.785	0.268
TX15A	0.244	0.893	0.785	0.424	0.616
TX16A	0.629	0.847	0.552	0.884	0.231
TX17A	0.074	0.957	0.460	0.673	0.334
TX18A	0.079	0.935	0.037	0.970	0.030
TX19A	0.332	0.883	0.855	0.389	0.675
TX19B	0.623	0.775	0.813	0.568	0.610

Table 4.20: Epidemiological criteria for the Five-Bin stratification method

4.5 Comparison of methods

There are essentially three criteria on which we can compare the methods developed in this chapter. They are (1) potential for burden reduction; (2) loss of information due reduced sample size receiving each question; and, (3) success in terms of tailoring the survey to the individual respondent. In addition, since the two primary gaps our research addresses with respect to split questionnaires are that (1) prior information on the sample unit is often ignored in the design and (2) the existing set of split questionnaire methods are ineffective for surveys collecting data on rare events, we classified the expenditures into categories on whether or not each was rare and/or recurrent. We use the classification system presented in Table 3.8 of Section 3.3.1.1. We utilize this classification system to summarize some of the results across the methods and to relate any trends to the research we address.

In Table 4.21, we present high-level summary of metrics associated with burden reduction and loss of information for each method. We also provide reference values for the full analysis file and the "one-half" condition. For comparisons along the dimension of potential burden reduction, we use the metrics displayed in the second through fifth columns of Table 4.21. We observe that the average number of questions asked to a respondent somewhat varies among the methods, e.g., from about 0.7 to 11. Under the PPS method, we ask, on average, fewer than one question to each respondent while for the Log1 and Two-Bin methods, we ask roughly the same number of questions. Under the Log2 method, we ask the most questions, on average, at slightly over 11.

Method	Asked	Length	Minutes per question	Overall % Reduction	Design effect Average	Design effect Median
PPS	0.73	1.13	0.66	97.12	2.46	1.98
Log1	9.43	12.18	1.27	68.96	1.32	0.99
Log2	11.18	15.29	1.32	61.03	1.45	0.98
Two-Bin	9.00	10.77	1.16	72.55	7.33	1.77
Five-Bin	5.91	8.72	1.41	77.78	2.59	1.83
One-Half	18.00	19.62	1.09	50.00	1.04	1.00
Full file	36.00	39.24	1.09	...	...	...

Table 4.21: Responsive split questionnaire methods simulation comparison

We speculate that the PPS method results in the smallest number of questions asked because it only (and often) asks about the respondent's largest expense (because of the manner in which the decision rules were developed). For many respondents, this would likely be housing expenditures (e.g., rent, mortgage payments). This may also be related to why the minutes per question is the lowest for this method. Housing expenditures are expenses that are incurred regularly, generally do not change from month-to-month, and as a consequence can be easily recalled. The easy recollection of these expenses likely contributes to "quick" reporting. Thus, we observe the lowest minutes per question under this method.

For the Consumer Expenditure Survey Program, it might be helpful to understand how implementation of one of the methods developed in this chapter might translate into "overall burden" reduction. Taking burden solely as the length of interview, we compute the percent reduction in interview length under each method. We present this in the fifth column of Table 4.21. As expected, the PPS method results in the greatest percentage reduction in interview length. This is only because

under this method, we ask, on average, fewer than one question to each respondent. We deem this method (as proposed in this chapter) not practical for implementation in a redesigned Consumer Expenditure Survey simply because we ask so few questions. Furthermore, it is not practical from a data collection cost perspective as it is cost-prohibitive to locate, contact, and survey respondents to collect information on one question. It is more promising, however, that the Log1, Log2, and Two-Bin methods yield percent reductions in interview length of about 60 to 70 percent. We believe these methods to be feasible for implementation given that we are still asking a reasonable amount of questions to each respondent. In addition, the average lengths of the interviews under these three methods are consistent with the recommended length for telephone surveys. Therefore, it might be feasible to consider these methods for telephone survey administration (in addition to personal visit).

In Figure 4.1, we expand on our comparison of the methods along the dimension of potential burden reduction. While the average minutes per question are provided in Table 4.21 for each method, we present box plots, based on the simulation data, associated with these quantities in Figure 4.1. It is important to note that the range of values under each method differs in their respective box plot, but the length of the range is constant across the box plots. What this graphical depiction highlights is that while the average minutes per question is the lowest with the PPS method, the associated range is the largest. This may be an indication that under this method, there is a greater amount of variability in the potential for burden reduction than the other methods. Furthermore, the Log2 method appears to have

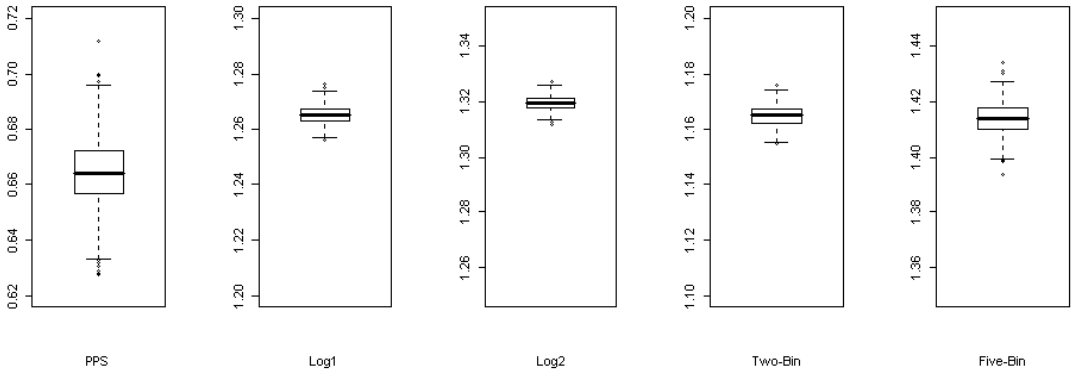

Figure 4.1: Minutes per sample unit per expenditure for the responsive split questionnaire methods

the tightest range of minutes per question. This suggests that under this design we can consistently rely on this method to produce fairly constant reductions in burden across implementations.

To assess the potential losses in precision across the methods, we can compare the average and median design effects (see Table 4.21). Since some of the methods yield large design effects for one or more of the mean expenditure estimates, the median might be a more appropriate measure to compare the methods. Both the Log1 and Log2, in general, do not appear to have adverse effects on the losses in precision (relative to a completely random split questionnaire design of the same subsample size) as the median design effects under these methods are about one. The other three methods, tend to produce larger design effects, on average. This may be an indication that these methods might result in greater losses of information if implemented.

Using the "rare by recurrent" classification system, we display the design effects for the mean expenditure estimates for the methods developed in Chapter 4 as

well as the "four-fifths" condition from Preliminary Study 2[18]. Figure 4.2 displays the full set of design effects while Figure 4.3 restricts the focus to design effects of two or less as the full set inhibits our ability to identify any prominent trends across the methods. For these graphs in addition to using the rare by recurrent classification, we sorted the expenditures in decreasing order with respect to their second interview prevalence within each classification.

In Figure 4.2, the key observations are as follows. In general, for the not rare expenditure categories, regardless of recurrence, the PPS and Two- and Five-Bin methods tend to result in the largest design effects. For the not recurrent, but rare expenditure categories, the Two-Bin method generally results in the largest design effects. Finally, for the rare and recurrent expenditure categories, the design effects range from less than one to about four with the Two-Bin method resulting in the largest design effects.

When looking at the restricted range of design effects presented in Figure 4.3, it appears that regardless of the rare by recurrence classification, the Log2 method tends to produce design effects that are around one. Perhaps the one exception might be for TX12C_1TO9 (other vehicle fuels). This is a very encouraging finding since this method also tends to outperform the other methods in terms of tailoring the survey to the individual respondent. We verify this claim in Figures 4.4–4.7. For the rare, but recurrent expenditure categories, the PPS method seems to result in the smallest design effects, but given that there are only four expenditure categories

[18]We chose not to display the design effects from the "one-half" condition since they were all around 1.

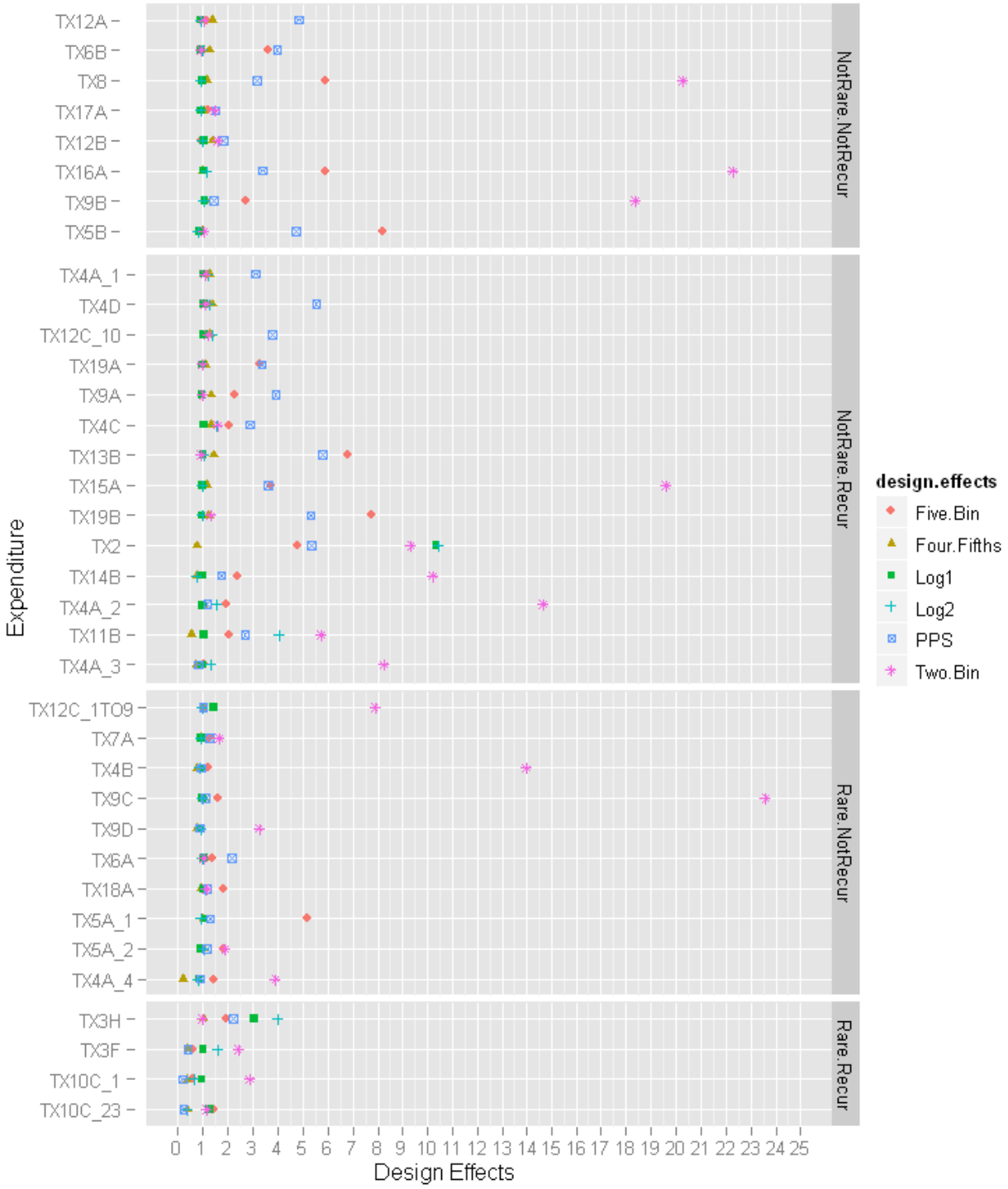

Figure 4.2: Full set of design effects for Chapter 4 methods

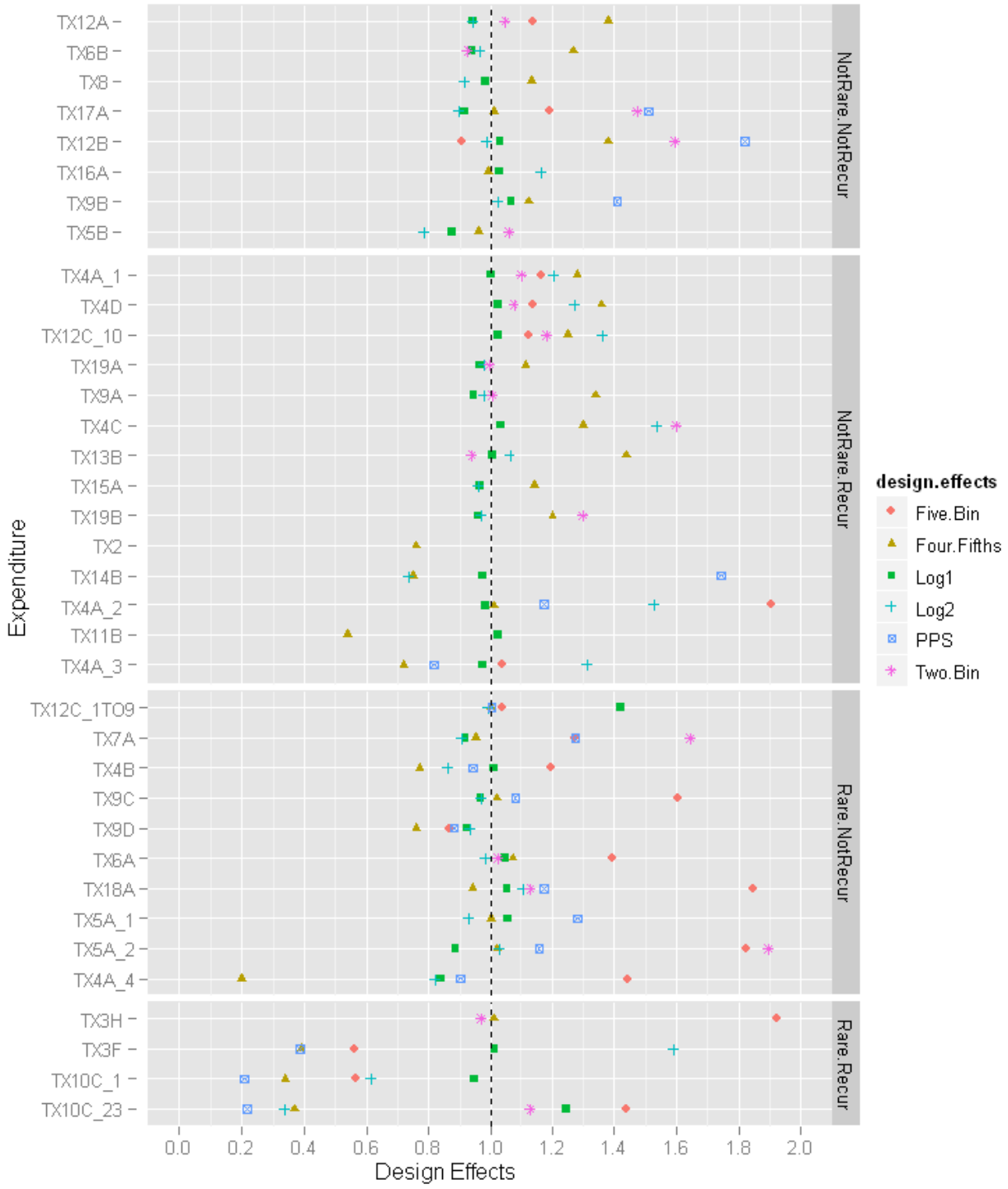

Figure 4.3: Restricted range of design effects for Chapter 4 methods

falling into this classification, the generalizability of this conclusion may be tenuous. Consistent with Figure 4.2, the Five-Bin method, regardless of the rare by recurrence classification, tends to produce the largest design effects among the methods considered. Perhaps by specifying certain precision targets on the key estimates of interest, we might be able to alleviate that to some degree.

The final way in which we compare the methods in terms of losses in information, is via a comparison of the root variance ratios (see Table 4.22). The comparison is to a completely random split questionnaire design. The simulation variances under each method are expressed as functions of the same sample size using the technique detailed in Section 2.6.1, in particular, equation (2.30). The same sample size we use is one-half the number of units in the analysis file. So, the quantities displayed in Table 4.22 are the square roots of the ratios of $V_{SQ,n^*,\text{alt}}(\cdot|S)$ to $V_{SQ,n^*,\text{one-half}}(\cdot|S)$ where alt is either PPS, Log1, Log2, Two-Bin, or Five-Bin.

In Table 4.22, the darker shaded cells correspond to the method that result in the smallest inflations in variance relative to a completely random split questionnaire design. The lighter shadings correspond to the methods that outperformed the completely random split questionnaire design (as indicated by a value of less than 1), but was not the superior method. The primary conclusion drawn from Table 4.22 is that while the root variance ratios vary across methods and expenditure categories, the Log2 method yields more instances of being superior when compared to the other Chapter 4 methods as well as the completely random split questionnaire design.

Expenditure	PPS	Log1	Log2	Two-Bin	Five-Bin
TX2	3.18	4.41	4.43	4.15	2.92
TX3F	0.88	1.44	1.81	2.06	0.84
TX3H	1.03	1.20	1.38	1.29	1.84
TX4A_1	2.43	0.41	0.74	1.40	1.46
TX4A_2	1.57	1.21	1.64	5.52	1.98
TX4A_3	1.30	1.32	1.57	4.06	1.44
TX4A_4	1.27	1.22	1.21	2.62	1.59
TX4B	1.37	1.37	1.25	5.22	1.52
TX4C	2.35	0.88	1.31	1.71	1.93
TX4D	3.22	0.45	0.82	1.35	1.34
TX5A_1	1.58	1.42	1.34	10.24	3.15
TX5A_2	1.45	1.25	1.36	1.81	1.80
TX5B	3.05	1.22	1.10	0.97	3.97
TX6A	2.18	1.50	1.44	1.41	1.63
TX6B	2.94	1.16	1.07	1.28	2.70
TX7A	1.58	1.29	1.26	1.75	1.55
TX8	2.42	1.15	1.01	6.11	3.24
TX9A	2.77	0.89	0.78	1.31	2.02
TX9B	1.69	1.35	1.28	6.07	2.33
TX9C	1.43	1.32	1.31	6.65	1.72
TX9D	1.36	1.36	1.36	2.59	1.32
TX10C_1	0.60	1.28	1.02	2.14	0.93
TX10C_23	0.61	1.50	0.76	1.12	1.50
TX11B	2.28	1.31	2.80	3.24	1.93
TX12A	3.03	1.06	0.87	1.31	1.17
TX12B	1.88	1.31	1.18	1.72	1.24
TX12C_1TO9	1.43	1.70	1.21	4.00	1.38
TX12C_10	2.74	0.52	0.98	1.47	1.49
TX13B	3.44	0.99	0.92	1.24	3.66
TX14B	1.90	1.28	0.99	4.55	2.18
TX15A	2.58	0.95	0.80	5.99	2.57
TX16A	2.66	1.34	1.41	6.78	3.44
TX17A	1.75	1.16	1.07	1.67	1.52
TX18A	1.54	1.45	1.48	1.44	1.90
TX19A	2.55	0.85	0.78	1.28	2.44
TX19B	3.15	0.93	0.82	1.29	3.70

Table 4.22: Comparison of root variance ratios of the responsive split questionnaire methods

Using the graphical displays presented in Section 2.6.2, we compare the five methods developed in Chapter 4 on the basis of the epidemiological criteria (see Figures 4.4 – 4.7) to assess their ability to tailor the survey to individual respondent. In these figures, we include references for the "one-half" and "four-fifths" conditions from Preliminary Study 2. We chose the "four-fifths" condition as the reference test for these graphs since this condition delineated the regions of superiority better than the "one-half" condition. With the "one-half" condition, we would essentially plot one dotted line. If the Chapter 4 method fell above the dotted line, then we would deem it overall superior to the "flipping the coin" method. If it fell below the dotted line, then we would deem it overall inferior.

While this discretization is easily interpretable, we do not have the nice characterization of a method being overall superior, superior for the presence of the event, superior for the absence of the event, or overall inferior. Thus, using the "four-fifths" condition as the reference test afforded us the capability to have the more complete characterization. Furthermore, in order to condense the number of graphs, we plotted multiple methods on each graph. Even though the reference test for each individual graph is the "four-fifths" condition, one can envision a corresponding set of dotted lines for any of the methods. This would allow them to make comparisons between each pair of methods.

In Figure 4.4, we offer comparisons of the responsive split questionnaire methods for the **not rare, but recurrent** expenditure categories. In general, it appears that for these expenditures, the Log2 method tends to be superior (either overall superior, superior for the presence, or superior for the absence) than the "four-fifths"

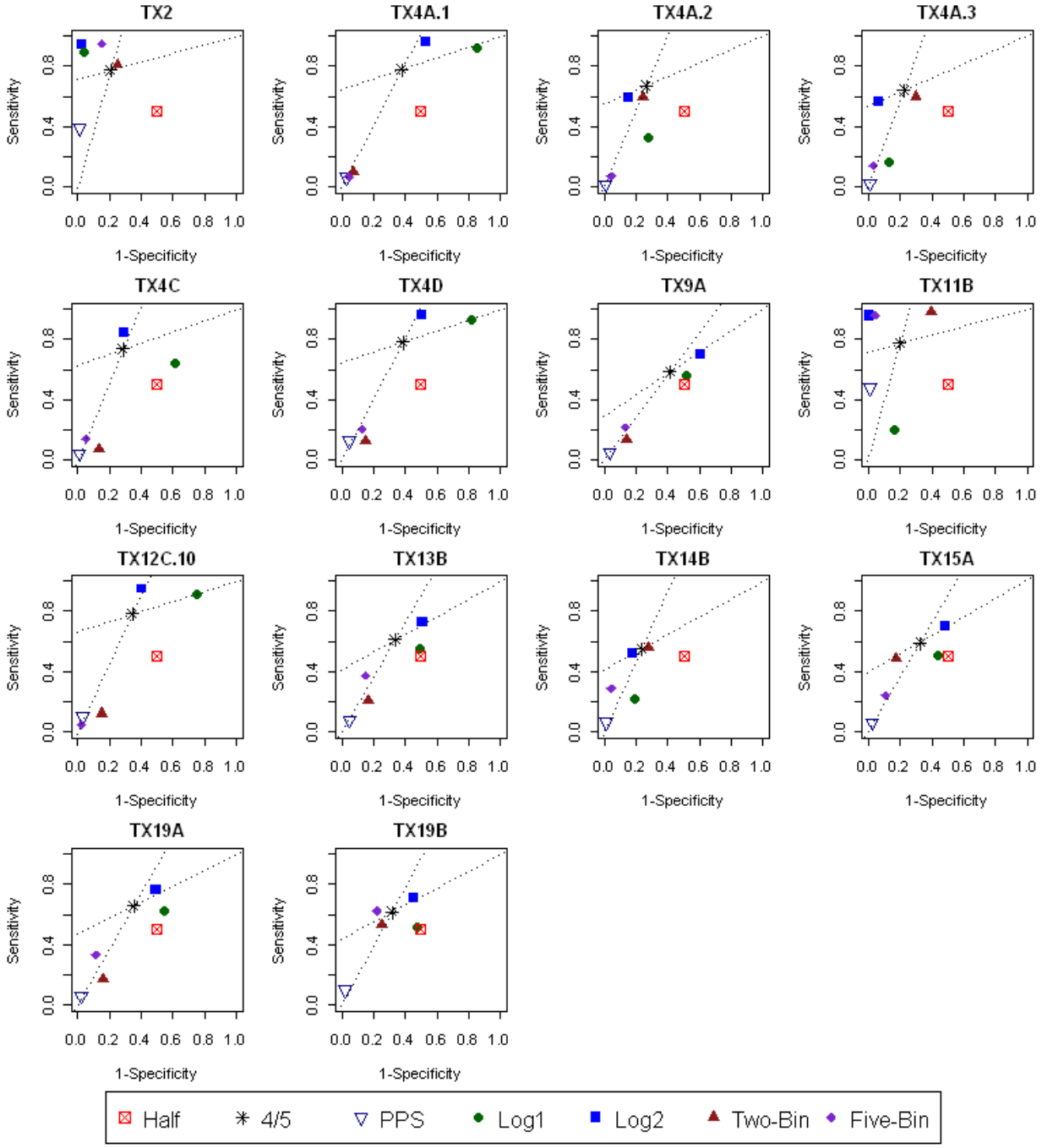

Figure 4.4: Diagnostic test comparisons for Chapter 4 methods (not rare, recurrent)

condition. Additionally, the Log2 method appears to outperform the other methods developed in Chapter 4. For some expenditure categories, there is evidence that the Two- and Five-Bin stratification methods are successful in terms of tailoring the survey questions to the respondent's purchase behavior. We would also argue that for the not rare, but recurrent expenditure categories, with the exception of, perhaps, the PPS method, all methods developed in this chapter are superior to the completely random split questionnaire design with respect to tailoring the survey to the individual respondent.

In Figure 4.5, we offer comparisons of the responsive split questionnaire methods for the **not rare, not recurrent** expenditure categories. For these expenditure categories, there is evidence that the Log1, Log2, and Five-Bin methods might outperform the "four-fifths" condition. Although there are no consistent trends, at least one of the methods is superior (in some way) to the "flipping the coin" method for each expenditure category. For instance, it appears that the Log2 method is superior for detecting the presence of the event relative to the "four-fifths" condition for TX5B (contractor labor, materials, and tools), TX16A (educational expenses), and TX17A (cost of subscriptions). On the other hand, it seems that the Two-Bin method is the preferred method for detecting the absence of the event for TX5B (contractor labor, materials, and tools) and TX16A (educational expenses).

For the **rare and recurrent** expenditure categories displayed in Figure 4.6, the Log2 method is overall superior to the "four-fifths" condition for all expenditures in this classification. There is also evidence that the PPS and Five-Bin methods are superior, either overall or for the presence of incurring the expense, to the "four-

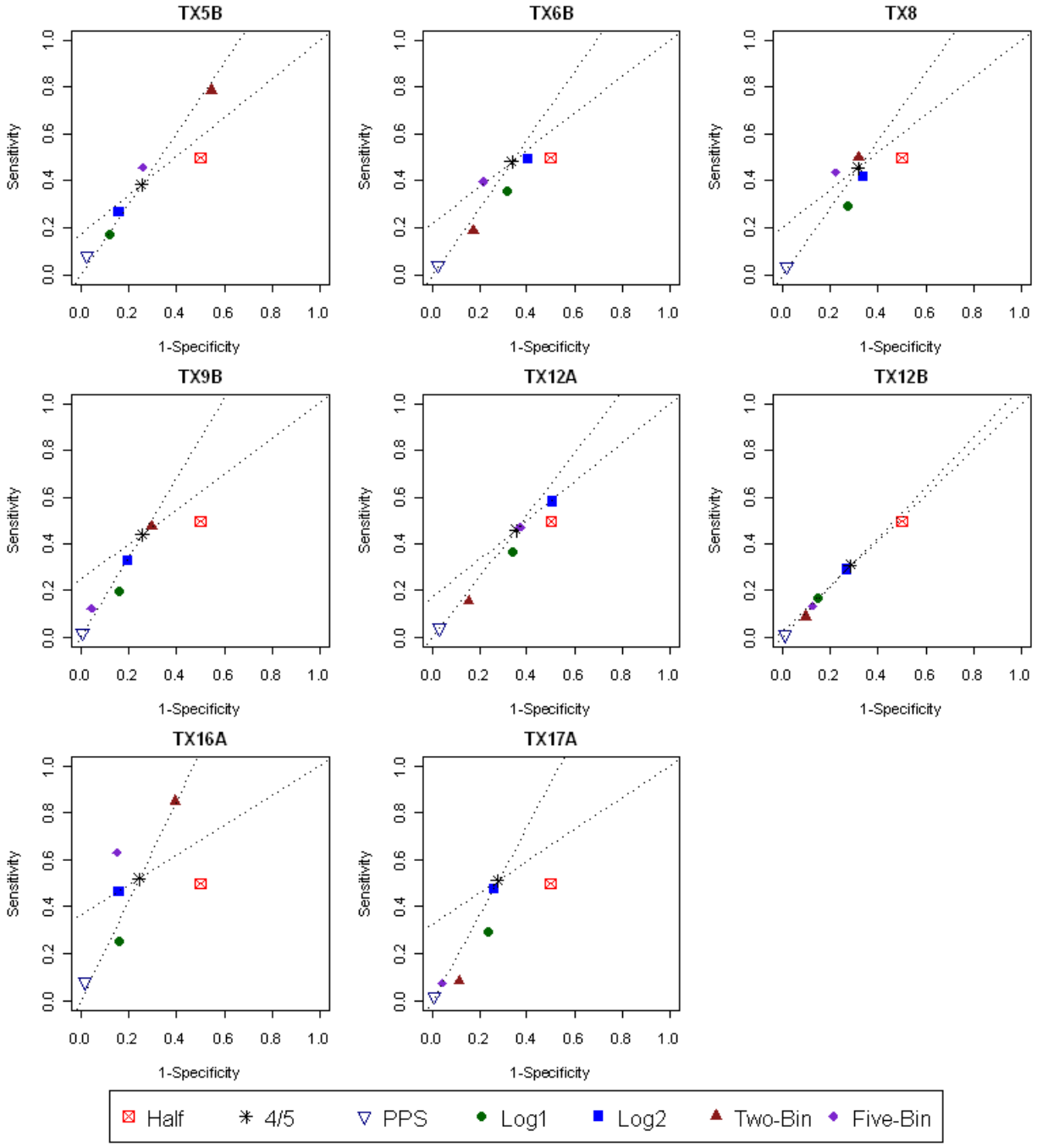

Figure 4.5: Diagnostic test comparisons for Chapter 4 methods (not rare, not recurrent)

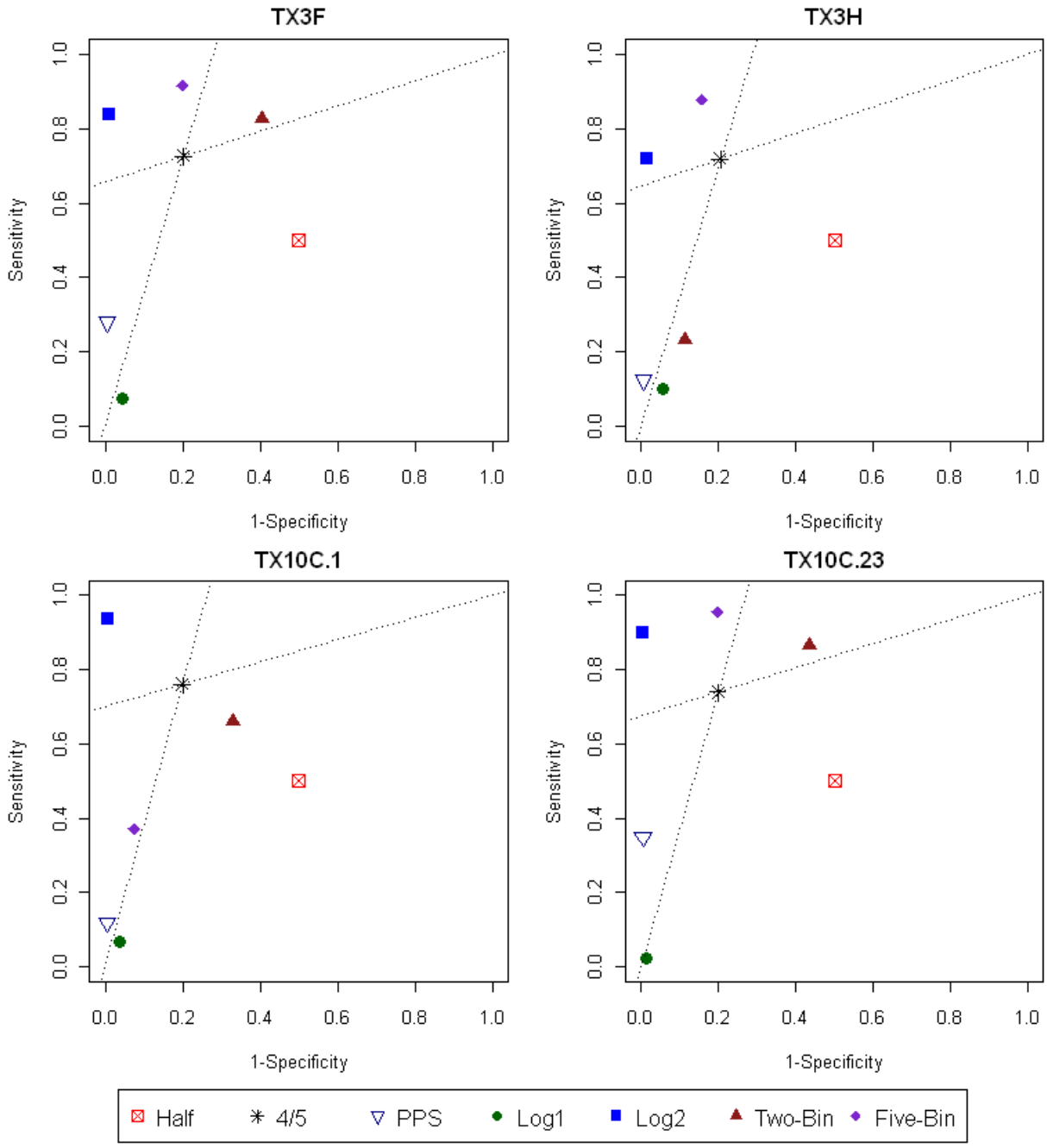

Figure 4.6: Diagnostic test comparisons for Chapter 4 methods (rare, recurrent)

fifths" condition.

For the last classification, **rare, but not recurrent**, displayed in Figure 4.7, there is little evidence of any of the methods being superior to the "four-fifths" condition. There is also a lack of evidence of any of these methods being superior to the completely random split questionnaire design. Considering that three of the five methods proposed in Chapter 4 use an indicator for whether the unit incurred the expense in the first interview reference period, it is not surprising that these methods produce little evidence of being superior (in any way). If anything, there is weak evidence that the Log2 method is superior for the "presence of the event" for TX4B (telephone cards, prepaid cell, public pay phone), TX7A (household item repair and service contracts), TX9C (clothing services), TX9D (sewing materials), and TX12C_1TO9 (other vehicle fuels). For these types of expenditures, in order to improve the responsive split questionnaire methods, it might be useful to include more relevant predictors and/or specialized subsets of predictors into the models predicting whether the unit incurred the expense in the second interview reference period.

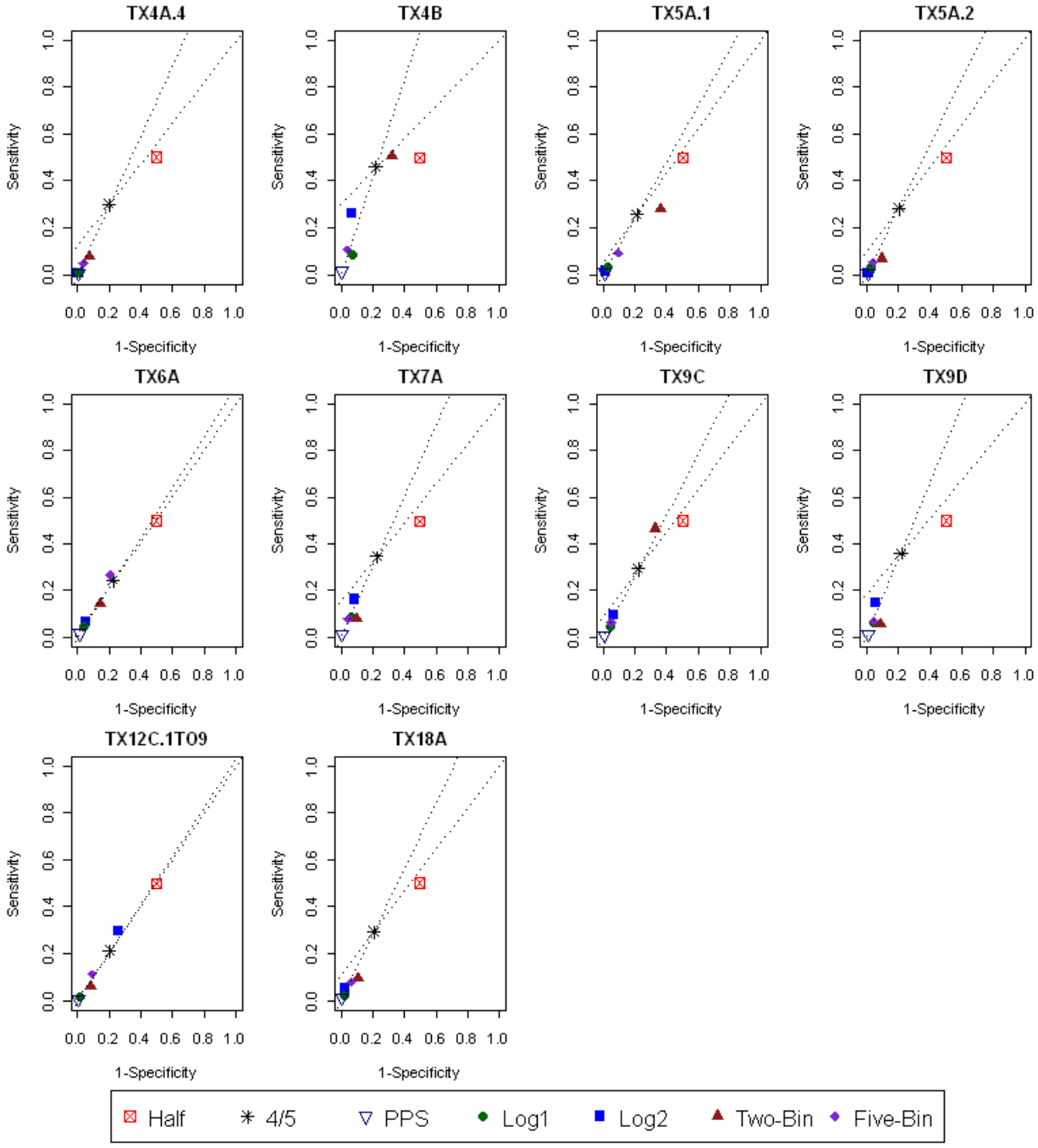

Figure 4.7: Diagnostic test comparisons for Chapter 4 methods (rare, not recurrent)

Chapter 5

Discussion

5.1 Overview

In this chapter, we conclude with a discussion of some general conclusions and lessons learned from our research. We offer general guidelines for survey programs wishing to utilize the methods (or similar methods) developed in this research. Finally, we highlight some limitations of this research by identifying areas for future investigation.

5.2 General conclusions

In this dissertation, we proposed various methods for incorporating prior information about the sample unit into the decision rules for a responsive split questionnaire. The implementation of a responsive split questionnaire that we considered was for the first interview to remain as is, but the second was subject to the split questionnaire design. The five methods we proposed were as follows: (1) the **PPS** method which represented a classical approach to sampling, e.g., probability proportional-to-size, where the measure-of-size was the sample unit's expenditure amount for a particular item reported in the first interview; (2) the **Log1** method was a logistic regression-based method where only first interview information was

available and an estimated propensity for purchasing the item in the first interview was used as a proxy for the propensity of purchasing the item in the second interview; (3) the **Log2** method was also a logistic regression-based method, but in addition to first interview information, "auxiliary data" were available to model the relationship of incurring the expense across two successive reference periods, and then an estimated, non-proxy, propensity for purchasing the item in the second interview was obtained; (4) the **Two-Bin** method employed stratification techniques for which information collected in the first interview was used to stratify the sample into two strata based on the sample unit's estimated propensity of incurring the expense in the second interview reference period. These strata represented a "Low" and "High" likelihood of incurring the expense in the second interview. Stratum standard deviations were estimated directly from expenditure information collected in the first interview. Mathematical programming methods were then used to determine the sampling fractions, i.e., decision rules, within each stratum subject to various design constraints; and, (5) the **Five-Bin** method extended the Two-Bin method to reflect a continuum of purchase behavior by stratifying the sample into five strata per expenditure, determined by the quintiles for the estimated propensities of incurring the expense in the second interview reference period.

To judge the overall effectiveness of each method, we evaluated each on the basis of two key dimensions: (1) the statistical properties of the design (e.g., precision of key estimates computed from the collected data), and (2) the successfulness of its ability to tailor the survey to the individual respondent (i.e., responsiveness). In Tables 5.1 and 5.2, we present a high-level qualitative summary of key findings

from our research related to these two dimensions. We display the recurrent expenditures in Table 5.1 and the not recurrent expenditures in Table 5.2. We also identify which expenditures were classified as rare using the classification system presented in Section 3.3.1.1. Using these tables, we can characterize the potential tradeoffs between satisfying precision requirements and being successful in terms of a responsive design. Furthermore, we can relate these characterizations to the general optimality framework outlined in Section 2.5.3.

In each table, for the **precision** dimension, a '+' was assigned to each expenditure category, for each method if (1) the simulation CV for the unconditional mean expenditure estimate was less than or equal to 0.1; (2) the design effect for the unconditional mean expenditure estimate was less than 1; (3) the simulation CV for the domain mean estimate was less than or equal to 0.1; or, (4) the root variance given in Table 4.22 was less than 1.

The criteria in the preceding paragraph are all related to the precision of key estimates and can be interpreted as follows. First, a simulation CV for the unconditional mean expenditure estimate of less than or equal to 0.1 implies that the precision requirement of various stakeholders will likely be satisfied. Second, a design effect for the unconditional mean expenditure estimate of less than 1 suggests a gain in precision of the responsive split questionnaire over a standard, completely random, split questionnaire design. Third, a simulation CV for the domain expenditure mean estimate (where the domain is defined by those units incurring the expense) of less than or equal to 0.1 implies that the precision requirement of various stakeholders will likely be satisfied. Finally, the root variance of less than 1 implies a gain

in precision of the responsive split questionnaire method relative to a completely random split questionnaire design after controlling for the sample size.

With respect to the **responsiveness** dimension, a '+' was assigned to each expenditure category, for each method if (1) the sensitivity was greater than 0.5; (2) the specificity was greater than 0.5; (3) the PPV was at least 20% greater than the prevalence; or (4) the NPV was at least 20% greater than one minus the prevalence.

The four criteria in the preceding paragraph are related to the epidemiological criteria that we used to assess the efficacy of the responsive design features of the split questionnaire and can be interpreted as follows. First, a sensitivity value greater than 0.5 implies that the responsive split questionnaire method is better than a completely random question asking procedure at asking about the expense of sample units who incurred the expense. Second, a specificity value greater than 0.5 suggests that the responsive split questionnaire method is better than a completely random question asking procedure at not asking about the expense of sample units who did not incur the expense. Third, a PPV of at least 20% greater than the prevalence implies that the responsive split questionnaire method will detect 20% more instances of incurring the expense than a completely random question asking procedure. Finally, a NPV of at least 20% greater than one minus the prevalence suggests that the responsive split questionnaire method will detect 20% more instances of not incurring the expense than a completely split questionnaire design. As a final note, for each dimension, there is a maximum possible number of four '+' that could be assigned, with the greater number of '+' being an indication that the method performs well for that expenditure with respect to the particular dimension.

Expenditure	Rare	Dimension	PPS	Log1	Log2	Two-Bin	Five-Bin
TX2	No	Precision	+	+	+	++	++
TX4A_1			++	+++	+++	++	++
TX4A_2			+	+++	++	+	+
TX4A_3			+	+++	++	+	+
TX4C			+	+++	++	++	++
TX4D			++	+++	+++	++	++
TX9A				++++	++++	++	++
TX11B			+	++	+	+	++
TX12C_10			++	+++	+++	++	++
TX13B			+	+++	+++	+++	++
TX14B				+++	++++		
TX15A				++++	++++		++
TX19A				++++	++++	++	++
TX19B				++++	++++	++	++
TX2	No	Responsiveness	++	++++	++++	++++	++++
TX4A_1			+	++	++	+	+
TX4A_2			++	+	+++	+++	++
TX4A_3			++	++	+++	+++	++
TX4C			++	+	++++	+	++
TX4D			+	++	++	+	+
TX9A			+	+	++	+	+
TX11B			++	+	++++	++++	++++
TX12C_10			+	+	+++	+	+
TX13B			+	++	++	+	+++
TX14B			++	+	+++	+++	++
TX15A			++	++	+++	+++	++
TX19A			++	+	+++	+	+++
TX19B			++	++	+++	++++	++++
TX3F	Yes	Precision	++			+	++++
TX3H						+	
TX10C_1			++	+	++	+	+++
TX10C_23			++		++	+	
TX3F	Yes	Responsiveness	++	++	+++	+++	+++
TX3H			++	++	+++	++	+++
TX10C_1			++	++	+++	+++	++
TX10C_23			++	++	+++	+++	+++

Table 5.1: Summary of research (**recurrent**)

The general trends observed from these two tables can be described as follows. For the **recurrent but not rare** expenditures, the Log2 method appears to perform the best, overall, with respect to both dimensions. Furthermore, for these types of expenditures (recurrent, but not rare), the tradeoffs between satisfying precision requirements and being successful in terms of tailoring the survey to the individual respondent are less stark than for the other types of expenditures. This may suggest that incorporating prior information into the design of a split questionnaire can enhance the split questionnaire without compromising other features of the survey design. On the other hand, for the **recurrent, but rare** expenditures, there is an indication that each method might be successful with respect to the responsiveness dimension, but deficient in terms of meeting the necessary precision requirements of key estimates. We still deem this an encouraging finding because it suggests that the **recurrent** expenditures are amenable to tailoring, regardless of the rarity of the expense, with the precision improving as the prevalence of the expense increases.

The tradeoffs between the two evaluative dimensions appear more evident for the **not recurrent** expenditures (see Table 5.2). In general, for the rare expenditure categories, each method shows promise with respect to the responsiveness dimension, but is inadequate for meeting certain precision requirements. As with the recurrent but rare expenditures, evidence that the methods are successful for designing a responsive split questionnaire for **not recurrent, but rare** expenditures is an encouraging finding. This suggests that implementation of a responsive split questionnaire has the potential to alleviate the concern of missing rare events (e.g., purchases). This represents the potential for a significant improvement over a

standard split questionnaire design since standard designs (of the same subsample size) would likely still fail to meet certain precision requirements.

There are four additional key findings that one can glean from this research. First, we demonstrated that incorporating prior information about the sample unit in the design of a split questionnaire can yield significant gains over standard split questionnaire designs. Standard designs may not meet certain precision requirements due to the reduced sample size receiving each question; thus, by incorporating prior information into the split questionnaire, we have improved another aspect of the design – its ability to customize the survey to the individual respondent. As a consequence, we can potentially rely on the methods to "screen out" sample units (i.e., not ask certain questions), with a high degree of accuracy, to save interview time and improve other aspects of the response process[1].

Second, there is evidence that the appropriate method for asking the question may depend on a variety of factors. This may include, but is not limited to the information available, characteristics of the underlying construct, and/or precision requirements on estimates derived from the collected data. In survey methodology, it is perhaps naive to believe that one method will outperform another with respect to every evaluation criteria. In other words, there will likely be tradeoffs. In our research, we demonstrated that sometimes a method might be successful in terms

[1]In this dissertation, we made no attempt to prove that other aspects of the response process will be improved by implementing a responsive split questionnaire. This is only a conjecture. We have, however, presented evidence that standard split questionnaire designs can improve the nonresponse and measurement properties of the survey (Creech et al., 2011). We acknowledge that an important and essential next step in this research is to explore the impact of a responsive split questionnaire on the quality of the questions asked. We discuss this suggested future research in Section 5.3.2.

Expenditure	Rare	Dimension	PPS	Log1	Log2	Two-Bin	Five-Bin
TX5B	No	Precision		++	+++	+++	
TX6B				+++	+++	+++	++
TX8				+++	+++		
TX9B				++	++		
TX12A				+++	++++	++	++
TX12B				++	+++	++	+++
TX16A				++	++		
TX17A				+++	+++	+	
TX5B	No	Responsiveness	++	++	++	++	++
TX6B			++	+	+	+	++
TX8			++	+	+	++	++
TX9B			++	+	++	++	++
TX12A			+	+	+	+	+
TX12B			+	+	+	+	+
TX16A			++	++	++	++++	+++
TX17A			++	+	++	+	++
TX4A_4	Yes	Precision	+	+	+		
TX4B			+		++		
TX5A_1					+		
TX5A_2				+			
TX6A					+		
TX7A				+	++		
TX9C				+	+		
TX9D			+	+	+		+
TX12C_1TO9					+++		++
TX18A							
TX4A_4	Yes	Responsiveness	+	++	+	+	++
TX4B			++	++	++	+++	++
TX5A_1			++	++	++	+	+
TX5A_2			++	++	+	+	++
TX6A			++	+	+	+	++
TX7A			++	++	++	+	++
TX9C			+	+	++	++	+
TX9D			++	++	++	+	++
TX12C_1TO9			+	+	+	+	+
TX18A			++	++	++	+	++

Table 5.2: Summary of research (**not recurrent**)

of customizing the survey to the individual respondent, but deficient in terms of satisfying stakeholder needs, and vice-versa. In general, an attempt to refine a method, e.g., by incorporating different covariates into the models, might improve some aspect of the design, but it may adversely affect another aspect. Therefore, it is essential to have a clear understanding of stakeholder utility functions, i.e., interests and objectives, costs, and tradeoffs among these, so that survey designers can make informed decisions regarding the "best" method.

Third, formulating the setup for the responsive split questionnaires using stratification methods deviates slightly from the approach taken in the PPS and the logistic regression methods. This is because the objective function minimized for the stratification methods sums across all expenditure categories. As a consequence, the decision to ask about a particular expense using stratification methods was influenced by characteristics of the other expenditures. We formulated the problem this way mostly out of convenience, essentially to solve one optimization problem. If we desired the stratification methods to mimic the PPS and logistic regression methods, then we might consider setting up a separate optimization problem for each expenditure category. Of course, as the number of expenditures increases so does the number of optimization problems; however, the primary advantage would be that the responsive split questionnaires using stratification methods would be responsive at the expenditure category-level.

Finally, our research may serve as a general guide of the steps a survey program should take if it wished to implement a responsive split questionnaire. First, and perhaps most importantly, the survey program must understand what data are

available to use in the design (i.e., can be collected in an initial phase of data collection), the nature of the relationships among the available data, and how those relationships can be used to make design decisions for subsequent phases of data collection. This requires a deep understanding of the underlying behavior the survey is attempting collect information about and could entail extensive model building to do so. Second, the survey program must propose appropriate evaluation criteria for key aspects of the new design. Each aspect will have a different set of evaluation criteria and it is important to acknowledge that each stakeholder may have a different ranking of these criteria. Third, it must conduct some sort of sensitivity analysis, for example, a simulation study or a field test, to assess whether those criteria are met and whether there are tradeoffs among stakeholder needs. Finally, it should consider a range of methods. As our research suggested, it is unlikely that one method will outperform all other methods with respect to every evaluation criteria. By considering a range of methods the survey program can make informed decisions regarding the optimal procedures to implement given their specific constraints.

5.3 Additional areas for future research

There are several additional issues pertaining to responsive split questionnaire methods that could be explored. In this dissertation, we focused our attention on a specific situation – a panel survey in which information from the first interview is used to determine which questions are administered to a sample unit in the second interview. While this represents a specific survey application, this is not an uncom-

mon situation in survey methodology, so our research still has broad applicability. However, to broaden the applicability of our methods and acknowledge the importance of conducting future research in this area, we identify and briefly discuss a few additional areas that may warrant further study.

5.3.1 Modeling and computing requirements

Our proposed methods rely heavily on the modeling of demographic and expenditure information collected in the first interview. In addition, since we developed models at the expenditure category-level, we engaged in a fairly lofty modeling exercise. However, we believe that it is not unreasonable to build models for each expenditure category because the computing requirements proposed in this research are somewhat typical of responsive designs. For instance, in the National Survey on Family Growth (NSFG), propensities for a case being interviewed on the next call are estimated nightly and then decisions are made about which sample units to target in a subsequent phase of data collection (Groves and Heeringa, 2006).

The primary goal of modeling expenditure and demographic information was to obtain a good prediction of whether the sample unit will incur a particular expense during the second interview reference period. Based on that prediction, we would more frequently (or not) ask about that expenditure category at the subsequent interview. We used a simple prediction model consisting of six demographic characteristics for all expenditure categories, most of which are currently included in the official CE nonresponse adjustment. While a simple model may suffer from

biased coefficients and a biased prediction, an overly complicated model may result in large variances, both in the coefficients and the prediction (Meyers, 1990). In this dissertation, we could have devoted significant efforts to building the best model for each expenditure category; however, we chose to consider a range of methods with a simple model. By demonstrating that a relatively simple model can still yield significant gains (e.g., in terms of tailoring the survey to the individual respondent) in a variety of situations, we can devote future research to more elaborate modeling exercises. Specifically, we can add complexity to the models for the various methods by including more and/or different covariates for each expenditure. In addition, we can consider other modeling techniques, such as classification and regression tress (Breiman et al., 1993), since logistic regression might not be the best technique for modeling rare events.

Another consequence of developing decision rules at the expenditure-level is that when estimating desired quantities about these expenditures, we needed to use a unique adjustment for each estimator. This adjustment was the inverse of the sample unit's subsampling probability for that expenditure category. That is, there would not be just a single weight for a consumer unit. For a responsive split questionnaire (as considered in this research), as the number of survey items increases, the modeling exercise becomes more extensive and the number of adjustments that data users need becomes quite large. Thus, future research could also include the development of a simpler, more manageable, and general adjustment procedure to use when these designs are implemented. Perhaps the survey methodological research pertaining to generalized variance functions (Jang et al., 1997; Johnson and King,

1987; Valliant, 1987) might help us gain traction on this problem. The purpose of generalized variance functions is to provide a quick and simple way to calculate standard errors for survey estimates, so our analogous method would be to provide a quick and simple way to provide weighted estimates of the desired quantities.

5.3.2 Context effects

Briefly defined, context effects in surveys suggest that responses to survey questions can be affected by prior items administered in the questionnaire as these prior items may provide cognitive cues to the respondent (Johnson et al., 1998). The methods developed in this research effectively assume that eliminating questions from the questionnaire will not negatively impact the respondent's ability to retrieve encoded events from their memory and report them for items actually asked. If they did negatively impact the response process, then we implicitly assumed that any measurement errors arising from context effects are offset by the benefits of (1) tailoring the questionnaire to the individual respondent and (2) reducing the length of the questionnaire. These assumptions may be tenuous; thus, it may be worth exploring the survey methodological research on context effects and conducting additional cognitive tests with responsive split questionnaires to gain an understanding of whether not asking certain questions adversely affects the survey responses. In other words, we will attempt to address the following research question: what will be the impact of implementing our methods on the measurement errors among the questions actually asked?

5.3.3 Effect on field staff

In the preceding section, we mentioned that a responsive split questionnaire might have an unanticipated negative impact on the respondent's recollection of certain expenditures. It is also possible that a responsive split questionnaire may adversely affect the interviewer's ability to administer the survey correctly. A dynamically changing interview may be quite challenging to administer to a respondent. Since the questions asked to a respondent could change across their tenure in the survey panel and the questions could be different across sample units within an interview, the interviewer will not be able to anticipate the next question in sequence for any given interview. This type of design requires a commitment from the field staff to actively pay attention to the survey instrument in order to administer it correctly to the respondent. Therefore, once the design is in place, usability testing, under a variety of scenarios, would be necessary before implementing it in the field. Extensive training would also need to be provided to the interviewers prior to allowing them to administer these instruments to actual sample units. Future research could focus on these operational issues.

5.3.4 Absence of panel survey design

Our research hinges on the fact that we have a panel survey in which we use the information collected in the first interview to determine the questions to ask at a subsequent interview. In the absence of this information, survey designers wishing to implement a responsive split questionnaire must look for other sources

of information for decision rule development. This is also the case for expenditure categories that are not asked about in the first interview[2]. Cochran (1977) identifies a few additional ways a survey designer may obtain the prior information necessary for the proposed methods. These include: (1) taking a smaller random sample from the population and estimating the desired quantities (e.g., estimating model parameters); (2) using the results of a pilot study, historical survey data from prior administrations of the survey, or other close sources of data; and, (3) guesswork about the structure of the population, aided by mathematical models and frame data.

One additional source of prior information that is absent from this list is the data users. Often, data users have a wealth of information that may provide survey designers with additional insight into the decision rule development process. Not only may data users provide the values of the various parameters for the models, but they may also be able to offer suggestions for the functional form of the regression models, covariates to include, constraints on the system, and evaluation criteria. Thus, future research to extend these methods to the non-panel surveys would require an exploration of alternative data sources to use when developing decision rules.

[2]See Table A.2 of Appendix A.2 for a listing of some expenditure categories that are only collected in the second interview.

5.3.5 First interview nonresponse

For the methods developed in this dissertation, data collected during the first interview were used to determine which questions were administered in the second interview. Since we evaluated our methods via simulation studies, we restricted our analysis file to only first and second interview respondents. In practice, however, if a sample unit fails to respond to the first interview survey request, then the amount of information which can be used to determine the decision rules for that unit is severely hampered. Thus, future research could focus on identifying the appropriate protocol or procedures to follow in the presence of first interview unit nonresponse for a responsive split questionnaire.

In some surveys, first interview nonrespondents are "lost to follow-up." This is the easiest solution because nonrespondents are dropped from the panel entirely and interviews are no longer attempted with these units. In this situation, first interview nonresponse would not pose problems beyond the usual issues associated with nonresponse (e.g., potential for bias in estimates, inflation in customary estimates of precision). In other surveys, including the Consumer Expenditure Surveys, first interview nonrespondents are contacted for a second interview, but instead of using the standard second interview questionnaire, a modified version of it is administered to the respondent. This instrument collects some information that should be collected during the first interview (e.g., inventory of household durables and household roster data) in addition to the full battery of second interview survey questions. So, each of these methods could be explored in future research to assess their value in

compensating for first interview nonresponse in a responsive split questionnaire.

Another solution might be to have interviewers collect auxiliary information about the sample unit from their surrounding environment (within the typical confidentiality constraints). This information could then be used as inputs into models and propensities/predictions based on these inputs could be obtained and used as decision rules. It is worth acknowledging that there is a growing body of literature on the value-added in using interviewer observed auxiliary information and this research may provide us with insight on how to address the issue of first interview nonresponse (Kreuter et al., 2010; West, 2010).

5.3.6 Extensions beyond the second interview

In our research, we limited our focus to the first two interviews; however, an obvious next step would be to extend our methods beyond the second interview. This would potentially require an exploration of alternative modeling techniques. Survival analysis techniques (Kalbfleisch and Prentice, 2002) might be worth exploring for modeling the occurrence of events after the second interview because sample units would likely have multiple, i.e., recurring events (e.g., purchases). Survival analysis techniques are flexible enough to handle these types of situations.

In addition to modeling expenditure and demographic information at later reference periods, there are a number of cognitive issues that also warrant investigation. One advantage of our setting is that the second interview remains bounded. In a bounded interview, interviewers typically employ dependent interviewing tech-

niques to review with the respondent his/her responses to questions in the previous interview (Groves et al., 2004). The purpose is to reduce telescoping errors, i.e., erroneously reporting events that occurred prior to the start of the reference period inquired about at the given interview (Groves et al., 2004). However, if a new question is asked in the third interview (e.g., about an expenditure category that was not asked in the second interview), then there is a potential for telescoping errors to occur. Thus, future research may focus on how to mitigate the potential occurrence of telescoping errors for unbounded questions in a responsive split questionnaire.

Another issue is that the current CE protocol encourages sample units to gather records (e.g., receipts, billing statements, checkbook registers) and use them during the interview to report expenditure information. A respondent might only gather records related to those expenditure categories asked in the second interview. So, if new or different questions are asked in a third interview, then the sample unit may not have easy access to records related to those expenditures. Thus, additional research might focus on how to encourage sample units to maintain a more extensive record system.

5.3.7 Incorporating data quality metrics into the decision rules

Although panel surveys provide unique measurement opportunities, e.g., the direct quantification of gross change, they pose additional concerns for data quality, in terms of nonresponse and measurement error, above and beyond cross-sectional or one-time surveys. The defining feature of all panel surveys is that observations

on the same sample members are taken through time. A problem with this is that it requires sample members to initially and continually respond to each survey request throughout their tenure in the panel. Although we already identified future research pertaining to first interview nonresponse in Section 5.3.5, it is important to understand that failure to observe all sample units at each prescribed time-point may result in nonresponse error if the values of the estimates derived only from the observed sample units differ from those based on the entire sample. Furthermore, there is the potential for nonresponse bias if there is a correlation between the sample unit's response propensity and their value on the substantive variable of interest.

Another problem with panel surveys is that there is a greater potential for respondents to learn and become familiar with the survey as they have multiple exposures to it over time. On the one hand, this can be beneficial because motivated respondents might learn what information is needed in the survey and therefore have the desired information ready at each subsequent request. On the other hand, cooperative yet less motivated respondents might learn how their responses affect the sequence of questions through the interview and thus modify their responses in hopes of easing the burden of the survey request (perhaps by reducing the length). Addressing this latter concern was used, in part, as motivation for our research. At any rate, the resulting mismatch between the respondent's true value and their actual response to the survey question results in measurement error. If these errors systematically depart from the truth, then measurement biases are likely to occur.

When the decision to implement a panel survey is made survey practitioners anticipate the occurrence of certain nonresponse and measurement errors and make

design decisions informed by those. Previous research on panel surveys has focused on the effects of increased effort, statistical adjustments, and questionnaire modifications on alleviating some of these concerns. We have suggested, throughout this research, that reducing the length of the questionnaire via split questionnaire methods may address some of these same nonresponse and measurement error concerns.

To further address these concerns, a reasonable extension of our methods might be to incorporate data quality metrics and/or indicators in the decision rules for a responsive split questionnaire. The primary goal would be to use information collected during Phase 1 (or some prior phase) to identify sample members that would be "good" reporters/respondents and administer a specialized subset of questions to them, or perhaps a more complete/full version of the questionnaire because these units may provide "good" responses regardless of the content, length, and/or other features of the survey. Similar methods can be devised when identifying "bad" reporters/respondents; however, their specialized subset of questions would likely be a small subset of questions for which reasonable responses could be provided rather quickly (e.g., global expenditure questions), of reasonable quality, and without much perceived burden.

5.3.8 Uncertainty in inputs

Survey organizations must consider a large number of design features that may have a substantial impact on both survey costs and data quality. However, decisions about particular design features are often complicated by the lack of relevant

information on the fixed and marginal costs of these design features as well as the relationship between the design features and data quality. To move forward with any survey operation, conjectures, sometimes based on limited and/or questionable information, about the cost and data quality properties of possible design features must be made. A further complication is that these conjectures oftentimes suggest that particular design features might have good data quality properties (from a total survey error perspective as well as other factors such as timeliness, accessibility, etc.) but they comprise a costly set of design features. Because of this conflict, the survey designers are then charged with the task of defining and determining an optimal set of cost-quality tradeoffs associated with choosing among a seemingly uncountable set of design features while using potentially imperfect information.

As survey designers, one of the first examples of a design-optimization problem involving tradeoffs that we encounter is to determine the sample allocation for a stratified sample while minimizing the sampling variance of an estimator for a population quantity subject to some cost constraint (and vice versa – minimize the cost of the survey subject to a specified sampling variance). We provided this example in Section 2.5.2. This is a gentle introduction to the broad class of design-optimization problems, but problems this simplistic are, by far, the exception in practice. In fact, if we approach any problem with this naivety after having any experience with designing surveys, then some unsettling characteristics of this problem should surface. One notable feature is that when carrying out this exercise we require at the very least (accurate/correct) information on the following: population counts for each stratum, stratum-specific variances of the characteristic of interest and stratum-

specific per unit costs. It is often the case that this information is gleaned from experiences with previous survey endeavors. In our research, we estimated these quantities from the first interview and assumed that they were accurate reflections of reality for the units for which the split questionnaire was designed. However, the conditions under which prior surveys and prior phases of data collection were conducted may not be applicable to the current survey. Thus, any evaluation of the tradeoffs between cost and quality incurred must account for the uncertainty in the information on which design decisions are based.

Although our example in Section 2.5.2 highlights the basic standard design-optimization approach, identifying the potential deficiency of using inputs wrought with error provides motivation for extending our research. In particular, future research may focus on accounting for the uncertainty in the information used as inputs to inform the design decisions, e.g., determine the subsampling probabilities. Special attention could be devoted to the level of precision required for the cost and quality information to adequately inform the design decisions and the sensitivity of cost-quality tradeoffs to changes in the assumptions regarding functional forms (e.g., the relationship between a target population and a design feature).

5.3.9 Meeting a variety of analytic objectives

In advance of designing a survey, survey designers cannot anticipate all potential uses of the collected data. It is possible that by implementing the methods developed in this dissertation, some data uses may not be met. For instance, some

higher-ordered interaction terms for a regression might not be estimable because the survey items associated with those might not be administered to the same set of sample members (see Figure 2.1[f]). Many of these problems can be circumvented by specifying that a subsample of the main sample would receive the full questionnaire (see Figure 2.1[c]). This feature would accommodate a vast array of data users with various analytic objectives. Of course, determining which sample units would be included in this full questionnaire subset becomes an interesting problem to explore. In Section 5.3.7, we provided a possible recommendation for which sample units to administer the full questionnaire. It may also be worthwhile exploring the use of imputation methods to "fill in" the data not collected.

Appendix A

Data Summary

A.1 Section listing of the CE

Section 1 - General Survey Information

- Part A - Reference Period
- Part B - General Housing Characteristics
- Part C - Major Household Appliances

Section 2 - Rented Living Quarters

Section 3 - Owned Living Quarters and Other Owned Real Estate

- Part A.1 - Screening Questions
- Part A.2 - Screening Questions: For New Households Only
- Part B - Detailed Property Description
- Part D - Disposed of Property
- Part E - Mortgage/Home Equity Loan Screening Questions
- Part F - Mortgages/Lump Sum Home Equity Loan
- Part H - Line of Credit Home Equity Loans
- Part I - Ownership Costs
- Part J - Change in Mortgage/Lump Sum Home Equity Loan

Section 4 - Utilities and Fuels for Owned and Rented Properties

- Part A - Telephone Expenses

Part B - Other Telephone Expenses

Part C - Internet Service Expenses

Part D - Utilities and Fuels for Owned and Rented Properties

Section 5 - Construction, Repairs, Alterations, and Maintenance of Property

Section 6 - Appliances, Household Equipment, and other Selected Items

Part A - Purchase of Household Appliances

Part B - Purchase of Household Appliances and Other Selected Items

Section 7 - Household Item, Repairs, and Service Contracts

Section 8 - Home Furnishings and Related Household Items

Part A - Purchases of Home Furnishings and Related Household Items

Part B - Rental, Leasing, or Repair of Furniture

Section 9 - Clothing and Sewing Materials

Part A - Clothing

Part B - Infants Clothing, Watches, Jewelry and Hairpieces

Part C - Clothing Services

Part D - Sewing Materials

Section 10 - Rented and Leased Vehicles

Section 11 - Owned Vehicles

Section 12 - Vehicle Operating Expenses

Part A - Vehicle Maintenance and Repair, Parts and Equipment

Part B - Licensing, Registration, and Inspection of Vehicles

Part C - Other Vehicle Operating Expenses

Section 13 - Insurance Other Than Health

Part A.1 - Screening Questions

Part A.2 - Screening Questions: For New Households Only

Part B - Detailed Questions

Part B.1 - Detailed Questions: For New Households Only

Section 14 - Hospitalization and Health Insurance

Part A.1 - Screening Questions

Part A.2 - Screening Questions: For New Households Only

Part B - Detailed Questions

Part C - Medicare, Medicaid, and Other Health Insurance Plans Not Directly Paid For By The Household

Section 15 - Medical and Health Expenditures

Part A - Screening Questions for Payments

Part B - Screening Questions for Reimbursements

Section 16 - Educational Expenses

Section 17 - Subscriptions, Memberships, Books, and Entertainment Expenses

Part A - Subscriptions and Memberships

Part B - Books and Entertainment

Section 18 - Trips and Vacations

Part A - Screening Questions

Part BC - Detailed Questions

Part E - Trip Expenses for Non-Household Members

Part F - Local Overnight Stays

Section 19 - Miscellaneous Expenses

Part A - Miscellaneous Expenses

Part B - Contributions

Section 20 - Expense Patterns for Food, Beverages, and Other Selected Items

Part A - Food and Beverages

Part B - Selected Services and Goods

Section 21 - Credit Liability

Part A.1 - Credit Balances - Second Quarter Only

Part A.2 - Credit Balances - Fifth Quarter Only

Part B - Finance Charges - Fifth Quarter Only

Section 22 - Work Experience and Income

Part A - Second Quarter, Fifth Quarter, or New Households Only

Part B - Second Quarter, Fifth Quarter, or New Households Only - Ask for entire Household as a group

Part G - Change In Assets - Fifth Quarter Only

A.2 Mapping of CE sections to expenditure variables

Section	Variable name	Variable description	Wave 2 only?
2	TX2	Rental payment	No
3F	TX3F	Mortgage/lump sum home equity loan (non-fixed rate mortgages)	No
3H	TX3H	Home loan line of credit	No
3J	TX3J_1	Principal and/or interest (fixed rate mortgages)	Yes
4A	TX4A_1	Telephone services (residential and mobile)	No
	TX4A_2	Internet access	No
	TX4A_3	Cable/satellite	No
	TX4A_4	Non-telephone (e.g., modem purchase, apps, ringtones)	No
4B	TX4B	Telephone cards, prepaid cell, public pay phone	No
4C	TX4C	Cable/satellite not reported (e.g., satellite radio, Internet caf)	No
4D	TX4D	Utilities and fuels for owned and rented properties	No
5	TX5A_1	Construction materials (for specific jobs not yet started)	No
	TX5A_2	Construction materials (for general jobs)	No
	TX5B	Contractor labor, materials, and tools	No
6A	TX6A	Major appliance installation, cost, and rental	No
6B	TX6B	Minor appliance price, rental, and installation	No
7	TX7A	Household item repair and service contracts	No
8	TX8	Furniture rental and repair	No
9A	TX9A	Clothing	No
9B	TX9B	Infant's clothing, watches, jewelry, and hairpieces	No
9C	TX9C	Clothing services	No
9D	TX9D	Sewing materials	No
10	TX10	Car lease termination fee and vehicle rental	Yes
	TX10C_1	Car monthly payment (for leased vehicles)	No
	TX10C_23	Car down payment and fees (for leased vehicles)	No
11	TX11B	Owned car down payment	No
12A	TX12A	Vehicle service and parts	No
12B	TX12B	Vehicle license fees	No
12C	TX12C_10	Average monthly gas expense	No
	TX12C_1TO9	Other vehicle fuels (e.g., tank gas, fluids) and fees	No
13B	TX13B	Non-health insurance expense (e.g., life, auto, homeowner)	No
14B	TX14B	Regular amount for health insurance	No
15A	TX15A	Medical service payments (e.g., eye exams, lab tests, x-rays)	No
16	TX16A	Educational expenses (e.g., tuition, recreational lessons)	No
17A	TX17A	Cost of subscriptions (e.g., newspapers, theater season tickets, gym)	No
18A	TX18A	Total amount paid by CU for a reimbursed trip	No
18B	TX18B	Package trips	Yes
19A	TX19A	Miscellaneous expenses (e.g., funerals, babysitting, flowers)	No
19B	TX19B	Cash contributions (e.g., child support, alimony, charities)	No
20A	TX20A_1	Weekly food, wine, and other food-type expenses	Yes
	TX20A_2	Only weekly food expenses	Yes

Table A.1: Mapping of expenditure variables to survey sections

A.3 Analysis file demographic characteristics

Variable name	Variable description	Levels
POVERTY	Indicates whether 20% of more of the persons in the tract are living in poverty for old construction	1: 20% or more of the population live in poverty 2: Less than 20% live in poverty
REG_OFF	Regional office	21: Boston 22: New York 23: Philadelphia 24: Detroit 25: Chicago 26: Kansas City 27: Seattle 28: Charlotte 29: Atlanta 30: Dallas 31: Denver 32: Los Angeles
URBAN	Urban/Rural for 2000 sample design	1: Urban 2: Rural
TENURE	Household tenure	1: Owner 2: Renter
RACE	Race of CU	1: Black 2: Non-black
SIZE	Number of CU members	1: 1 CU member 2: 2 CU members 3: 3 or 4 CU members 4: 5 or more CU members
CONVREF	Was this a converted refusal?	1: Yes 2: No
LANGUAGE	In what language was this interview conducted?	1: English 2: Spanish 3: Other

Table A.2: Listing of demographic characteristics

Appendix B

Summary statistics for expenditure information prior to data cleaning

Expenditure	Interview 1	Interview 2
TX2	2,250	6,480
TX3F	6,000	19,869
TX3H	3,200	12,000
TX4A_1	346	1,020
TX4A_2	70	195
TX4A_3	130	375
TX4A_4	110	140
TX4B	150	220
TX4C	180	327
TX4D	750	1,781
TX5A_1	3,000	2,500
TX5A_2	1,250	700
TX5B	12,499	16,054
TX6A	2,900	2,947
TX6B	2,047	2,500
TX7A	700	709
TX8	2,604	2,880
TX9A	1,026	1,443
TX9B	750	1,042
TX9C	200	379
TX9D	230	374
TX10C_1	1,571	1,571
TX10C_23	12,000	14,000
TX11B	18,000	18,000
TX12A	1,500	1,900
TX12B	384	476
TX12C_1TO9	80	616
TX12C_10	700	700
TX13B	2,000	2,162
TX14B	1,700	1,500
TX15A	1,640	2,520
TX16A	6,100	8,855
TX17A	760	1,134
TX18A	3,000	2,000
TX18B	...	7,458
TX19A	1,645	2,375
TX19B	2,200	3,007

Table B.1: 97.5th percentiles for interviews 1 and 2 (before top-coding)

Expenditure	n reporting	Mean	Std Dev	Min	Q1	Median	Q3	95^{th} Pctl	99^{th} Pctl	Max
TX2	2,921	219.65	490.83	0.00	0.00	0.00	242.00	1,130.00	2,000.00	13,000.00
TX3F	473	72.85	530.29	0.00	0.00	0.00	0.00	0.00	2,085.00	25,000.00
TX3H	604	71.56	3,206.64	0.00	0.00	0.00	0.00	150.00	900.00	325,000.00
TX4A_1	9,606	127.99	1,160.39	0.00	50.00	100.00	160.00	290.00	424.00	118,585.00
TX4A_2	3,023	10.16	18.58	0.00	0.00	0.00	20.00	45.00	60.00	400.00
TX4A_3	1,364	7.85	22.84	0.00	0.00	0.00	0.00	60.00	100.00	280.00
TX4A_4	42	0.10	2.77	0.00	0.00	0.00	0.00	0.00	0.00	196.00
TX4B	728	21.54	1,380.27	0.00	0.00	0.00	0.00	15.00	60.00	99,999.00
TX4C	6,634	50.88	54.84	0.00	0.00	49.00	84.00	146.00	200.00	1,118.00
TX4D	9,619	231.73	203.95	0.00	100.00	196.00	309.00	578.00	986.00	3,316.00
TX5A_1	254	10.34	140.56	0.00	0.00	0.00	0.00	0.00	200.00	6,000.00
TX5A_2	94	2.36	104.34	0.00	0.00	0.00	0.00	0.00	0.00	10,000.00
TX5B	1,374	259.11	1,806.62	0.00	0.00	0.00	0.00	860.00	6,000.00	50,000.00
TX6A	444	30.49	237.13	0.00	0.00	0.00	0.00	0.00	1,000.00	7,200.00
TX6B	3,466	115.51	400.88	0.00	0.00	0.00	49.00	601.00	1,955.00	8,000.00
TX7A	686	12.17	244.50	0.00	0.00	0.00	0.00	36.00	250.00	24,000.00
TX8	2,948	106.55	968.96	0.00	0.00	0.00	15.00	428.00	2,020.00	84,500.00
TX9A	5,719	129.81	298.63	0.00	0.00	22.00	150.00	520.00	1,200.00	7,620.00
TX9B	1,764	26.62	247.46	0.00	0.00	0.00	0.00	100.00	370.00	13,200.00
TX9C	444	1.95	16.34	0.00	0.00	0.00	0.00	0.00	51.00	500.00
TX9D	473	2.35	21.20	0.00	0.00	0.00	0.00	0.00	60.00	850.00
TX10C_1	391	31.25	927.30	0.00	0.00	0.00	0.00	0.00	556.00	87,000.00
TX10C_23	158	44.87	527.48	0.00	0.00	0.00	0.00	0.00	1,500.00	23,143.00
TX11B	1,781	683.34	2,566.93	0.00	0.00	0.00	0.00	5,000.00	11,000.00	80,000.00
TX12A	3,700	98.58	299.68	0.00	0.00	0.00	40.00	600.00	1,400.00	7,000.00
TX12B	1,591	14.69	58.00	0.00	0.00	0.00	0.00	90.00	270.00	2,175.00
TX12C_1TO9	130	0.26	4.17	0.00	0.00	0.00	0.00	0.00	5.00	306.00
TX12C_10	9,358	201.58	212.87	0.00	70.00	150.00	290.00	560.00	900.00	6,000.00
TX13B	5,533	216.05	685.44	0.00	0.00	30.00	215.00	910.00	2,250.00	40,200.00
TX14B	2,081	64.34	397.53	0.00	0.00	0.00	0.00	300.00	1,000.00	18,000.00
TX15A	5,052	132.94	455.97	0.00	0.00	0.00	100.00	580.00	1,910.00	17,000.00
TX16A	1,898	176.19	1,837.75	0.00	0.00	0.00	0.00	624.00	3,100.00	138,800.00
TX17A	2,671	32.79	276.90	0.00	0.00	0.00	4.00	110.00	471.00	15,002.00
TX18A	202	7.03	106.33	0.00	0.00	0.00	0.00	0.00	150.00	4,900.00
TX19A	6,276	182.13	1,022.27	0.00	0.00	25.00	130.00	649.00	2,185.00	55,660.00
TX19B	5,252	240.04	852.56	0.00	0.00	1.00	200.00	1,200.00	2,450.00	25,400.00

Table B.2: Summary statistics for interview 1 expenditures (before top-coding, non-reports treated as zeros)

Expenditure	n reporting	Mean	Std Dev	Min	Q1	Median	Q3	95^{th} Pctl	99^{th} Pctl	Max
TX2	2,941	644.40	1,352.24	0.00	0.00	0.00	750.00	3,300.00	5,748.00	21,000.00
TX3F	491	232.34	1,850.58	0.00	0.00	0.00	0.00	0.00	6,300.00	112,800.00
TX3H	589	149.86	1,947.76	0.00	0.00	0.00	0.00	375.00	2,919.00	89,185.00
TX4A_1	9,710	355.43	273.50	0.00	153.00	300.00	497.00	867.00	1,194.00	4,320.00
TX4A_2	2,811	28.00	51.68	0.00	0.00	0.00	57.00	135.00	180.00	600.00
TX4A_3	1,367	24.08	69.20	0.00	0.00	0.00	0.00	195.00	300.00	720.00
TX4A_4	24	0.11	3.07	0.00	0.00	0.00	0.00	0.00	0.00	140.00
TX4B	866	5.17	25.66	0.00	0.00	0.00	0.00	30.00	126.00	720.00
TX4C	6,839	63.45	78.47	0.00	0.00	50.00	95.00	199.00	390.00	868.00
TX4D	9,662	623.19	501.13	0.00	300.00	558.00	844.00	1,420.00	2,201.00	16,089.00
TX5A_1	185	7.23	121.29	0.00	0.00	0.00	0.00	0.00	96.00	6,100.00
TX5A_2	98	4.22	293.59	0.00	0.00	0.00	0.00	0.00	0.00	30,000.00
TX5B	1,905	463.77	3,601.57	0.00	0.00	0.00	0.00	1,771.00	9,825.00	160,400.00
TX6A	634	43.76	479.20	0.00	0.00	0.00	0.00	100.00	1,174.00	42,000.00
TX6B	4,609	203.02	630.56	0.00	0.00	0.00	142.00	1,040.00	2,651.00	22,250.00
TX7A	1,027	17.53	89.46	0.00	0.00	0.00	0.00	105.00	388.00	3,546.00
TX8	3,887	155.15	736.13	0.00	0.00	0.00	50.00	705.00	2,696.00	32,000.00
TX9A	7,022	218.49	400.04	0.00	0.00	84.00	285.00	830.00	1,759.00	7,000.00
TX9B	2,383	42.77	226.14	0.00	0.00	0.00	0.00	212.00	672.00	10,000.00
TX9C	714	5.01	61.61	0.00	0.00	0.00	0.00	15.00	107.00	4,240.00
TX9D	672	4.35	37.23	0.00	0.00	0.00	0.00	10.00	107.00	2,000.00
TX10C_1	418	32.74	927.97	0.00	0.00	0.00	0.00	0.00	600.00	87,000.00
TX10C_23	176	53.53	658.18	0.00	0.00	0.00	0.00	0.00	2,000.00	35,600.00
TX11B	1,864	731.28	2,695.53	0.00	0.00	0.00	0.00	5,000.00	11,000.00	80,000.00
TX12A	5,741	182.05	481.56	0.00	0.00	25.00	150.00	879.00	2,086.00	12,000.00
TX12B	2,928	31.72	91.77	0.00	0.00	0.00	20.00	169.00	400.00	3,000.00
TX12C_1TO9	2,852	30.14	173.97	0.00	0.00	0.00	5.00	131.00	469.00	11,715.00
TX12C_10	9,365	195.49	201.03	0.00	70.00	150.00	250.00	520.00	900.00	4,000.00
TX13B	6,815	384.62	608.59	0.00	0.00	213.00	541.00	1,358.00	2,560.00	16,401.00
TX14B	2,814	75.33	407.47	0.00	0.00	0.00	10.00	361.00	1,100.00	19,000.00
TX15A	6,463	315.41	3,103.22	0.00	0.00	48.00	267.00	1,210.00	3,178.00	302,950.00
TX16A	2,425	277.93	1,974.82	0.00	0.00	0.00	0.00	1,144.00	5,615.00	140,450.00
TX17A	3,502	61.36	282.56	0.00	0.00	0.00	39.00	251.00	980.00	14,060.00
TX18A	316	10.29	109.94	0.00	0.00	0.00	0.00	0.00	225.00	3,000.00
TX19A	7,085	285.01	1,312.88	0.00	0.00	51.00	241.00	1,065.00	3,305.00	100,750.00
TX19B	6,400	448.34	2,498.39	0.00	0.00	50.00	500.00	1,800.00	3,827.00	204,300.00

Table B.3: Summary statistics for interview 2 expenditures (before top-coding, non-reports treated as zeros)

Appendix C

Supplemental analyses for Chapter 3

C.1 Bivariate cross-interview correlations of expenditures across the two interviews (after top-coding, non-reports treated as zeros)

Wave 2 (across) Wave 1 (down)	TX2	TX3F	TX3H	TX3J_1	TX4A_1	TX4A_2	TX4A_3	TX4A_4	TX4B	TX4C	TX4D
TX2	0.929	-0.068	-0.070	-0.031	-0.067	-0.065	-0.037	-0.007	0.044	-0.043	**-0.263**
TX3F	-0.070	0.882	0.058	**0.289**	0.076	0.027	0.013	0.002	-0.007	0.039	0.118
TX3H	-0.079	0.041	0.568	-0.003	0.091	0.036	0.051	-0.006	-0.014	0.041	0.106
TX4A_1	-0.060	0.061	0.084	0.027	0.812	0.365	0.336	0.028	-0.059	0.044	**0.321**
TX4A_2	-0.058	0.014	0.058	0.010	0.401	0.622	0.373	0.047	-0.014	-0.131	0.166
TX4A_3	-0.021	0.009	0.030	0.011	0.356	0.391	0.678	0.064	-0.028	**-0.246**	0.137
TX4A_4	-0.009	0.002	0.045	-0.003	0.035	0.046	0.067	0.016	-0.001	-0.019	0.027
TX4B	0.043	-0.018	-0.013	-0.008	-0.058	-0.021	-0.028	0.010	0.386	-0.035	-0.016
TX4C	-0.058	0.041	0.049	0.030	0.060	-0.138	**-0.233**	-0.019	-0.040	0.683	0.182
TX4D	**-0.260**	0.101	0.103	0.018	**0.305**	0.134	0.121	0.022	-0.020	0.156	0.661
TX5A_1	-0.035	0.038	0.002	-0.006	0.042	0.021	0.011	-0.003	0.009	0.008	0.033
TX5A_2	-0.016	-0.003	0.013	-0.003	0.028	0.026	0.010	0.000	-0.002	0.000	0.009
TX5B	-0.067	0.013	0.045	-0.001	0.070	0.044	0.050	-0.005	0.008	0.040	0.087
TX6A	-0.028	0.002	0.013	0.020	0.050	0.036	0.030	-0.005	-0.002	0.020	0.065
TX6B	-0.030	0.045	0.023	-0.006	0.152	0.076	0.039	0.014	0.001	0.074	0.109
TX7A	-0.064	0.013	0.031	0.026	0.083	0.054	0.038	0.000	-0.017	0.050	0.109
TX8	-0.012	0.030	0.026	0.012	0.108	0.051	0.055	0.005	-0.012	0.035	0.096
TX9A	-0.004	0.068	0.073	0.020	**0.247**	0.105	0.087	0.011	0.009	0.089	0.187
TX9B	0.010	0.020	0.008	0.035	0.103	0.052	0.037	-0.002	0.014	0.048	0.079
TX9C	0.002	0.044	0.094	-0.007	0.068	0.035	0.033	0.003	-0.008	0.028	0.047
TX9D	-0.035	-0.009	0.018	-0.008	0.030	0.019	0.024	0.050	0.000	0.008	0.048
TX10C_1	-0.007	0.070	0.040	0.020	0.098	0.040	0.041	-0.005	-0.023	0.053	0.103
TX10C_23	0.001	0.062	0.029	0.012	0.075	0.041	0.047	-0.002	-0.008	0.023	0.055
TX11B	-0.020	0.018	0.001	0.011	0.136	0.073	0.056	0.014	-0.019	0.058	0.084
TX12A	-0.039	0.021	0.042	0.007	0.108	0.045	0.025	0.000	0.010	0.044	0.083
TX12B	-0.018	0.051	0.009	0.004	0.063	0.057	0.014	0.002	-0.003	0.038	0.062
TX12C_1TO9	-0.006	-0.001	0.008	-0.005	-0.003	-0.009	-0.007	0.001	0.001	0.002	0.009
TX12C_10	-0.108	0.071	0.073	0.000	**0.345**	0.153	0.076	0.006	0.000	0.124	**0.301**
TX13B	-0.115	0.042	0.065	0.003	0.178	0.097	0.073	0.005	-0.021	0.077	0.184
TX14B	-0.050	0.033	0.038	-0.005	0.033	0.026	0.023	0.012	-0.018	0.040	0.070
TX15A	-0.081	0.019	0.062	0.005	0.126	0.084	0.044	0.010	-0.001	0.067	0.140
TX16A	-0.015	0.029	0.050	0.015	0.122	0.063	0.040	0.020	0.002	0.035	0.109
TX17A	-0.051	0.032	0.067	0.006	0.101	0.053	0.042	-0.005	-0.019	0.086	0.097
TX18A	0.020	0.070	0.069	0.001	0.035	-0.004	-0.006	0.000	-0.001	0.026	0.027
TX19A	-0.053	0.081	0.089	0.009	0.184	0.092	0.075	-0.002	-0.018	0.102	**0.208**
TX19B	-0.059	0.040	0.066	-0.005	0.137	0.082	0.047	0.007	-0.013	0.057	0.134

Table C.1: Bivariate cross-interview correlations

Wave 2 (across) Wave 1 (down)	TX5A_1	TX5A_2	TX5B	TX6A	TX6B	TX7A	TX8	TX9A	TX9B	TX9C	TX9D
TX2	-0.034	-0.027	-0.099	-0.049	-0.046	-0.088	-0.046	-0.019	0.021	0.005	-0.037
TX3F	0.007	0.001	0.060	0.014	0.018	0.027	0.048	0.096	0.024	0.054	0.003
TX3H	0.029	0.011	0.021	0.012	0.054	0.043	0.030	0.088	0.007	0.069	0.036
TX4A_1	0.025	0.023	0.058	0.043	0.162	0.104	0.091	**0.261**	0.090	0.068	0.038
TX4A_2	0.019	0.002	0.054	0.017	0.092	0.058	0.057	0.140	0.050	0.047	0.026
TX4A_3	0.030	-0.005	0.021	0.010	0.063	0.031	0.042	0.097	0.048	0.025	0.030
TX4A_4	0.041	0.001	0.004	0.008	0.004	-0.006	-0.009	0.017	0.003	-0.005	0.009
TX4B	0.001	-0.006	-0.010	-0.002	0.009	0.009	-0.007	0.012	0.016	-0.004	-0.001
TX4C	0.006	0.017	0.069	0.025	0.103	0.082	0.083	0.116	0.045	0.045	0.014
TX4D	0.013	0.040	0.112	0.053	0.104	0.131	0.099	0.191	0.049	0.050	0.033
TX5A_1	0.024	0.022	0.069	0.014	0.079	0.034	0.045	0.045	0.041	0.016	0.007
TX5A_2	0.001	0.007	0.073	-0.004	0.006	0.013	0.013	0.030	0.021	0.006	0.005
TX5B	0.007	0.036	0.179	0.044	0.067	0.042	0.096	0.089	0.018	0.024	0.034
TX6A	0.020	0.010	0.011	0.039	0.055	0.024	0.068	0.022	-0.003	0.017	0.022
TX6B	0.032	0.010	0.052	0.027	0.129	0.042	0.097	0.168	0.058	0.043	0.052
TX7A	0.012	0.005	0.060	0.037	0.038	0.111	0.051	0.075	0.016	0.044	0.024
TX8	0.046	0.020	0.059	0.032	0.087	0.062	0.152	0.144	0.058	0.025	0.022
TX9A	0.041	0.025	0.064	0.047	0.143	0.068	0.125	0.365	0.111	0.112	0.047
TX9B	-0.001	0.001	0.031	0.031	0.049	0.016	0.072	0.127	0.219	0.048	0.018
TX9C	0.004	-0.007	0.023	-0.002	0.042	0.055	0.046	0.108	0.035	0.144	0.003
TX9D	0.004	0.008	0.003	0.008	0.059	-0.002	0.019	0.020	0.031	0.004	0.385
TX10C_1	0.012	-0.003	0.045	-0.009	0.029	0.008	0.070	0.120	0.062	0.038	-0.009
TX10C_23	0.008	-0.002	0.016	-0.011	0.042	0.012	0.050	0.081	0.033	0.018	0.014
TX11B	0.011	0.022	-0.004	0.024	0.055	0.046	0.050	0.093	0.048	0.013	0.029
TX12A	0.003	0.025	0.044	0.025	0.055	0.040	0.031	0.097	0.034	0.036	0.027
TX12B	0.010	0.002	0.026	0.004	0.065	0.028	0.036	0.053	0.036	0.006	0.024
TX12C_1TO9	0.002	-0.004	0.012	0.002	0.028	0.000	-0.003	0.023	0.014	-0.010	0.029
TX12C_10	0.029	0.033	0.056	0.049	0.156	0.081	0.101	**0.241**	0.104	0.040	0.045
TX13B	0.018	0.028	0.069	0.029	0.082	0.107	0.062	0.118	0.029	0.044	0.026
TX14B	0.019	-0.001	0.044	0.015	0.042	0.046	0.034	0.049	-0.006	0.057	0.015
TX15A	0.009	0.015	0.032	0.022	0.070	0.053	0.055	0.128	0.039	0.066	0.056
TX16A	-0.001	0.004	0.031	-0.010	0.077	0.047	0.033	0.139	0.036	0.039	0.016
TX17A	0.008	0.000	0.064	0.039	0.082	0.083	0.081	0.153	0.036	0.050	0.040
TX18A	0.006	0.051	0.003	0.007	0.048	0.030	0.034	0.059	0.004	0.029	0.004
TX19A	0.028	0.032	0.093	0.035	0.151	0.097	0.141	**0.212**	0.099	0.087	0.038
TX19B	0.040	0.014	0.066	0.036	0.090	0.100	0.054	0.150	0.045	0.073	0.041

Table C.2: Bivariate cross-interview correlations (2)

Wave 2 (across) Wave 1 (down)	TX10	TX10C_1	TX10C_23	TX11B	TX12A	TX12B	TX12C_1TO9	TX12C_10	TX13B	TX14B
TX2	0.011	-0.003	-0.004	-0.022	-0.038	-0.021	-0.017	-0.113	-0.146	-0.059
TX3F	0.042	0.065	0.077	0.024	0.043	0.043	0.046	0.066	0.051	0.030
TX3H	0.011	0.062	0.025	0.020	0.045	0.017	0.046	0.081	0.086	0.048
TX4A_1	0.027	0.108	0.069	0.134	0.161	0.116	0.136	**0.327**	**0.270**	0.038
TX4A_2	0.020	0.037	0.043	0.081	0.098	0.055	0.077	0.152	0.143	0.042
TX4A_3	0.005	0.044	0.044	0.062	0.050	0.023	0.046	0.079	0.100	0.012
TX4A_4	-0.002	0.001	0.002	-0.004	0.010	-0.005	0.018	0.009	0.024	0.017
TX4B	0.013	-0.012	0.001	-0.014	-0.005	-0.002	0.010	0.003	-0.010	-0.012
TX4C	0.029	0.072	0.038	0.075	0.073	0.056	0.076	0.154	0.150	0.055
TX4D	0.009	0.096	0.063	0.078	0.114	0.070	0.098	**0.270**	**0.246**	0.082
TX5A_1	0.011	0.006	-0.007	0.007	0.012	0.027	0.029	0.059	0.037	0.004
TX5A_2	-0.001	0.018	0.003	-0.010	0.016	0.042	0.012	0.039	0.024	0.002
TX5B	-0.002	0.021	0.010	0.014	0.035	0.006	0.061	0.052	0.064	0.013
TX6A	-0.005	0.012	-0.003	0.009	0.017	-0.009	0.023	0.047	0.032	-0.010
TX6B	0.002	0.023	0.001	0.058	0.092	0.045	0.059	0.137	0.104	0.031
TX7A	0.004	0.024	0.003	0.015	0.042	0.024	0.038	0.060	0.085	0.045
TX8	0.010	0.045	0.039	0.038	0.056	0.013	0.058	0.073	0.073	0.014
TX9A	0.020	0.110	0.065	0.076	0.131	0.053	0.142	**0.206**	0.162	0.045
TX9B	0.026	0.048	0.049	0.068	0.039	0.014	0.049	0.090	0.079	0.038
TX9C	0.009	0.033	0.037	0.031	0.043	0.042	0.051	0.049	0.063	0.049
TX9D	-0.007	-0.006	0.002	0.028	0.038	0.027	0.029	0.039	0.022	0.021
TX10C_1	0.087	0.968	0.387	-0.007	0.000	0.028	0.059	0.074	0.073	0.032
TX10C_23	0.086	0.418	0.914	0.016	0.004	0.016	0.057	0.050	0.061	0.006
TX11B	0.019	-0.010	0.008	0.965	0.072	0.056	0.055	0.137	0.093	0.002
TX12A	0.025	0.007	0.012	0.028	0.143	0.037	0.072	0.162	0.099	0.029
TX12B	0.033	0.023	0.039	0.047	0.053	0.018	0.065	0.101	0.092	0.034
TX12C_1TO9	-0.002	-0.008	0.004	-0.014	0.044	0.019	0.011	0.060	0.021	-0.014
TX12C_10	0.027	0.102	0.048	0.151	0.191	0.127	0.111	0.690	**0.247**	0.028
TX13B	0.025	0.072	0.053	0.059	0.082	0.061	0.086	0.158	0.266	0.059
TX14B	0.003	0.014	0.001	-0.006	0.022	0.028	0.033	0.029	0.077	0.794
TX15A	0.006	0.027	0.039	0.034	0.072	0.054	0.053	0.093	0.136	0.064
TX16A	0.008	0.044	0.017	0.046	0.052	0.033	0.092	0.100	0.097	0.049
TX17A	0.008	0.112	0.132	0.033	0.063	0.052	0.093	0.065	0.130	0.075
TX18A	0.001	0.009	0.000	0.001	0.053	0.014	0.024	0.019	0.024	0.023
TX19A	0.036	0.112	0.079	0.091	0.104	0.061	0.126	0.160	0.153	0.082
TX19B	0.038	0.070	0.086	0.046	0.102	0.066	0.065	0.135	0.152	0.056

Table C.3: Bivariate cross-interview correlations (3)

Wave 2 (across) Wave 1 (down)	TX15A	TX16A	TX17A	TX18A	TX18B	TX19A	TX19B	TX20A_1	TX20A_2
TX2	-0.104	-0.044	-0.067	-0.005	-0.022	-0.082	-0.077	-0.037	-0.073
TX3F	0.040	0.045	0.063	0.051	0.034	0.069	0.048	0.090	0.076
TX3H	0.063	0.090	0.104	0.045	0.041	0.109	0.091	0.105	0.089
TX4A_1	0.120	0.140	0.112	0.043	0.049	0.184	0.129	**0.332**	**0.316**
TX4A_2	0.082	0.058	0.079	0.024	0.028	0.106	0.074	0.177	0.156
TX4A_3	0.037	0.037	0.047	0.006	0.027	0.080	0.030	0.109	0.100
TX4A_4	0.008	0.009	0.040	0.004	-0.001	0.046	-0.007	0.025	0.006
TX4B	-0.024	-0.005	-0.017	-0.008	0.001	-0.014	0.002	0.002	0.010
TX4C	0.096	0.079	0.121	0.031	0.027	0.125	0.076	**0.210**	0.158
TX4D	0.137	0.109	0.126	0.019	0.051	**0.208**	0.116	**0.313**	**0.329**
TX5A_1	0.019	0.010	0.028	0.008	0.002	0.027	0.020	0.033	0.031
TX5A_2	0.000	0.009	0.001	-0.003	0.003	0.026	0.031	0.013	0.011
TX5B	0.061	0.039	0.055	-0.006	0.013	0.093	0.034	0.081	0.074
TX6A	0.032	0.015	0.016	0.006	0.011	0.031	0.009	0.058	0.053
TX6B	0.081	0.059	0.091	0.018	0.022	0.107	0.089	0.167	0.140
TX7A	0.072	0.034	0.061	0.016	-0.001	0.090	0.053	0.080	0.062
TX8	0.043	0.048	0.093	0.030	0.020	0.125	0.044	0.121	0.085
TX9A	0.097	0.140	0.133	0.056	0.055	0.188	0.097	**0.289**	**0.228**
TX9B	0.030	0.031	0.057	0.035	0.002	0.115	0.048	0.112	0.094
TX9C	0.036	0.043	0.066	0.040	0.033	0.078	0.079	0.084	0.047
TX9D	0.038	0.013	0.042	0.012	0.009	0.039	0.063	0.029	0.030
TX10C_1	0.032	0.065	0.074	0.005	0.034	0.099	0.046	0.131	0.082
TX10C_23	0.024	0.054	0.075	0.007	-0.002	0.065	0.074	0.077	0.048
TX11B	0.033	0.035	0.061	0.024	0.022	0.072	0.048	0.129	0.096
TX12A	0.079	0.048	0.055	0.011	0.000	0.072	0.067	0.112	0.100
TX12B	0.041	0.033	0.040	0.041	0.009	0.044	0.060	0.079	0.065
TX12C_1TO9	0.006	0.011	0.008	0.013	-0.006	0.001	-0.003	0.021	0.030
TX12C_10	0.085	0.121	0.086	0.020	0.050	0.159	**0.208**	**0.361**	**0.341**
TX13B	0.118	0.070	0.121	0.019	0.043	0.142	0.098	0.175	0.141
TX14B	0.109	0.062	0.072	0.026	0.036	0.064	0.073	0.043	0.032
TX15A	0.336	0.067	0.111	0.008	0.009	0.141	0.107	0.129	0.124
TX16A	0.041	0.221	0.093	0.039	0.037	0.093	0.059	0.149	0.131
TX17A	0.107	0.072	0.361	0.031	0.060	0.138	0.121	0.148	0.092
TX18A	0.021	0.034	0.068	0.049	-0.004	0.040	0.036	0.043	0.031
TX19A	0.132	0.121	0.175	0.043	0.040	0.375	0.137	**0.234**	0.178
TX19B	0.119	0.099	0.125	0.052	0.034	0.130	0.536	0.154	0.131

Table C.4: Bivariate cross-interview correlations (4)

Appendix D

Supplemental analyses for Chapter 4

D.1 Probability proportional-to-size

	Before Constraint						After Constraint					
Expenditure	Min	Q1	Median	Mean	Q3	Max	Min	Q1	Median	Mean	Q3	Max
TX2	0.0000	0.0000	0.0000	0.1123	0.1310	0.9996	0.0050	0.0050	0.0050	0.1159	0.1310	0.9996
TX3F	0.0000	0.0000	0.0000	0.0141	0.0000	0.9007	0.0050	0.0050	0.0050	0.0189	0.0050	0.9007
TX3H	0.0000	0.0000	0.0000	0.0085	0.0000	0.7628	0.0050	0.0050	0.0050	0.0132	0.0050	0.7628
TX4A_1	0.0000	0.0199	0.0421	0.0600	0.0778	0.9901	0.0050	0.0199	0.0421	0.0604	0.0778	0.9901
TX4A_2	0.0000	0.0000	0.0000	0.0046	0.0041	0.2899	0.0050	0.0050	0.0050	0.0082	0.0050	0.2899
TX4A_3	0.0000	0.0000	0.0000	0.0034	0.0000	0.2180	0.0050	0.0050	0.0050	0.0078	0.0050	0.2180
TX4A_4	0.0000	0.0000	0.0000	0.0000	0.0000	0.0548	0.0050	0.0050	0.0050	0.0050	0.0050	0.0548
TX4B	0.0000	0.0000	0.0000	0.0016	0.0000	0.5000	0.0050	0.0050	0.0050	0.0063	0.0050	0.5000
TX4C	0.0000	0.0000	0.0145	0.0284	0.0395	0.9821	0.0050	0.0050	0.0145	0.0303	0.0395	0.9821
TX4D	0.0000	0.0357	0.0817	0.1207	0.1602	0.9977	0.0050	0.0357	0.0817	0.1212	0.1602	0.9977
TX5A_1	0.0000	0.0000	0.0000	0.0024	0.0000	0.7810	0.0050	0.0050	0.0050	0.0072	0.0050	0.7810
TX5A_2	0.0000	0.0000	0.0000	0.0004	0.0000	0.5094	0.0050	0.0050	0.0050	0.0054	0.0050	0.5094
TX5B	0.0000	0.0000	0.0000	0.0324	0.0000	0.9818	0.0050	0.0050	0.0050	0.0368	0.0050	0.9818
TX6A	0.0000	0.0000	0.0000	0.0076	0.0000	0.8825	0.0050	0.0050	0.0050	0.0124	0.0050	0.8825
TX6B	0.0000	0.0000	0.0000	0.0288	0.0166	0.9606	0.0050	0.0050	0.0050	0.0322	0.0166	0.9606
TX7A	0.0000	0.0000	0.0000	0.0030	0.0000	0.7021	0.0050	0.0050	0.0050	0.0076	0.0050	0.7021
TX8	0.0000	0.0000	0.0000	0.0212	0.0044	0.9967	0.0050	0.0050	0.0050	0.0249	0.0050	0.9967
TX9A	0.0000	0.0000	0.0089	0.0421	0.0565	0.9868	0.0050	0.0050	0.0089	0.0444	0.0565	0.9868
TX9B	0.0000	0.0000	0.0000	0.0063	0.0000	0.5525	0.0050	0.0050	0.0050	0.0105	0.0050	0.5525
TX9C	0.0000	0.0000	0.0000	0.0006	0.0000	0.1520	0.0050	0.0050	0.0050	0.0054	0.0050	0.1520
TX9D	0.0000	0.0000	0.0000	0.0009	0.0000	0.4073	0.0050	0.0050	0.0050	0.0057	0.0050	0.4073
TX10C_1	0.0000	0.0000	0.0000	0.0047	0.0000	0.6516	0.0050	0.0050	0.0050	0.0095	0.0050	0.6516
TX10C_23	0.0000	0.0000	0.0000	0.0058	0.0000	0.8117	0.0050	0.0050	0.0050	0.0108	0.0050	0.8117
TX11B	0.0000	0.0000	0.0000	0.0846	0.0000	0.9728	0.0050	0.0050	0.0050	0.0888	0.0050	0.9728
TX12A	0.0000	0.0000	0.0000	0.0312	0.0165	0.7891	0.0050	0.0050	0.0050	0.0345	0.0165	0.7891
TX12B	0.0000	0.0000	0.0000	0.0055	0.0000	0.4785	0.0050	0.0050	0.0050	0.0098	0.0050	0.4785
TX12C_1TO9	0.0000	0.0000	0.0000	0.0001	0.0000	0.0935	0.0050	0.0050	0.0050	0.0051	0.0050	0.0935
TX12C_10	0.0000	0.0279	0.0630	0.0929	0.1245	0.9979	0.0050	0.0279	0.0630	0.0934	0.1245	0.9979
TX13B	0.0000	0.0000	0.0101	0.0663	0.0833	0.8997	0.0050	0.0050	0.0101	0.0687	0.0833	0.8997
TX14B	0.0000	0.0000	0.0000	0.0195	0.0000	0.8433	0.0050	0.0050	0.0050	0.0236	0.0050	0.8433
TX15A	0.0000	0.0000	0.0000	0.0402	0.0405	0.9827	0.0050	0.0050	0.0050	0.0429	0.0405	0.9827
TX16A	0.0000	0.0000	0.0000	0.0266	0.0000	0.9585	0.0050	0.0050	0.0050	0.0307	0.0050	0.9585
TX17A	0.0000	0.0000	0.0000	0.0081	0.0009	0.7605	0.0050	0.0050	0.0050	0.0119	0.0050	0.7605
TX18A	0.0000	0.0000	0.0000	0.0018	0.0000	0.7632	0.0050	0.0050	0.0050	0.0067	0.0050	0.7632
TX19A	0.0000	0.0000	0.0104	0.0432	0.0502	0.9524	0.0050	0.0050	0.0104	0.0453	0.0502	0.9524
TX19B	0.0000	0.0000	0.0004	0.0649	0.0704	0.9259	0.0050	0.0050	0.0050	0.0675	0.0704	0.9259

Table D.1: PPS propensity summaries before and after constraint

D.2 Logistic regression methods

	Before Constraint						After Constraint					
Expenditure	Min	Q1	Median	Mean	Q3	Max	Min	Q1	Median	Mean	Q3	Max
TX2	0.0005	0.0029	0.0047	0.2783	0.8615	0.9725	0.0050	0.0050	0.0050	0.2794	0.8615	0.9725
TX3F	0.0006	0.0042	0.0454	0.0451	0.0678	0.1705	0.0050	0.0050	0.0454	0.0457	0.0678	0.1705
TX3H	0.0001	0.0024	0.0676	0.0576	0.0969	0.1613	0.0050	0.0050	0.0676	0.0586	0.0969	0.1613
TX4A_1	0.6219	0.8920	0.9419	0.9153	0.9619	0.9869	0.6219	0.8920	0.9419	0.9153	0.9619	0.9869
TX4A_2	0.0523	0.2107	0.2860	0.2880	0.3630	0.6106	0.0523	0.2107	0.2860	0.2880	0.3630	0.6106
TX4A_3	0.0207	0.0823	0.1157	0.1300	0.1605	0.4504	0.0207	0.0823	0.1157	0.1300	0.1605	0.4504
TX4A_4	0.0000	0.0012	0.0032	0.0040	0.0053	0.0305	0.0050	0.0050	0.0050	0.0061	0.0053	0.0305
TX4B	0.0174	0.0435	0.0586	0.0694	0.0839	0.3087	0.0174	0.0435	0.0586	0.0694	0.0839	0.3087
TX4C	0.3468	0.5902	0.6432	0.6321	0.6939	0.7319	0.3468	0.5902	0.6432	0.6321	0.6939	0.7319
TX4D	0.4899	0.8810	0.9628	0.9165	0.9751	0.9900	0.4899	0.8810	0.9628	0.9165	0.9751	0.9900
TX5A_1	0.0020	0.0097	0.0249	0.0242	0.0365	0.0550	0.0050	0.0097	0.0249	0.0243	0.0365	0.0550
TX5A_2	0.0004	0.0036	0.0067	0.0090	0.0122	0.0324	0.0050	0.0050	0.0067	0.0097	0.0122	0.0324
TX5B	0.0176	0.0351	0.1581	0.1309	0.1822	0.2535	0.0176	0.0351	0.1581	0.1309	0.1822	0.2535
TX6A	0.0181	0.0309	0.0399	0.0423	0.0499	0.0844	0.0181	0.0309	0.0399	0.0423	0.0499	0.0844
TX6B	0.0977	0.2596	0.3263	0.3303	0.3955	0.5585	0.0977	0.2596	0.3263	0.3303	0.3955	0.5585
TX7A	0.0074	0.0242	0.0685	0.0654	0.0898	0.1661	0.0074	0.0242	0.0685	0.0654	0.0898	0.1661
TX8	0.1481	0.2298	0.2729	0.2809	0.3216	0.4813	0.1481	0.2298	0.2729	0.2809	0.3216	0.4813
TX9A	0.2897	0.4655	0.5480	0.5449	0.6235	0.7863	0.2897	0.4655	0.5480	0.5449	0.6235	0.7863
TX9B	0.0524	0.1096	0.1537	0.1681	0.2201	0.4036	0.0524	0.1096	0.1537	0.1681	0.2201	0.4036
TX9C	0.0078	0.0295	0.0409	0.0423	0.0567	0.0740	0.0078	0.0295	0.0409	0.0423	0.0567	0.0740
TX9D	0.0066	0.0264	0.0391	0.0451	0.0576	0.1558	0.0066	0.0264	0.0391	0.0451	0.0576	0.1558
TX10C_1	0.0011	0.0152	0.0278	0.0373	0.0462	0.1987	0.0050	0.0152	0.0278	0.0373	0.0462	0.1987
TX10C_23	0.0019	0.0084	0.0123	0.0151	0.0180	0.0709	0.0050	0.0084	0.0123	0.0151	0.0180	0.0709
TX11B	0.0361	0.1198	0.1570	0.1697	0.2127	0.4871	0.0361	0.1198	0.1570	0.1697	0.2127	0.4871
TX12A	0.1456	0.2977	0.3547	0.3525	0.4012	0.5219	0.1456	0.2977	0.3547	0.3525	0.4012	0.5219
TX12B	0.0405	0.1151	0.1502	0.1516	0.1843	0.2882	0.0405	0.1151	0.1502	0.1516	0.1843	0.2882
TX12C_1TO9	0.1538	0.8624	0.9441	0.8917	0.9679	0.9911	0.1538	0.8624	0.9441	0.8917	0.9679	0.9911
TX12C_10	0.0014	0.0068	0.0109	0.0124	0.0154	0.0818	0.0050	0.0068	0.0109	0.0126	0.0154	0.0818
TX13B	0.2070	0.4496	0.5445	0.5272	0.6086	0.7460	0.2070	0.4496	0.5445	0.5272	0.6086	0.7460
TX14B	0.0401	0.1451	0.1985	0.1983	0.2631	0.2925	0.0401	0.1451	0.1985	0.1983	0.2631	0.2925
TX15A	0.1612	0.3969	0.5015	0.4814	0.5846	0.6756	0.1612	0.3969	0.5015	0.4814	0.5846	0.6756
TX16A	0.0230	0.0889	0.1491	0.1808	0.2511	0.5112	0.0230	0.0889	0.1491	0.1808	0.2511	0.5112
TX17A	0.0251	0.1791	0.2552	0.2545	0.3444	0.4203	0.0251	0.1791	0.2552	0.2545	0.3444	0.4203
TX18A	0.0026	0.0125	0.0166	0.0193	0.0236	0.0523	0.0050	0.0125	0.0166	0.0193	0.0236	0.0523
TX19A	0.1719	0.5189	0.6296	0.5980	0.6947	0.7993	0.1719	0.5189	0.6296	0.5980	0.6947	0.7993
TX19B	0.2445	0.4207	0.5315	0.5004	0.5770	0.6436	0.2445	0.4207	0.5315	0.5004	0.5770	0.6436

Table D.2: Log1 propensity summaries before and after constraint

Expenditure	Intercept	Poverty	Region	Urbanicity	Tenure	Race	CU Size	Indicator
TX2	-15.127	0.372	0.009	-1.003	7.543	-0.149	-0.102	4.455
TX3F	2.582	0.070	-0.028	-0.165	-8.699	0.582	0.174	8.821
TX3H	-5.625	2.420	-0.044	-0.186	-9.781	3.563	0.116	6.437
TX4A_1	0.047	0.326	-0.024	-0.035	-0.674	0.027	0.308	4.048
TX4A_2	-3.094	0.399	-0.014	-0.234	-0.318	0.278	0.211	3.317
TX4A_3	-2.899	0.472	-0.034	-0.423	-0.466	0.130	0.233	4.222
TX4A_4	-7.428	5.018	-0.885	-9.038	-7.682	5.378	0.055	5.703
TX4B	-3.169	-0.474	-0.043	-0.032	0.231	0.783	0.188	2.985
TX4C	-2.421	0.186	0.014	0.192	-0.084	0.147	-0.076	3.952
TX4D	-1.203	0.522	0.013	0.347	-0.736	-0.237	0.313	3.961
TX5A_1	-9.494	2.189	-0.026	0.019	-4.589	2.945	0.223	-1.318
TX5A_2	-34.092	8.997	-0.012	-2.896	-6.936	10.833	-0.022	-0.539
TX5B	-0.480	0.271	0.005	0.042	-2.013	0.220	-0.004	1.173
TX6A	-3.444	-0.161	0.007	0.253	-0.562	0.316	0.218	0.340
TX6B	-2.297	0.425	-0.002	-0.075	-0.161	0.353	0.262	1.044
TX7A	-2.769	0.273	0.022	0.035	-1.219	0.168	0.125	1.685
TX8	-1.915	0.293	-0.011	0.033	-0.172	0.275	0.199	1.103
TX9A	-0.978	0.292	-0.015	-0.149	0.003	0.294	0.298	1.069
TX9B	-3.225	0.197	0.003	-0.096	0.075	0.131	0.389	1.710
TX9C	-2.709	0.528	-0.016	-0.499	-0.683	0.422	-0.002	1.508
TX9D	-8.973	0.433	0.009	0.168	-0.440	2.443	0.184	2.396
TX10C_1	-43.594	8.241	0.423	-14.272	-5.019	9.361	0.677	63.724
TX10C_23	-39.615	10.249	-0.018	-14.935	-9.052	11.828	1.654	67.133
TX11B	-26.747	6.785	0.079	-3.906	-1.812	5.261	0.521	42.420
TX12A	-1.947	0.482	0.016	0.072	-0.466	0.303	0.231	0.764
TX12B	-2.544	0.387	0.008	0.167	-0.402	0.287	0.171	0.147
TX12C_1TO9	-3.132	0.402	0.054	0.025	-0.618	0.136	0.324	4.359
TX12C_10	-0.188	0.373	-0.054	-0.498	-0.447	0.408	0.140	0.870
TX13B	-0.813	0.373	0.001	0.185	-0.477	-0.072	0.183	1.927
TX14B	-1.793	0.457	-0.033	-0.082	-0.412	0.352	-0.026	3.094
TX15A	-0.558	0.485	-0.015	0.057	-0.709	0.292	0.050	1.832
TX16A	-4.432	0.560	-0.005	-0.224	-0.116	0.226	0.693	2.388
TX17A	-1.182	0.664	-0.028	-0.262	-0.848	0.494	-0.059	1.913
TX18A	-10.695	1.401	0.049	-0.895	0.089	1.885	0.035	2.254
TX19A	-1.710	0.400	-0.020	0.058	-0.509	0.737	0.183	2.102
TX19B	0.063	0.274	-0.013	-0.073	-0.499	-0.017	0.028	2.199

Table D.3: Parameter estimates for Log2 propensity model

	Before Constraint						After Constraint					
Expenditure	Min	Q1	Median	Mean	Q3	Max	Min	Q1	Median	Mean	Q3	Max
TX2	0.0001	0.0003	0.0003	0.2795	0.9637	0.9851	0.0050	0.0050	0.0050	0.2827	0.9637	0.9851
TX3F	0.0000	0.0000	0.0041	0.0457	0.0050	0.9811	0.0050	0.0050	0.0050	0.0477	0.0050	0.9811
TX3H	0.0000	0.0000	0.0085	0.0533	0.0110	0.9131	0.0050	0.0050	0.0085	0.0553	0.0110	0.9131
TX4A_1	0.1886	0.9677	0.9808	0.9266	0.9858	0.9920	0.1886	0.9677	0.9808	0.9266	0.9858	0.9920
TX4A_2	0.0234	0.0703	0.0943	0.2672	0.6529	0.8283	0.0234	0.0703	0.0943	0.2672	0.6529	0.8283
TX4A_3	0.0076	0.0273	0.0388	0.1293	0.0548	0.8638	0.0076	0.0273	0.0388	0.1293	0.0548	0.8638
TX4A_4	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0050	0.0050	0.0050	0.0050	0.0050	0.0050
TX4B	0.0138	0.0386	0.0474	0.0804	0.0589	0.7671	0.0138	0.0386	0.0474	0.0804	0.0589	0.7671
TX4C	0.1134	0.1962	0.9160	0.6518	0.9267	0.9466	0.1134	0.1962	0.9160	0.6518	0.9267	0.9466
TX4D	0.1551	0.9635	0.9796	0.9221	0.9854	0.9943	0.1551	0.9635	0.9796	0.9221	0.9854	0.9943
TX5A_1	0.0000	0.0002	0.0143	0.0110	0.0189	0.0310	0.0050	0.0050	0.0143	0.0129	0.0189	0.0310
TX5A_2	0.0000	0.0000	0.0005	0.0043	0.0094	0.0105	0.0050	0.0050	0.0050	0.0070	0.0094	0.0105
TX5B	0.0201	0.0343	0.2042	0.1793	0.2099	0.4734	0.0201	0.0343	0.2042	0.1793	0.2099	0.4734
TX6A	0.0189	0.0413	0.0563	0.0566	0.0700	0.1681	0.0189	0.0413	0.0563	0.0566	0.0700	0.1681
TX6B	0.1440	0.3189	0.3992	0.4390	0.5707	0.7461	0.1440	0.3189	0.3992	0.4390	0.5707	0.7461
TX7A	0.0158	0.0348	0.0924	0.0948	0.1065	0.4606	0.0158	0.0348	0.0924	0.0948	0.1065	0.4606
TX8	0.1420	0.2711	0.3204	0.3695	0.4737	0.6884	0.1420	0.2711	0.3204	0.3695	0.4737	0.6884
TX9A	0.2998	0.5429	0.7136	0.6691	0.7938	0.8798	0.2998	0.5429	0.7136	0.6691	0.7938	0.8798
TX9B	0.0738	0.1175	0.1582	0.2252	0.2242	0.7030	0.0738	0.1175	0.1582	0.2252	0.2242	0.7030
TX9C	0.0097	0.0421	0.0535	0.0654	0.0802	0.3036	0.0097	0.0421	0.0535	0.0654	0.0802	0.3036
TX9D	0.0016	0.0336	0.0511	0.0601	0.0624	0.5166	0.0050	0.0336	0.0511	0.0603	0.0624	0.5166
TX10C_1	0.0000	0.0000	0.0000	0.0373	0.0000	1.0000	0.0050	0.0050	0.0050	0.0421	0.0050	1.0000
TX10C_23	0.0000	0.0000	0.0000	0.0151	0.0000	1.0000	0.0050	0.0050	0.0050	0.0200	0.0050	1.0000
TX11B	0.0000	0.0001	0.0033	0.1725	0.0101	1.0000	0.0050	0.0050	0.0050	0.1749	0.0101	1.0000
TX12A	0.1884	0.4599	0.5450	0.5467	0.6572	0.8143	0.1884	0.4599	0.5450	0.5467	0.6572	0.8143
TX12B	0.1029	0.2230	0.2885	0.2768	0.3284	0.4566	0.1029	0.2230	0.2885	0.2768	0.3284	0.4566
TX12C_1TO9	0.0581	0.2082	0.2663	0.2694	0.3261	0.6408	0.0581	0.2082	0.2663	0.2694	0.3261	0.6408
TX12C_10	0.0881	0.9600	0.9751	0.8935	0.9819	0.9914	0.0881	0.9600	0.9751	0.8935	0.9819	0.9914
TX13B	0.2397	0.4445	0.7593	0.6493	0.8595	0.9195	0.2397	0.4445	0.7593	0.6493	0.8595	0.9195
TX14B	0.0457	0.1160	0.1557	0.2668	0.1867	0.8473	0.0457	0.1160	0.1557	0.2668	0.1867	0.8473
TX15A	0.1702	0.4356	0.5428	0.6171	0.8637	0.8919	0.1702	0.4356	0.5428	0.6171	0.8637	0.8919
TX16A	0.0221	0.0669	0.1237	0.2292	0.2247	0.8639	0.0221	0.0669	0.1237	0.2292	0.2247	0.8639
TX17A	0.0367	0.1613	0.2756	0.3327	0.3760	0.7841	0.0367	0.1613	0.2756	0.3327	0.3760	0.7841
TX18A	0.0004	0.0102	0.0244	0.0242	0.0310	0.2850	0.0050	0.0102	0.0244	0.0244	0.0310	0.2850
TX19A	0.1207	0.4608	0.8109	0.6758	0.8894	0.9300	0.1207	0.4608	0.8109	0.6758	0.8894	0.9300
TX19B	0.2259	0.4065	0.7379	0.6107	0.8668	0.8868	0.2259	0.4065	0.7379	0.6107	0.8668	0.8868

Table D.4: Log2 propensity summaries before and after constraint

D.3 Stratification methods

Expenditure	Stratum	N_{hk}	S_{hk}	c_h	$n_{k,min}$	n_{hk}	f_{hk}	Have
TX2	Low	5,258	0.00	9	400	30	0.0057	11
	High	5,237	521.27	1		4,238	0.8093	2,930
TX3F	Low	5,322	0.00	9	400	30	0.0056	16
	High	5,173	593.10	1		4,422	0.8549	475
TX3H	Low	5,353	16.71	9	400	30	0.0056	22
	High	5,142	286.63	1		1,237	0.2406	567
TX4A_1	Low	5,315	75.27	9	400	343	0.0646	4,619
	High	5,180	83.71	1		669	0.1291	5,091
TX4A_2	Low	5,311	0.00	9	400	30	0.0056	331
	High	5,184	19.69	1		3,482	0.6718	2,480
TX4A_3	Low	5,252	0.00	9	400	30	0.0057	137
	High	5,243	29.56	1		3,484	0.6645	1,230
TX4A_4	Low	5,250	0.32	9	400	755	0.1438	12
	High	5,245	3.19	1		33	0.0063	12
TX4B	Low	5,329	0.00	9	400	30	0.0056	221
	High	5,166	17.40	1		3,482	0.6741	645
TX4C	Low	5,258	40.45	9	400	762	0.1448	1,971
	High	5,237	39.96	1		208	0.0397	4,868
TX4D	Low	5,252	154.35	9	400	845	0.1609	4,539
	High	5,243	164.91	1		452	0.0863	5,123
TX5A_1	Low	5,340	162.68	9	400	3,751	0.7024	72
	High	5,155	0.00	1		30	0.0058	113
TX5A_2	Low	5,264	12.51	9	400	756	0.1436	28
	High	5,231	38.09	1		203	0.0388	70
TX5B	Low	5,288	259.06	9	400	961	0.1817	500
	High	5,207	2,527.69	1		5,207	1.0000	1,405
TX6A	Low	5,299	80.50	9	400	781	0.1473	240
	High	5,196	262.39	1		692	0.1331	394
TX6B	Low	5,273	20.56	9	400	757	0.1435	1,634
	High	5,222	399.98	1		1,105	0.2116	2,975
TX7A	Low	5,353	1.53	9	400	755	0.1411	283
	High	5,142	72.70	1		253	0.0493	744
TX8	Low	5,294	0.00	9	400	30	0.0057	1,394
	High	5,201	439.43	1		4,022	0.7734	2,493
TX9A	Low	5,255	74.20	9	400	776	0.1478	2,857
	High	5,240	235.27	1		629	0.1201	4,165

Table D.5: Two-Bin stratification classification and associated parameters

Expenditure	Stratum	N_{hk}	S_{hk}	c_h	$n_{k,min}$	n_{hk}	f_{hk}	Have
TX9B	Low	5,250	0.00	9	400	30	0.0057	692
	High	5,245	100.90	1		3,512	0.6696	1,691
TX9C	Low	5,299	0.00	9	400	30	0.0057	219
	High	5,196	17.60	1		3,482	0.6702	495
TX9D	Low	5,271	3.11	9	400	755	0.1433	178
	High	5,224	21.06	1		126	0.0241	494
TX10C_1	Low	5,249	0.00	9	400	30	0.0057	11
	High	5,246	156.27	1		3,554	0.6775	407
TX10C_23	Low	5,260	0.00	9	400	30	0.0057	3
	High	5,235	647.38	1		4,612	0.8809	173
TX11B	Low	5,325	0.00	9	400	30	0.0056	42
	High	5,170	2,925.51	1		5,170	1.0000	1,822
TX12A	Low	5,319	89.31	9	400	787	0.1479	2,348
	High	5,176	310.03	1		812	0.1568	3,393
TX12B	Low	5,302	35.18	9	400	760	0.1434	1,187
	High	5,193	55.93	1		235	0.0452	1,741
TX12C_10	Low	5,301	140.37	9	400	831	0.1568	4,293
	High	5,194	170.71	1		463	0.0891	5,072
TX12C_1TO9	Low	5,249	0.96	9	400	755	0.1439	1,083
	High	5,246	3.86	1		34	0.0065	1,769
TX13B	Low	5,252	65.38	9	400	772	0.1469	2,372
	High	5,243	436.81	1		1,257	0.2397	4,443
TX14B	Low	5,255	0.00	9	400	30	0.0057	589
	High	5,240	253.58	1		3,671	0.7005	2,225
TX15A	Low	5,253	0.00	9	400	30	0.0057	2,164
	High	5,242	337.68	1		3,813	0.7273	4,299
TX16A	Low	5,287	0.00	9	400	30	0.0057	371
	High	5,208	822.24	1		5,208	1.0000	2,054
TX17A	Low	5,322	1.24	9	400	755	0.1419	901
	High	5,173	113.15	1		323	0.0624	2,601
TX18A	Low	5,349	6.92	9	400	755	0.1412	101
	High	5,146	131.87	1		364	0.0706	215
TX19A	Low	5,249	129.78	9	400	819	0.1561	2,460
	High	5,246	341.71	1		915	0.1744	4,625
TX19B	Low	5,248	23.23	9	400	757	0.1443	1,952
	High	5,247	522.97	1		3,649	0.6954	4,448

Table D.6: Two-Bin stratification classification and associated parameters (2)

Expenditure	Stratum	N_{hk}	S_{hk}	c_h	$n_{k,min}$	n_{hk}	f_{hk}	Have
	Lowest	2102	0.00	9		30	0.0143	3
	Low	2155	0.00	7		30	0.0139	3
TX2	Medium	2049	0.00	5	400	30	0.0146	8
	High	2107	401.07	3		1,789	0.8490	900
	Highest	2082	492.07	1		2,082	1.0000	2,027
	Lowest	2107	0.00	9		30	0.0142	2
	Low	2099	0.00	7		30	0.0143	7
TX3F	Medium	2127	16.46	5	400	118	0.0553	14
	High	2113	62.55	3		228	0.1079	21
	Highest	2049	900.77	1		2,049	1.0000	447
	Lowest	2108	0.00	9		30	0.0142	1
	Low	2099	26.67	7		32	0.0150	9
TX3H	Medium	2097	0.00	5	400	30	0.0143	28
	High	2175	0.00	3		30	0.0138	21
	Highest	2016	438.02	1		1,964	0.9744	530
	Lowest	2102	60.74	9		80	0.0378	1,502
	Low	2057	73.95	7		108	0.0526	1,991
TX4A_1	Medium	2215	74.04	5	400	110	0.0496	2,162
	High	2080	80.98	3		109	0.0522	2,055
	Highest	2041	87.22	1		281	0.1379	2,000
	Lowest	2134	0.00	9		30	0.0141	84
	Low	2119	0.00	7		30	0.0142	138
TX4A_2	Medium	2057	0.00	5	400	30	0.0146	233
	High	2087	18.78	3		385	0.1845	807
	Highest	2098	13.09	1		63	0.0302	1,549
	Lowest	2099	0.00	9		30	0.0143	34
	Low	2129	0.00	7		30	0.0141	69
TX4A_3	Medium	2087	0.00	5	400	30	0.0144	77
	High	2111	0.00	3		30	0.0142	123
	Highest	2069	35.69	1		364	0.1762	1,064
	Lowest	2113	0.22	9		47	0.0222	5
	Low	2098	0.00	7		30	0.0143	3
TX4A_4	Medium	2097	0.89	5	400	56	0.0268	4
	High	2159	2.91	3		52	0.0241	5
	Highest	2028	4.08	1		227	0.1120	7

Table D.7: Five-Bin stratification classification and associated parameters

Expenditure	Stratum	N_{hk}	S_{hk}	c_h	$n_{k,min}$	n_{hk}	f_{hk}	Have
	Lowest	2180	0.00	9		30	0.0138	69
	Low	2076	0.00	7		30	0.0145	103
TX4B	Medium	2064	0.00	5	400	30	0.0145	87
	High	2090	0.00	3		30	0.0144	101
	Highest	2085	25.70	1		359	0.1722	506
	Lowest	2101	0.00	9		30	0.0143	361
	Low	2099	30.55	7		139	0.0663	670
TX4C	Medium	2060	42.72	5	400	272	0.1320	1,854
	High	2166	40.91	3		124	0.0570	2,018
	Highest	2069	37.95	1		590	0.2853	1,936
	Lowest	2103	97.92	9		212	0.1009	1,471
	Low	2128	135.24	7		337	0.1584	2,067
TX4D	Medium	2087	161.70	5	400	471	0.2256	2,035
	High	2087	166.41	3		467	0.2240	2,034
	Highest	2090	169.49	1		591	0.2828	2,055
	Lowest	2132	71.59	9		161	0.0757	7
	Low	2079	109.32	7		241	0.1158	34
TX5A_1	Medium	2152	220.31	5	400	536	0.2491	44
	High	2146	0.00	3		30	0.0140	45
	Highest	1986	0.00	1		30	0.0151	55
	Lowest	2107	10.05	9		33	0.0156	7
	Low	2092	9.54	7		57	0.0272	11
TX5A_2	Medium	2099	61.61	5	400	274	0.1303	32
	High	2235	0.00	3		30	0.0134	26
	Highest	1962	0.00	1		30	0.0153	22
	Lowest	2138	0.00	9		30	0.0140	48
	Low	2110	409.11	7		907	0.4301	252
TX5B	Medium	2122	0.00	5	400	30	0.0141	446
	High	2010	0.00	3		30	0.0149	412
	Highest	2115	3,848.30	1		2,115	1.0000	747
	Lowest	2104	26.30	9		78	0.0371	54
	Low	2096	121.30	7		265	0.1263	113
TX6A	Medium	2025	104.64	5	400	249	0.1228	123
	High	2241	132.07	3		369	0.1646	162
	Highest	2029	376.54	1		1,258	0.6199	182

Table D.8: Five-Bin stratification classification and associated parameters (2)

Expenditure	Stratum	N_{hk}	S_{hk}	c_h	$n_{k,min}$	n_{hk}	f_{hk}	Have
	Lowest	2116	0.00	9		30	0.0142	526
	Low	2119	32.01	7		96	0.0454	704
TX6B	Medium	2112	56.13	5	400	157	0.0743	907
	High	2054	365.49	3		1,045	0.5086	1,069
	Highest	2094	473.70	1		1,759	0.8399	1,403
	Lowest	2109	0.00	9		30	0.0142	57
	Low	2108	0.44	7		41	0.0192	117
TX7A	Medium	2229	2.37	5	400	55	0.0247	203
	High	1956	1.80	3		69	0.0354	233
	Highest	2093	108.53	1		320	0.1528	417
	Lowest	2099	0.00	9		30	0.0143	458
	Low	2121	0.00	7		30	0.0141	572
TX8	Medium	2112	36.46	5	400	115	0.0545	720
	High	2079	324.67	3		920	0.4424	892
	Highest	2084	576.58	1		2,084	1.0000	1,245
	Lowest	2103	0.00	9		30	0.0143	959
	Low	2110	45.19	7		123	0.0581	1,197
TX9A	Medium	2103	184.74	5	400	435	0.2068	1,479
	High	2095	224.79	3		601	0.2868	1,638
	Highest	2084	252.89	1		778	0.3733	1,749
	Lowest	2116	0.00	9		30	0.0142	207
	Low	2085	0.00	7		30	0.0144	289
TX9B	Medium	2113	0.00	5	400	30	0.0142	376
	High	2084	0.00	3		30	0.0144	476
	Highest	2097	142.65	1		543	0.2587	1,035
	Lowest	2080	0.00	9		30	0.0144	66
	Low	2150	0.00	7		30	0.0140	99
TX9C	Medium	2076	2.93	5	400	267	0.1288	123
	High	2169	1.50	3		47	0.0219	180
	Highest	2020	27.14	1		181	0.0896	246
	Lowest	2146	1.41	9		46	0.0217	45
	Low	2119	4.69	7		62	0.0292	91
TX9D	Medium	2130	0.26	5	400	56	0.0264	106
	High	2092	1.35	3		62	0.0295	131
	Highest	2008	32.92	1		242	0.1208	299

Table D.9: Five-Bin stratification classification and associated parameters (3)

Expenditure	Stratum	N_{hk}	S_{hk}	c_h	$n_{k,min}$	n_{hk}	f_{hk}	Have
	Lowest	2088	0.00	9		30	0.0144	5
	Low	2112	0.00	7		30	0.0142	2
TX10C_1	Medium	2180	0.00	5	400	30	0.0138	7
	High	2094	0.00	3		30	0.0143	7
	Highest	2021	240.06	1		792	0.3919	397
	Lowest	2123	0.00	9		30	0.0141	1
	Low	2092	0.00	7		30	0.0143	0
TX10C_23	Medium	2103	0.00	5	400	30	0.0143	3
	High	2084	0.00	3		30	0.0144	4
	Highest	2093	1,010.38	1		2,093	1.0000	168
	Lowest	2115	0.00	9		30	0.0142	15
	Low	2074	0.00	7		30	0.0145	22
TX11B	Medium	2138	0.00	5	400	30	0.0140	13
	High	2142	0.00	3		30	0.0140	27
	Highest	2026	3,865.26	1		2,026	1.0000	1,787
	Lowest	2100	57.17	9		308	0.1467	682
	Low	2098	94.53	7		437	0.2082	1,035
TX12A	Medium	2101	139.60	5	400	784	0.3731	1,198
	High	2097	250.32	3		1,803	0.8599	1,321
	Highest	2099	372.12	1		1,145	0.5455	1,505
	Lowest	2100	22.97	9		192	0.0915	332
	Low	2108	39.39	7		279	0.1322	569
TX12B	Medium	2184	39.20	5	400	268	0.1226	628
	High	2053	46.99	3		278	0.1353	697
	Highest	2050	69.79	1		331	0.1617	702
	Lowest	2118	110.26	9		46	0.0216	1,202
	Low	2151	136.66	7		61	0.0284	2,084
TX12C_10	Medium	2026	148.86	5	400	69	0.0339	1,975
	High	2131	161.66	3		74	0.0349	2,080
	Highest	2069	178.16	1		218	0.1052	2,024
	Lowest	2105	0.07	9		114	0.0544	365
	Low	2103	0.15	7		169	0.0804	460
TX12C_1TO9	Medium	2214	1.48	5	400	169	0.0763	606
	High	2099	1.25	3		199	0.0949	624
	Highest	1974	6.10	1		386	0.1956	797

Table D.10: Five-Bin stratification classification and associated parameters (4)

Expenditure	Stratum	N_{hk}	S_{hk}	c_h	$n_{k,min}$	n_{hk}	f_{hk}	Have
	Lowest	2105	0.00	9		30	0.0143	653
	Low	2105	0.00	7		30	0.0143	1,068
TX13B	Medium	2097	255.17	5	400	625	0.2979	1,533
	High	2092	477.05	3		1,100	0.5260	1,715
	Highest	2096	439.94	1		1,328	0.6337	1,846
	Lowest	2111	0.00	9		30	0.0142	165
	Low	2181	0.00	7		30	0.0138	280
TX14B	Medium	2011	0.00	5	400	30	0.0149	324
	High	2057	0.00	3		30	0.0146	403
	Highest	2135	344.43	1		1,027	0.4813	1,642
	Lowest	2099	0.00	9		30	0.0143	589
	Low	2123	0.00	7		30	0.0141	991
TX15A	Medium	2089	193.08	5	400	471	0.2254	1,330
	High	2105	348.58	3		712	0.3382	1,767
	Highest	2079	354.02	1		763	0.3668	1,786
	Lowest	2136	0.00	9		30	0.0140	129
	Low	2068	0.00	7		30	0.0145	133
TX16A	Medium	2093	0.00	5	400	30	0.0143	232
	High	2091	282.28	3		565	0.2702	567
	Highest	2107	1,171.18	1		2,107	1.0000	1,364
	Lowest	2153	0.00	9		30	0.0139	229
	Low	2047	0.02	7		47	0.0227	376
TX17A	Medium	2123	4.05	5	400	42	0.0198	588
	High	2074	62.61	3		254	0.1224	826
	Highest	2098	152.40	1		190	0.0905	1,483
	Lowest	2140	0.00	9		30	0.0140	36
	Low	2153	10.91	7		30	0.0139	43
TX18A	Medium	2024	0.00	5	400	267	0.1322	39
	High	2099	3.31	3		42	0.0200	85
	Highest	2079	205.89	1		316	0.1521	113
	Lowest	2127	0.00	9		30	0.0141	634
	Low	2071	30.39	7		51	0.0249	1,039
TX19A	Medium	2112	279.07	5	400	662	0.3135	1,693
	High	2116	330.84	3		893	0.4219	1,850
	Highest	2069	368.66	1		1,115	0.5388	1,869
	Lowest	2102	0.00	9		30	0.0143	632
	Low	2098	0.00	7		30	0.0143	840
TX19B	Medium	2098	345.16	5	400	921	0.4391	1,300
	High	2102	522.80	3		1,830	0.8704	1,805
	Highest	2095	558.53	1		2,095	1.0000	1,823

Table D.11: Five-Bin stratification classification and associated parameters (5)

Bibliography

[1] Adams, L. M. and Darwin, G. (1982). Solving the Quandary Between Questionnaire Length and Response Rate in Educational Research. *Research in Higher Education*, **17**, 231–40.

[2] Adiguzel, F. and Wedel, M. (2008). Split Questionnaire Design for Massive Surveys. *Journal of Marketing Research*, **25**(5), 608–17.

[3] Agresti, A. (2002). *Categorical Data Analysis*. New York: Wiley-Interscience.

[4] Askegaard, L. D. and Umila, B. V. (1982). An Empirical Investigation of the Applicability of Multiple Matrix Sampling to the Method of Rank Order. *Journal of Educational Measurement*, **19**(3), 193–7.

[5] Berger, J. O. (1980). *Statistical Decision Theory: Foundations, Concepts, and Methods*. New York: Springer-Verlag.

[6] Biggerstaff, B. J. (2000). Comparing Diagnostic Tests: A Simple Graphic Using Likelihood Ratios. *Statistics in Medicine*, **19**(5), 649–63.

[7] Bogen, K. (1996). The Effect of Questionnaire Length on Response Rates – A Review of the Literature. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 1020–25.

[8] Brackstone, G. (1999). Managing Data Quality in a Statistical Agency. *Survey Methodology*, **25**, 139–49.

[9] Bradburn, N. M. (1978). Respondent Burden. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 35–40.

[10] Breiman, L., Friedman, J., Stone, C., and Olshen, R. (1993). *Classification and Regression Trees*. London: Chapman and Hall.

[11] Bureau of Labor Statistics, U.S. Department of Labor, *Handbook of Methods*, Chapter 16, April 2007 edition, Consumer Expenditures and Income. http://www.bls.gov/opub/hom/pdf/homch16.pdf (visited September 10, 2009).

[12] Bureau of Labor Statistics, U.S. Department of Labor, *Current Standard Error Tables*. http://www.bls.gov/cex/#tables (visited June 20, 2012).

[13] Bureau of Labor Statistics, Consumer Expenditure Survey (2010). *Consumer Expenditure Surveys Quarterly Interview CAPI Survey (2010).* http://www.bls.gov/cex/capi/2010/cecapihome.htm (visited January 27, 2011).

[14] Casella, G. and Berger, R. L. (2002). *Statistical Inference*, 2^{nd} edition. Duxbury.

[15] Chipperfield, J. O. and Steel, D. G., (2009). Design and Estimation for Split Questionnaire Surveys. *Journal of Official Statistics*, **25**(2), 227-44.

[16] Cochran, W. (1977). *Sampling Techniques*, 3^{rd} edition. Wiley: New York.

[17] Conrad, F. G. and Schober M. F. (2000). Clarifying Question Meaning in a Household Telephone Survey. *Public Opinion Quarterly*, 64, 1–28.

[18] Cornell, J. A. (1990). *Experiments with Mixtures: Designs, Models, and the Analysis of Mixture Data*, 2^{nd} edition. New York: John Wiley & Sons, Inc.

[19] Creech, B., Smith, M., Davis, J., Tan, L., To, N., Fricker, S., and Gonzalez, J. M. (2011). Measurement Issues Study Final Report. *BLS Internal Report.*

[20] Czajka, J. L., Hirabayashi, S. M., Little, R. J. A., and Rubin, D. B. (1992). Projecting from Advance Data Using Propensity Modeling: An Application to Income and Tax Statistics. *Journal of Business and Economic Statistics*, **60**(2), 117–32.

[21] De Leeuw, E. and de Heer, W. (2002). Trends in Household Survey Nonresponse: A Longitudinal and International Perspective. Chapter 3 in Groves, et al. (eds.) *Survey Nonresponse*. New York: Wiley, 41–54.

[22] Fedorov, V. V. (1972). *Theory of Optimal Experiments*. New York: Academic Press.

[23] Fricker, S., Gonzalez, J. M., and Tan, L. (2011). Are You Burdened? Let's Find Out. *A paper presented at the 2011 Annual Conference for the American Association for Public Opinion Research*, Phoenix, AZ.

[24] Galesic, M. and Bosnjak, M. (2009). Effects of Questionnaire Length on Participation and Indicators of Response Quality in a Web Survey. *Public Opinion Quarterly*, **73**(2), 349–60.

[25] Garner, T. I., Janini, G., Passero, W., Paszkiewicz, L., and Vendemia, M. (2006). The CE and the PCE: A Comparison. *Monthly Labor Review*, 20–46.

[26] Gonzalez, J. M. and Eltinge, J. L. (2008). Adaptive Matrix Sampling for the Consumer Expenditure Quarterly Interview Survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 2081–8.

[27] Gonzalez, J. M. and Eltinge, J. L. (2010). Optimal Survey Design: A Review. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 4970–83.

[28] Gordis, L. (2000). *Epidemiology*, 2^{nd} edition. W. B. Saunders Company.

[29] Groves, R. M. (1989). *Survey Errors and Survey Costs*. John Wiley and Sons, Inc., Hoboken, New Jersey.

[30] Groves, R. M., Cialdini, R. B., and Couper, M. P. (1992). Understanding the Decision to Participate in a Survey. *Public Opinion Quarterly*, **56**, 475–95.

[31] Groves, R. M., Fowler J., Couper M. P., Lepkowski J. M., Singer E., and Tourangeau R. (2004). *Survey Methodology*. New York: Wiley.

[32] Groves, R. M. and Heeringa, S. G. (2006). Responsive Design for Household Surveys: Tools for Actively Controlling Survey Errors and Costs. *Journal of the Royal Statistical Society, Series A*, **169**(3), 439–57.

[33] Groves, R. M., Singer, E., and Corning, A. (2000). Leverage-Saliency Theory of Survey Participation: Description and an Illustration. *Public Opinion Quarterly*, **64**, 299–308.

[34] Heeringa, S. G. and Groves, R. M. (2004). Responsive Design for Household Surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 3644–51.

[35] Herzog, A. R. and Bachman, J. G. (1981). Effects of Questionnaire Length on Response Quality. *The Public Opinion Quarterly*, **45**, 549–59.

[36] Hinkins, S. M. (1984). Matrix Sampling and the Effects of Using Hot Deck Imputation. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 415–20.

[37] Horvitz, D. G. and Thompson, D. J. (1952). A Generalization of Sampling Without Replacement from a Finite Universe. *Journal of the American Statistical Association*, **47**, 663–85.

[38] Jang, D. S., Cox, B. G., and Edson, D. J. (1997). Generalized Variance Functions for Data from Multi-frame Surveys: The SESTAT Experience. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 158–63.

[39] Johnson, E. G. and King, B. F. (1987). Generalized Variance Functions for a Complex Sample. *Journal of Official Statistics*, **3**(3), 235–50.

[40] Johnson, T., O'Rourke, D., and Severns, E. (1998). Effects of Question Context and Response Order on Attitude Questions. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 857–60.

[41] Johnson, W. R., Sieveking, N. A., and Clanton, E. S. (1974). Effects of Alternative Positioning of Open-Ended Questions in Multiple-Choice Questionnaires. *Journal of Applied Psychology*, **59**, 776–8.

[42] Kalton, G. and Anderson, D. W. (1986). Sampling Rare Populations. *Journal of the Royal Statistical Society, Series A*, 149, 65–82.

[43] Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*, 2^{nd} edition. New York: John Wiley and Sons.

[44] Kish, L. (1965). *Survey Sampling*. New York: John Wiley and Sons.

[45] Kish, L. (1988). Multipurpose Sample Designs. *Survey Methodology*, **14**(1), 19-32.

[46] Kraut, A. I., Wolfson, A. D., and Rothenberg, A. (1975). Some Effects of Position on Opinion Survey Items. *Journal of Applied Pyschology*, **60**, 774–6.

[47] Kreuter, F., McCulloch, S., Presser, S., and Tourangeau, R. (2011). The Effects of Asking Filter Questions in Interleafed Versus Grouped Format. *Sociological Methods and Research*, **40**(1), 88–104.

[48] Kreuter, F., Olson, K., Wagner, J., Yan, T., Ezzati-Rice, T. M., Casas-Cordero, C., Lemay, M., Peytchev, A., Groves, R. M., and Raghunathan, T. E.. (2010). Using Proxy Measures and Other Correlates of Survey Outcomes to Adjust for Nonresponse: Examples from Multiple Surveys. *Journal of the Royal Statistical Society, Series A*, **173**(3), 389–407.

[49] Krosnick, J. A., Holbrook, A. L., Berent, M. K., Carson, R. T., Hanemann, W. M., Kopp, R. J., Mitchell, R. C., Presser, S., Ruud, P. A., Smith, V. K., Moody, W. R., Green, M. C., and Conaway, M. (2002). The Impact of "No

Opinion" Response Options on Data Quality: Non-Attitude Reduction or an Invitation to Satisfice? *Public Opinion Quarterly*, **66**, 371–403.

[50] Leaver, S. G., Johnson, W. H., Baskin, R., Scarlett, S., and Morse, R. (1996). Commodities and Services Sample Redesign for the 1998 Consumer Price Index Revision. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 239–44.

[51] Little, R. J. A. (1986). Survey Nonresponse Adjustments for Estimates of Means. *International Statistical Review*, **54**, 139–57.

[52] Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*, 2^{nd} edition. Hoboken, New Jersey, Wiley.

[53] Lohr, S. L. (1999). *Sampling: Design and Analysis*. Duxbury Press.

[54] Love, L. T. and Turner, A. G. (1975). The Census Bureau Experience: Respondent Availability and Response Rates. *Proceedings of the Business and Economics Section, American Statistical Association*, 76–85.

[55] Maindonald, J. and Braun, J. (2003). *Data Analysis and Graphics Using R: An Example-based Approach*. Cambridge University Press.

[56] Matthews, B. W. (1975). Comparison of the Predicted and Observed Secondary Structure of T4 Phage Lysozyme. *Biochimica et Biophysica Acta*, **405**, 442–51.

[57] McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, 2^{nd} edition. Chapman and Hall.

[58] Meyers, R. H. (1990). *Classical and Modern Regression with Applications*, 2^{nd} edition. Duxbury.

[59] Munger, G. F., and Lloyd, B. H. (1988). The Use of Multiple Matrix Sampling for Survey Research. *Journal of Experimental Education*, **56**, 187–191.

[60] National Institutes of Health (2002). *Rare Diseases Act of 2002*. Bethesda, MD. http://history.nih.gov/research/downloads/PL107-280.pdf (visited February 23, 2012).

[61] Office of Management and Budget (2006). *Standards and Guidelines for Statistical Surveys*. Washington, DC. http://www.whitehouse.gov/omb/inforeg/statpolicy/standards.pdf (visited September 10, 2009).

[62] Pugh, R. C. (1971). Empirical Evidence on the Application of Lord's Sampling Technique to Likert Items. *The Journal of Experimental Education*, **39**(3), 54–7.

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

[64] Robinson, J. P. and Godbey, G. (1997). *Time for Life: The Surprising Ways Americans Use Their Time*, 2^{nd} Edition. The Pennsylvania State University Press.

[65] Rosenbaum, P. R. and Rubin, D. B. (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika*, **70**, 41–55.

[66] Roszkowski, M. J. and Bean, A. G. (1990). Believe It or Not: Longer Questionnaires Have Lower Response Rates. *Journal of Business and Psychology*, **4**, 495–509.

[67] Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley: New York.

[68] Rutter, C. M., Zaslavsky, A. M., Feuer, E. J. (2010). Dynamic Microsimulation Models for Health Outcomes: A Review. *Medical Decision Making, Sage Publication*, 10–8.

[69] Särndal, C. E., Swensson, B., and Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer–Verlag.

[70] Sharp, L. M. and Frankel, J. (1983). Respondent Burden: A Test of Some Common Assumptions. *Public Opinion Quarterly*, **47**, 36–53.

[71] Shields, J. and To, N. (2005). Learning to Say No: Conditioned Underreporting in an Expenditure Survey. *American Association for Public Opinion Research – Proceedings of the Section on Survey Research Methods, American Statistical Association*, 3963–8.

[72] Shoemaker, D. M. (1973a). *Principles and Procedures of Multiple Matrix Sampling*. Cambridge, MA: Ballinger Publishing Company.

[73] Shoemaker, D. M. (1973b). A Note on Allocating Items to Subsets in Multiple Matrix Sampling and Approximating Standard Errors of Estimates with the Jackknife. *Journal of Educational Measurement*, **10**, 211–9.

[74] Silvey, S. D. (1980). *Optimal Design*. New York: Chapman and Hall.

[75] Sirotnik, K. A. (1970). An Investigation of the Context Effect in Matrix Sampling. *Journal of Educational Measurement*, **7**(3), 199–207.

[76] Sudman, S. (1967). *Reducing the Costs of Surveys*. Chicago: Aldine.

[77] Thomas, N., Raghunathan, T. E., Schenker, N., Katzoff, M. J., and Johnson, C. L. (2006). An Evaluation of Matrix Sampling Methods Using Data from the National Health and Nutrition Examination Survey. *Survey Methodology*, **32**, 217–31.

[78] To, N., Davis, J., and Creech, B. (2011). *Comparison of Consumer Expenditure Survey Designs*, U.S. Bureau of Labor Statistics. Internal document, February, 2011.

[79] Varberg, D. and Purcell, E. J. (1997). *Calculus*, 7^{th} edition. Prentice Hall: New Jersey.

[80] Valliant, R. (1987). Generalized Variance Functions in Stratified Two-Stage Sampling. *Journal of the American Statistical Association*, **82**(398), 499–508.

[81] Valliant, R. and Gentle, J. E. (1997). An Application of Mathematical Programming to Sample Allocation. *Computational Statistics and Data Analysis*, **25**(3), 337–60.

[82] West, B. (2010). An Examination of the Quality and Utility of Interviewer Estimates of Household Characteristics in the National Survey of Family Growth (NSFG). *A presentation given for the University of Michigan's Program in Survey Methodology Brown Bag Seminar Series, November 2010*.