Multiple Matrix Sampling: A Review December 2007

Jeffrey M. Gonzalez and John L. Eltinge
U.S. Bureau of Labor Statistics, Office of Survey Methods Research
2 Massachusetts Avenue NE, Washington, DC 20212
Gonzalez.Jeffrey@bls.gov

Abstract

The Consumer Expenditure Quarterly Interview Survey (CEQ) is an ongoing panel survey of U.S. households in which detailed information on an estimated 60 to 70 percent of total expenditures for a consumer unit is collected. The CEQ is generally administered face-to-face and takes about 65 minutes to complete. One proposed method to decrease the length of a given interview is to use multiple matrix sampling. This would involve dividing the questionnaire into sections of questions and then administering these sections to subsamples of the main sample. We provide an overview of the current research on multiple matrix sampling. We review its origins, highlight the fields in which it has received the most application and discuss how it has been applied to problems in surveys. We then discuss the phases of the survey process that require consideration when implementing a multiple matrix sampling design and conclude with a few mathematical considerations and an identification of future work.

KEY WORDS: Split questionnaire; Respondent burden; Nonresponse; Sample survey; Selection probability; Variance estimation

1. Introduction and Motivation

The Consumer Expenditure Quarterly Interview Survey (CEQ) is an ongoing panel survey of U.S. households in which detailed information on an estimated 60 to 70 percent of total expenditures for a consumer unit (CU) is collected.[1] The CEQ is generally administered face-to-face and on average takes 65 minutes to complete (Bureau of Labor Statistics, U.S. Department of Labor, *Handbook of Methods*, Chapter 16, April 2007 edition, Consumer Expenditures and Income). Data from the CEQ are used in calculating the cost weights for the Consumer Price Index (CPI), one of the nation's leading economic indicators; thus, unbiased and precise estimates of family expenditures are essential for this calculation.

The statistical properties of these estimators can be influenced by the quality of the data collected, which in turn may be affected by characteristics of the survey instrument itself. Previous research has indicated that a lengthy questionnaire can have adverse effects on data quality (Kraut *et al.* 1975; Johnson *et al.* 1974; Herzog and Bachman 1981). For example, using a survey involving a nationally representative sample of high school seniors, Herzog and Bachman (1981) concluded that the probability of providing accurate responses is likely to decline if the survey process extends beyond some optimal length. They determined that this was due to a decrease in motivation to continue to comply with the survey request. Respondents with a reduced motivation may be more likely to look for easier ways to respond to questions (e.g., straight-line responding - an increased propensity to give identical responses for questions or items with similar response categories) or even prematurely terminate the survey request. For instance, anecdotal evidence from interviewers for the CEQ suggests that respondents learn to report no expenditures of a certain type (e.g., vacation expenditures) so that they will not get asked subsequent, more specific questions about that expenditure type (Shields and To, 2005).

A lengthy survey may not only affect the quality of data via premature termination of participation or straight-line responding, but it may also affect a sample member's decision to respond (Burchell and Marsh 1992; Groves, Singer and Corning 2000). It has been well-documented that household survey response rates have been steadily declining over recent years (de Leeuw and de Heer 2002). For instance, since July 2000, the response rate for CEQ has been gradually declining from about 80% to about 76% in July 2006.[2] Because of this, there is an increasing concern about the effect of nonresponse bias on estimates from these surveys. Furthermore, a potential respondent may be less inclined to participate in a survey in the absence of an intrinsic interest in the survey (Groves, Singer and Corning 2000). Thus, without prior knowledge of a sample member's intrinsic interest in the survey, it may be in the survey organization's best interest to administer a shorter questionnaire in hopes of obtaining a higher response rate and higher quality data from the questions asked.

Beyond obtaining higher response rates and quality data, another advantage of using a shorter questionnaire is a potential reduction in data collection costs. A shorter questionnaire should require less interviewing time; thus, each interviewer should be able to handle a larger caseload. This should result in a smaller interviewing staff which may reduce data collec-

[1] A consumer unit is the unit for which expenditure reports are collected. It is defined as: "(1) all members of a particular housing unit who are related by blood, marriage, adoption, or some other legal arrangement, such as foster children; (2) a person living alone or sharing a household with others, or living as a roomer in a private home, lodging house, or in permanent living quarters, in a hotel or motel, but who is financially independent; or (3) two or more unrelated persons living together who pool their income to make joint expenditure decisions. Students living in university-sponsored housing are also included in the sample as separate consumer units." (Bureau of Labor Statistics, U.S. Department of Labor, *Handbook of Methods*, Chapter 16, April 2007 edition, Consumer Expenditures and Income)

[2] These response rates were computed internally.

tion costs. It should also be noted that the development and implementation of any design modifications to an existing survey may require potentially high initiation costs, but the costs incurred during these initial stages could be offset by the reduction in future data collection costs.

However, there are tradeoffs to administering a shorter questionnaire. The organization conducting the survey must decide which questions to eliminate from the questionnaire. In a survey like the CEQ, eliminating questions altogether may be difficult since this survey has an obligation to provide the basis for revising the CPI cost weights and to collect detailed information on family expenditures. With this in mind, researchers at the Bureau of Labor Statistics (BLS) are investigating survey design approaches to shorten the questionnaire while still collecting the necessary expenditure information from at least some of the respondents. One proposed method, multiple matrix sampling, is drawn from the educational assessment literature. Multiple matrix sampling, sometimes referred to as a split questionnaire, is a technique that involves dividing the questionnaire into sections of questions, possibly overlapping, and then administering these sections to distinct subsamples of the main sample. While this approach ensures that all of the necessary questions are asked of at least some of the respondents, administering each question to fewer sample members may reduce the efficiency of any estimate calculated from those questions (i.e., as the sample size for measuring a characteristic decreases, its sampling variance increases).

This paper provides an overview of the past research on multiple matrix sampling. First, we review the origins of this method and provide examples from various fields in which multiple matrix sampling designs have been examined. We then discuss considerations for splitting a questionnaire, collecting data using a split questionnaire and processing and analyzing the collected data. We conclude with a discussion of mathematical issues, proposed future work and a summary of main points.

2. Origins and Previous Applications

2.1 The Origins: Educational Assessment

Multiple matrix sampling appears to have originated in the early 1950s when Turnbull, Ebel and Lord, researchers at the Educational Testing Service, looked favorably on this technique for sampling items for and estimating the normative distribution of standardized tests (Shoemaker 1973). However, it was Lord, Hooke and Tukey who developed the early statistical procedures for estimating population moments and other quantities from multiple matrix sampling designs (Shoemaker 1973). Then Shoemaker (1973), in the first text solely devoted to multiple matrix sampling, summarized the statistical methodology, including estimation and hypothesis testing, used in multiple matrix sampling designs and highlighted some of the procedural guidelines for implementing this technique.

Multiple matrix sampling was an attractive option for addressing problems in the educational assessment field for myriad reasons. First, there were often several questions available that could be asked of a student in order to assess knowledge of a concept, but it was infeasible to ask all of these questions of a single student. Multiple matrix sampling was viewed as a plausible option for randomly sampling questions from a large universe of questions, administering these samples to students and accurately measuring students' knowledge. A second benefit was that researchers saw a potential reduction in testing time, since classroom testing time was limited within the school day. Furthermore, if test administrators proposed conducting a shorter test, then the school may have been more likely to comply with the request. Thus, a shorter test would be beneficial not only because testing time would be reduced but also because school participation would likely increase. Finally, if students were administered different tests assessing knowledge of the same concepts, then there might be a higher likelihood of capturing that student's true educational attainment by mitigating the possibility of students copying each other's examinations.

In summary, multiple matrix sampling was first utilized in the educational assessment field to select subsets of items and examinees in order to estimate the normative distribution of standardized tests. In achieving this analytic objective with shorter tests, educational researchers also likely obtained higher school participation rates and a more accurate assessment of a student's educational attainment. Survey researchers in other fields, such as the government, public health and business fields, recognized that declining response rates and poor data quality were also problems encountered in their own work and began to explore how multiple matrix sampling techniques could remedy these problems. The next section identifies a few examples of how these researchers have explored multiple matrix sampling designs.

2.2 Previous Applications: Government and Public Health

Applications of multiple matrix sampling by government agencies began as early as 1970, when the U.S. Census Bureau used a nested sampling design with two long forms and administered one to 15 percent and the other to 5 percent of the population (Navarro and Griffin, 1993). The U.S. Census Bureau also revisited the idea of multiple matrix sampling when they looked at this method as a viable option for reducing respondent burden and differential undercount as well as improving coverage in the 2000 Decennial Census. Navarro and Griffin (1993) identified five multiple matrix sampling schemes for potential implementation in the 2000 Decennial Census and addressed issues of reliability and respondent burden associated with each design. The designs were developed paying special attention to cross-tabulation of data items, sample size and small area estimation. For instance, data items, like "place of work" and "journey to work", which required cross-tabulations, were almost always put on the same sample form.

The five matrix sampling designs they proposed varied in the number of sample forms, items per form and sampling rate associated with each form. For example, one design consisted of three forms in which each form contained economic ques-

tions and had only two of the following blocks of questions: "Soc I", "Soc II" and "Housing". The overall sampling rate for this design was 20% of the population so the sampling rate associated with the economic questions was 20% while each of the other blocks collected data from 13.3% of the population. Navarro and Griffin then compared the five designs using coefficient of variation and a crude measure of respondent burden. Coefficient of variation (CV), the ratio of the standard error of the estimate to the estimate, was used to assess reliability. The crude measure of respondent burden, calculated by multiplying the number of items on the form by the overall sampling rate and then adjusting for item nonresponse, was used to evaluate the reduction in burden from the 1990 Decennial Census among the five designs. A preliminary comparison of the five designs revealed that, in terms of reliability, the designs provided acceptable or adequate estimates for most areas and, in all but one design, the level of respondent burden decreased from that of the 1990 Decennial Census.

The above results were encouraging, but they also identified areas that needed further research if multiple matrix sampling was to be used in the 2000 Decennial Census. Navarro and Griffin's proposed next step was to conduct a correlation analysis to determine the optimal grouping among items on different forms. The intent was that highly correlated items would go on different forms so that imputation models could be developed to predict the items not contained on a particular form. Navarro and Griffin also planned to conduct simulations, using 1990 Decennial Census data, to assess the potential loss in accuracy of estimates and to evaluate the utility of their imputation models. In our review of the literature, we were unable to find any information as to whether any of these designs were actually implemented in the 2000 Decennial Census.

Other applications of multiple matrix sampling within governmental agencies were explored in the 1980s by the Internal Revenue Service (IRS). Hinkins (1983) described the IRS's implementation of this technique for reviewing and editing corporate tax returns. For certain items on tax returns, taxpayers are required to supply supplemental information on an attached schedule. On corporate tax returns, one example would be "Other Income". A taxpayer who reports "Other Income" attaches a schedule detailing what this income is. The IRS then reviews all schedules and potentially corrects the information to make sure this income should not be reported elsewhere on the tax form or combined with other income items. Reviewing forms requiring no changes or edits would be highly inefficient. Thus, to reduce costs and save time, they used the methods of multiple matrix sampling to determine which forms to subsample for review and then to estimate the edited income on the forms not sampled.

Ideally, the IRS only wanted to conduct a detailed review of forms that would result in a change so they quickly reviewed all corporate tax forms and grouped them into two strata - Total Assets of $250 million or more and Total Assets of less than $250 million. IRS researchers believed that a review of corporate tax forms with higher assets would more likely result in a correction than forms with lower assets. Therefore, they reviewed all forms in the first stratum and subsampled forms and items in the second stratum for review. This method falls naturally into the double sampling framework, but they determined that using the estimation procedures developed for double sampling was not feasible since each form had multiple items that potentially needed editing. Thus, for the forms that were not subsampled for editing, they used a hot-deck imputation method for estimating the edited income amounts. This procedure involved creating adjustment cells in which a record with schedules to be imputed was matched with schedules within that same cell. It should be noted that instead of imputing "amount of change" they imputed "percent change".

After the IRS implemented their proposed multiple matrix sampling design, they applied it to a sample of approximately 3,000 records to assess the effectiveness of their design. It appears that stratification was successful in identifying tax forms that resulted in changes to the "Other Income" schedule. However, they were unsuccessful in predicting records without changes. They also revisited their assessment of the effectiveness of the design by investigating the effects of using the hot deck imputation procedure (Hinkins, 1984), and concluded that the hot-deck imputation procedure would not significantly affect important population and subpopulation estimates. Although these results were encouraging, they also cited the importance of refining their imputation models and making sure there was an adequate sample size within each adjustment cell.

Finally, one of the most notable applications of multiple matrix sampling was presented in 1995 by Raghunathan and Grizzle. Using a public health survey, the Cancer Risk Behavior Survey, they addressed the following question: what modification can be made to the design stage of a survey in order to lower respondent burden and possibly raise the response rate?

A typical interview for the Cancer Risk Behavior Survey takes about 30 minutes, but can take as long as 50 minutes. Due to the perceived high respondent burden as a consequence of the length of the interview, they proposed splitting the questionnaire and administering a random sample of questions to randomly sampled individuals. Under a multiple matrix sampling design, if questions are randomly assigned to respondents, then analysts can assume that the questions not asked are missing completely at random (MCAR). Using the MCAR assumption, it is possible to estimate characteristics of the marginal distributions of the variables, fit regression models and perform categorical data analysis. However, Raghunathan and Grizzle noted that one of the main disadvantages with this method is that some combinations of questions may never be asked together on the same questionnaire; therefore, their corresponding associations cannot be estimated using traditional statistical methods. To address this problem, they developed a multiple imputation method for analyzing the data from the split questionnaire which creates a complete data set so that correlations and other quantities can be estimated. They compared their inferences from that of the original survey to those from the split questionnaire, investigated the loss of efficiency in using this design and evaluated the robustness of the multiple imputation procedures.

Using existing data from the full questionnaire, they assessed the quality of the multiple imputation method by com-

paring point estimates of proportions and the associated standard errors of twelve variables of interest from the full questionnaire to the multiple imputation method and the available case method (the available case method uses only the data collected from that split form). They based their comparisons on a discrepancy measure - the coefficient of variation of the split data estimates (e.g., multiple imputation and available case methods) around the "true" value. The "true" value was based on estimates from the full questionnaire. They found that, in general, the estimates of the proportions for the twelve items obtained using either the available case method or the multiple imputation method were very similar to those obtained from the full questionnaire. Overall, the standard error estimates from both of these methods were larger than those obtained from the full questionnaire, but the multiple imputation method resulted in smaller standard error estimates than the available case method for all variables of interest. They also compared the results obtained from various linear regression analyses and found that the multiple imputation method resulted in narrower confidence intervals than the available case method, an indication that the multiple imputation method performed better than the available case method.

3. The Survey Process: Considerations when Splitting a Questionnaire

From the literature, we have identified three phases of the survey process that may guide the implementation of a multiple matrix sampling design. They are questionnaire development, data collection and processing and analysis. Our discussion of these phases is from the perspective of modifying the design of an existing survey via multiple matrix sampling. We focus on features of the CEQ related to these phases and then discuss how they may guide implementation of multiple matrix sampling in the CEQ. Questionnaire development would involve determining how to split the original questionnaire, data collection would entail deciding which sample members are administered each sub-questionnaire and processing and analysis would involve any post-data collection procedures, including analyzing the collected data.

3.1 Development of a Split Questionnaire

The first phase of the survey process that requires consideration is questionnaire development. Expanding on the definition above, splitting the questionnaire entails allocation of survey items to each sub-questionnaire as well as the optimal number of sub-questionnaires. These decisions should be consistent with the objectives of the original survey and informed by characteristics of that existing survey. For example, the two main objectives of the CEQ are to provide the basis for the CPI cost weights revision and to collect detailed family expenditure information. Therefore, any decision regarding the number of forms and items per form should meet these goals.

The mode of data collection and content of the original survey are also related to the design decisions involved in splitting a questionnaire. The proliferation of computer technology in survey operations has resulted in designing survey instruments with more intricate skip patterns and logic because surveys now generally make use of a computer assisted interview (CAI) instrument to handle these features. These CAI instruments route the interviewer and respondent through the questionnaire. Because the CEQ makes use of a CAI instrument for data collection, it is naturally more complex, in terms of logic and question dependencies, than surveys conducted via paper-and-pencil. Couple this with the fact that the survey is already designed to collect very detailed expenditure information, then one may fully understand the difficulty of attempting to implement a multiple matrix sampling design on the CEQ.

We illustrate how challenging splitting a questionnaire could be with the following example. When multiple matrix sampling was first used in educational assessment, there was a universe of questions to select from. This universe contained disjoint sets of questions measuring different concepts. As a simple example, when assessing mathematical aptitude, the universe of questions may contain a set of addition problems and a set of multiplication problems. Educational researchers could randomly select any addition problem as well as any multiplication problem. These two questions are seemingly unrelated, i.e., asking one does not depend on asking the other, and could be easily allocated to different tests. In contrast, most surveys, such as the CEQ, have questions that are related both contextually and logically to other questions. This increases the complexity of assigning questions to different forms because if two questions are logically related, then they must appear on the same form.

As demonstrated in the example above, the allocation of questions to different forms is a nontrivial task. Allocation of items can occur in several ways. The simplest method of allocation is to randomly sample questions and place them on different forms. Because of the complexity of most surveys and interrelated-nature of their questions, this method may need modifications. A modification to this approach may be to construct thematically (and by default, most likely logically related) blocks of questions. For instance, in the CEQ, all questions pertaining to health care expenditures may form one block while all data items about cash contributions would compose another block. It is then possible to randomly sample these blocks and then distribute them among the sub-questionnaires.

Survey designers can also base the allocation of questions on various statistical criteria. One method is to examine correlations among questions on the original survey and identify those that are most related. Questions with high correlations would then be allocated to different sub-questionnaires. This method was proposed under the assumption that multiple imputation techniques would be used to analyze the data collected from the multiple matrix sampling forms. The idea was that the questions not asked on one form could be predicted (imputed) by highly correlated items on that form (Raghunathan and Grizzle 1995). A second method would be to develop an algorithm that would automatically distribute items among a set number of forms. Thomas *et al.* (2006) proposed one technique that utilized an index of predictive value. This index represented the proportion of the difference between the

variance of the no-imputation estimator and the variance that would have been obtained with complete data that is recovered by the multiple-imputation estimator.

There are also questions that may not be randomly or statistically allocated to different forms. Researchers often have a "high priority" list of questions, a core, that are of special interest and/or may require more precision than other questions. In order to meet these demands, it would be beneficial to ask these questions of all sample members. Thus, one multiple matrix sampling design could be to have a core set of questions appear on every sub-questionnaire, with each version having a distinct set of allocated questions using one of the techniques described above. An example of "high priority" questions may be those questions that are most predictive of other questions. The core, in combination with the allocated questions to that form could then be used to predict the information not collected by that sub-questionnaire (Raghunathan and Grizzle 1995). With respect to the CEQ, since one of the main goals is to provide the basis for the CPI cost weights revision, one could identify those questions that directly relate to the CPI revision process and include those in the core.

Finally, the survey practitioner must also determine an optimal number of forms. Since the motivation for using multiple matrix sampling is to improve data quality and reduce nonresponse by shortening the length of the interview and reducing respondent burden, then the number of forms should be chosen so that each achieves a balance between length and cognitive demand. If the time required to complete a form is severely disproportionate across forms, then the survey organization might have achieved little in improving overall data quality and response rates. In addition, the content of various questions may impose differing cognitive requirements on the respondent. As an extreme example from the CEQ, if the allocation method placed all recurring expenditures (e.g., monthly bills) on one form and very detailed expenditures (e.g., clothing items) on another, then the cognitive demand on the respondent from the latter would likely be higher than that from the former because respondents could have greater difficulty in recalling detailed expenditures over those that tend to be relatively constant throughout the reference period. Thus, the number of forms is guided by the motivating reasons for implementing a multiple matrix sampling design.

3.2 Data Collection using a Split Questionnaire

The next phase of the survey process that requires consideration is data collection. This involves determining which sample members receive which form. Similar to developing the split questionnaire, the design decisions for data collection should be consistent with objectives of the original survey. For the CEQ, the data collection procedures must meet the needs of the CPI cost weights revision as well as continue to collect detailed expenditure information.

The sample design of the original survey is also related to data collection using the split questionnaire. For cross-sectional surveys, data collection for each sub-questionnaire could be a microcosm of the sample design of the original survey. For example, if the split questionnaire consisted of five sub-questionnaires, then each sub-questionnaire could be administered to each of five replicates of the main sample. A similar method could be implemented for panel surveys (i.e., surveys in which sample units are interviewed more than once over a period of time); however, these types of surveys pose some unique and interesting challenges when collecting data. Before discussing these challenges, we provide an example of a panel survey to illustrate how the sample design affects the data collection decisions made when using a split questionnaire.

The current CEQ is a rotating panel survey. For this survey, each CU in the sample is contacted for an interview every three months over five calendar quarters. The sample for each quarter is then divided into three panels, each corresponding to a month within the quarter. The initial interview is a bounding interview for collecting demographic characteristics, an inventory of major durable goods and expenditures using a one-month recall period. This interview is thought to reduce any biases that may be introduced via forward telescoping; thus, the expenditure information captured during this interview is not used in any BLS published estimates. The second through the fifth interviews use a uniform questionnaire to collect expenditures in the previous quarter using a direct monthly recall approach or a quarterly recall method. It should be noted that in all five interviews the major expenditure categories (e.g., housing, transportation, health care, etc.) for which expenditure information is collected are identical and only the manner in which the information is collected varies between the initial interview and the remaining four (e.g., one month recall for the initial interview versus a combination of one and three month recall for the remaining four). In addition, some topics are only collected in the second and fifth interviews.

Given this type of design, a survey practitioner must decide whether the form administered in the initial interview dictates the forms administered in subsequent interviews. With regard to the CEQ, if we assume that each form has a core in addition to distinct subsets of other questions, then one approach would be to administer the same subsets in all five interviews. This approach could be the easiest to implement in that it requires minimal interim data collection monitoring and has some immediate benefits. First, the second through fifth interviews are still bounded. If the questions changed in subsequent interviews, then forward telescoping may occur. Also, varying the expenditure categories may increase the cognitive burden on the respondent by requiring the respondent to think about and report different expenditures each interview. Occurrences of both of these may adversely affect data quality. Finally, if the analyst uses imputation techniques, then the requirement that the data are MAR is still met.

By contrast, rotating questions across interviews could enhance the multiple matrix sampling design. For instance, information obtained during the initial interview may help identify the optimal subset of questions to administer in subsequent interviews. Building on the research from Hinkins (1983), suppose that during the first interview of the CEQ, certain CU characteristics were identified that are known predictors of certain types of expenditures. For example, in the initial interview, it was discovered that a CU contained chil-

dren. That CU would potentially have a different health care expenditure pattern than a single-person CU; thus, we could obtain more efficient estimates of quarterly health care expenditures by collecting that information more frequently from CUs with children.

3.3 Processing and Analysis from a Split Questionnaire

The last phase of the survey process we consider is the processing and analysis phase. This phase would involve any post-data collection procedures, including but not limited to consistency checks, weighting, imputation and analysis. The objectives of the existing survey play a major role in how the data from a split questionnaire are processed and analyzed. Also, the design decisions made during the other two phases guide what decisions practitioners make regarding the post-data collection procedures.

The CPI is a major stakeholder for CEQ data; accordingly, an understanding of how it uses the current CEQ data for revising the cost weights is essential when deciding how to process and analyze the data from a split CEQ questionnaire. Information from the CEQ is also used in the calculation of the sampling variance for the 6-month price change for the commodities and services portion of the CPI (Bureau of Labor Statistics, U.S. Department of Labor, *Handbook of Methods*, Chapter 17, June 2007 edition, The Consumer Price Index). More specifically, the relative importance of certain expenditure item groups are directly incorporated into this calculation. Under a multiple matrix sampling design, not all of the expenditure item groups may be collected from all CUs. In order to obtain a more realistic estimate of the relative importance for these expenditure item groups, the post-data collection procedures must be altered to reflect these changes. One solution may be to incorporate imputation methods into the processing systems of the modified survey.

Incorporating imputation methods may not only meet the CPI requirements, but it may meet the requirements of other users. The BLS has an obligation to meet the needs of external data users such as academic researchers and public and private institutions. For both statistical and non-statistical purposes, these users often want complete records for each CU. Imputing the data that are not captured under the multiple matrix sampling design is a method that can create complete records for every CU. Furthermore, these methods, in general and under the multiple matrix sampling framework, are actively being researched. Thus, general acceptance of these methods as well as our understanding of how to utilize them and their implications may increase in the future.

As noted, the design decisions made in the other two phases may affect the implemented post-data collection procedures. For instance, similar to a design that was explored by Navarro and Griffin (1993) a split questionnaire may consist of several sub-questionnaires being administered to distinct subsamples with the original full questionnaire being administered to one subsample. If imputation methods were used to recapture the information not asked on the sub-questionnaires, then administering the original full questionnaire to a subsample would assist in imputation model development and validation.

4. Discussion

Our proposed future work is motivated by the mathematical considerations related to implementing a multiple matrix sampling design. These considerations can be classified into three main categories - population quantities of interest, estimation procedures and evaluation criteria. When modifying existing surveys, these tend to be related to the analysis goals of the original survey.

For many surveys, some population quantities of interest are univariate statistics such as, means and totals. For instance, using CEQ data, the BLS estimates and publishes the average quarterly expenditure on a particular item per CU (Bureau of Labor Statistics, U.S. Department of Labor, *Handbook of Methods*, Chapter 16, April 2007 edition, Consumer Expenditures and Income). If the CEQ employed a multiple matrix sampling design, then there would be at least two techniques available to continue to accomplish these goals (e.g., available case method and multiple imputation). These goals are easily accomplished by constructing an estimate using the available case method (i.e., use only the cases in which the expenditure information was collected directly when estimating the desired quantity). If, however, there are other population quantities of interest (e.g., coefficients of a generalized linear model), then the available case method would only be an option for expenditures appearing on the same form. Thus, multiple imputation might prove to be the more appropriate procedure. It is with these types of decisions that we focus our future work.

We plan to explore the mathematical properties of changes to the current sample design and analytic procedures. We will investigate alternative sample designs to sample both expenditures and CUs. We will also investigate how to modify selection probabilities, post-stratification and calibration weights and variance estimation procedures as a result of the changes to the sample design. We will empirically evaluate using existing CEQ data trade-offs in estimation efficiency for various quantities (e.g., expenditure means).

The review of the literature on multiple matrix sampling was motivated by the potential for applying this technique to the CEQ. As noted, we hope that modifying the current CEQ via multiple matrix sampling will improve data quality by decreasing respondent burden, lower nonresponse rates and decrease long-term data collection costs (aside from any initiation costs incurred). With examples drawn from the educational assessment, government, and public health fields, we have highlighted the previous research conducted on multiple matrix sampling. We identified three phases of the survey process that guide the implementation of these designs and concluded with summary of our future work that will be motivated by the various open mathematical problems associated with multiple matrix sampling. Finally, we hope that we have conveyed that implementation on an existing survey must be consistent with the main objectives of the original survey and that changes in the current structure must be further developed and studied.

Acknowledgments

The views expressed in this paper are those of the authors and do not necessarily reflect the policies of the U.S. Bureau of Labor Statistics. The authors would like to thank Lucilla Tan, Jennifer Edgar, Karen Goldenberg, David McGrath, Steve Cohen, Sioux Groves, Bill Mockovak, Clyde Tucker and Ashley Bowers for helpful discussions of the Consumer Expenditure Quarterly Survey, multiple matrix sampling and earlier versions of this paper.

REFERENCES

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**, 193-197.

Beaton, A.E. and Zwick, R. (1992). Overview of the National Assessment of Educational Progress. *Journal of Educational Statistics*, **17**, 95-109.

Burchell, B. and Marsh, C. (1992). The Effect of Questionnaire Length on Survey Response. *Quality and Quantity*, **26**, 233-44.

Bureau of Labor Statistics, U.S. Department of Labor, *Handbook of Methods*, Chapter 16, April 2007 edition, Consumer Expenditures and Income, on the Internet at http://www.bls.gov/opub/hom/pdf/homch16.pdf (visited June 21, 2007).

Bureau of Labor Statistics, U.S. Department of Labor, *Handbook of Methods*, Chapter 17, June 2007 edition, The Consumer Price Index, on the Internet at http://www.bls.gov/opub/hom/pdf/homch17.pdf (visited August 23, 2007).

Cochran, W.G. (1977). *Sampling Techniques*, Third Edition. New York, Wiley.

De Leeuw, E. and de Heer, W. (2002). Trends in Household Survey Nonresponse: A Longitudinal and International Perspective. Chapter 3 in R.M. Groves *et al.* (eds.) *Survey Nonresponse*, New York: Wiley, 41-54.

Ghosh, D. and Vogt, A. (2000). Determining an Optimal Split for a Lengthy Questionnaire. *American Statistical Association, Proceedings of the Section on Survey Research Methods*, 693-6.

Gressard, R.P. and Loyd, B.H. (1991). A Comparison of Item Sampling Plans in the Application of Multiple Matrix Sampling. *Journal of Educational Measurement*, **28**, 119-130.

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.

Herzog, A.R. and Bachman, J.G. (1981). Effects of Questionnaire Length on Response Quality. *The Public Opinion Quarterly*, **45**, 549-59.

Hinkins, S.M. (1983). Matrix Sampling and the Related Imputation of Corporate Income Tax Returns. *American Statistical Association, Proceedings of the Section on Survey Research Methods*, 427-33.

Hinkins, S.M. (1984). Matrix Sampling and the Effects of Using Hot Deck Imputation. *American Statistical Association, Proceedings of the Section on Survey Research Methods*, 415-20.

Hinkins, S., Parsons, V., and Scheuren, F. (1999). Inverse Sampling Algorithm for NHIS Confidentiality Protection. *American Statistical Association, Proceedings of the Section on Survey Research Methods*, 485-502.

Houseman, E.A. and Milton, D.K. (2005). Partial Questionnaire Designs, Questionnaire Non-response, and Attributable Fraction: Applications to Adult Onset Asthma. *Statistics in Medicine*, **25**, 1499-1519.

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.

Kennickell, A.B. (1997). Multiple Imputation and Disclosure Protection: The Case of the 1995 Survey of Consumer Finances. In W. Alvey and B. Jamerson (eds.) *Record Linkage Techniques*, 248-67. National Academy Press, Washington, D.C.

Kraut, A.I., Wolfson, A.D. and Rothenberg, A. (1975). Some Effects of Position on Opinion Survey Items. *Journal of Applied Psychology*, **60**, 774-6.

Little, R.J.A. (1993). Statistical Analysis of Masked Data. *Journal of Official Statistics*, **9**, 407-26.

Mislevy, R.J. (1983). Item Response Models for Grouped Data. *Journal of Educational Statistics*, **8**, 271-288.

Myerber, N.J. (1979). The Effect of Item Stratification on the Estimation of the Mean and Variance of Universe Scores in Multiple Matrix Sampling. *Educational and Psychological Measurement*, **39**, 57-68.

Navarro, A. and Griffin, R.A. (1993). Matrix Sampling Designs for the Year 2000 Census. *American Statistical Association, Proceedings of the Section on Survey Research Methods*, 480-5.

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

Rässler, S., Koller F. and Mäenpää C. (2001). A Split Questionnaire Survey Design applied to German Media and Consumer Surveys. *Paper presented at The International Conference on Improving Surveys*, Copenhagen, Denmark.

Rubin, D.B. (1993). Discussion: Statistical Disclosure Limitation. *Journal of Official Statistics*, **9**, 462-468.

Schafer, J.L. and Schenker, N. (2000). Inference with Imputed Conditional Means. *Journal of the American Statistical Association*, **95**, 144-54.

Scheetz, J.P. and Forsyth, R.A. (1977). A Comparison of Simple Random Sampling versus Stratification for Allocating Items to Subtests in Multiple Matrix Sampling. *Paper presented at The Annual Meeting of the National Council on Measurement in Education*, New York.

Shields, J. and To, N. (2005). Learning to Say No: Conditioned Underreporting in an Expenditure Survey. *American Association for Public Opinion Research - American Statistical Association, Proceedings of the Section on Survey Research Methods*, 3963-8.

Shoemaker, D.M. (1973). A Note on Allocating Items to Subsets in Multiple Matrix Sampling and Approximating Standard Errors of Estimate with the Jackknife. *Journal of Educational Measurement*, **10**, 211-219.

Shoemaker, D.M. and Knapp, T.R. (1974). A Note on Terminology and Notation in Matrix Sampling. *Journal of Educational Measurement*, **11**, 59-61.

Shoemaker, D.M. (1973). *Principles and Procedures of Multiple Matrix Sampling*. Cambridge, MA: Ballinger Publishing Company.

Thomas, N. and Gan, N. (1997). Generating Multiple Imputations for Matrix Sampling Data Analyzed with Item Response Models. *Journal of Educational and Behavioral Statistics*, **22**, 425-445.

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**, 217-231.

Zeger, L.M. and Thomas, N. (1997). Efficient Matrix Sampling Instruments for Correlated Latent Traits: Examples from the National Assessment of Educational Progress. *Journal of the American Statistical Association*, **92**, 416-425.