MATRIX SAMPLING DESIGNS FOR THE YEAR 2000 CENSUS

Alfredo Navarro and Richard A. Griffin[1]
Alfredo Navarro, Bureau of the Census, Washington DC 20233

I. Introduction and Background

Over the past fifty years, the Bureau of the Census has transformed the decennial census from a 100 percent data collection activity into an operation which collects the bulk of census data on a sample basis. In spite of this, the demand for new and additional information have resulted in a steady increase in respondent burden since 1970.

A major goal of the year 2000 census is to keep the "average" burden on respondents at the 1990 census level while at the same time meet as many data needs as possible. The use of multiple sample forms, also known as matrix sampling, is being considered to help meet this goal. Matrix sampling involves dividing up the full set of sample data items among several sample questionnaires or forms. This is a major departure from 1980 and 1990 when a single sample form was used to collect the sample content. However this will not be the first time the Bureau of the Census employs a multiple sample forms design to collect sample data. The 1970 census used a nested sampling design based on two long forms administered to samples of 15 and 5 percent of the population.

If the overall respondent burden is held to 1990 levels, then with an increase in sample content (even if the sample data items are spread over multiple forms), the reliability of estimates for any one item likely will be lower than for 1990. This may be a major concern for small area data users. Conversely, if we specify the reliability levels needed for small area estimates, the burden will increase compared to 1990. Related to this issue, are questions about the need for, and reliability of, estimates based on cross-tabulations of items. Cross-tabulations of items that do not appear on the same sample form cannot be directly estimated, and so model based estimates would have to be considered to prepare this type of estimates.

A very important goal of the 2000 census is to improve coverage and reduce the differential undercount. If content is essentially kept the same as in 1990, then spreading this content over several sample forms will likely reduce respondent burden while providing sample forms that are shorter than the 1990 sample form. This could increase mail return rates. Results from the 1990 census evaluation studies indicate that the quality of data, particularly in terms of coverage, is somewhat better for mail return questionnaires than for those not returned by mail and subsequently completed by enumerators during follow-up operations (Griffin and Moriarity, 1992). Therefore, the use of shorter multiple sample forms could help to improve coverage for the 2000 census. This paper describes and discusses reliability and respondent burden issues related to five alternative matrix sampling plans, the first four could be used for sample data collection for the Year 2000 Census. We are now in the process of validating these designs based on results from a study that will produce information about what variables are highly correlated. Results from this work will help to determine an optimal way to "split up" content across multiple forms.

The matrix sampling designs presented here are the result of the Census Bureau statistical staff work. This early phase of the work limits our attention to 1990 content.

Each matrix sampling plan is defined by three sets of parameters, number of forms, number of data items per form, and the sampling fraction corresponding to each form. These three sets of parameters are the basis for the calculation of each design respondent burden. A "crude" measure of respondent burden is used to compare the matrix sampling plans to the 1990 census sample design. Section II.B discusses respondent burden in great detail.

The issue of reliability is discussed in detail in Section II.A. For the purpose of this work we assume a 20 percent sample except for the fifth design. Under this plan, 8 forms are used to collect data from the total population, each form used at a 12.5percent sampling rate. A second version of this design uses 16 forms, each form used at a 6.25

[1] Alfredo Navarro is Supervisory Mathematical Statistician in the Decennial Statistical Studies Division of the Census Bureau. Richard Griffin is now Chief Methodologist at Chilton Research Services. This paper reports the general results of research undertaken by Census Bureau Staff. The views are attributable to the authors and do not necessarily reflect those of the Census Bureau.

percent sampling rate.

Each design is described and assessed in Section III.B. A summary of the basic results of this work is included in Section IV. Research plans about census sample correlations and indirect (or model based) estimation of cross-tabulations are briefly discussed in section V.

II. Reliability and Respondent Burden

The use of a matrix sampling design provides an adequate and feasible way by which more data may be collected and more data needs can be met. This can be accomplished in order to minimize the reduction in the reliability of sample estimates by carefully designing the sample forms. A second consideration in the use of a matrix sampling plan is the reduction in respondent burden. These two issues are discussed next.

A. Coefficient of Variation

The issue of reliability of sample estimates is be discussed based on the concept of coefficient of variation, or simply the CV. The CV of an estimate is the ratio of the standard error to the expected value of the estimate. There is no specific rule to determine if a given CV is good or not. This determination is based on several considerations, use of the data, consequences of making the wrong decision, and so forth. In practice, a CV of 10 percent or less is often consider to be adequate, between 10 and 50 percent to be acceptable, and 50 percent or more to be not desirable. Assuming normality, a CV of 50 percent or more implies that the 95 percent confidence interval about an estimate includes zero. This is a highly undesirable situation for many possible uses of the data. For instance, the CV of an estimate for a 10 percent population characteristic in a tract with 2500 population sampled at 6 percent is about 24 percent. The CV of an estimate for a 25 percent housing characteristic in a place with 4000 housing units sampled at 12 percent is about 7.4 percent. The type and size values given are for illustration purposes. Note that for a given proportion (i.e., p-value) the CV decreases as the area size and sampling fraction increases. For example, keeping sampling fraction and area size fixed, the CV decreases as the item proportion increases.

If one wants to calculate the CV for an estimated proportion, say p=.13, for a place with 2689 population sampled at 25 percent, use the following formula.

$$CV(P) = \sqrt{\frac{(1-f)}{f} * \frac{(100-P\ )}{B * P}} * DE$$

P is the estimated proportion, B is the base of the percentage, and a 1-in-6 sampling fraction is assumed. In this example B is the total population. DE is commonly referred to as the design factor. For this example, the use of the formula results in a CV of about

$$CV(P) = \sqrt{\frac{3 * (100-13)}{2689 * 13}} \doteq 8.6\ \%$$

if the design factor, DE, is 1.0. One may want to calculate the CV for an estimate of a 20 percent characteristic for the 18 years old and above population, in this case B is defined as the size of the 18 years old and above population.

The actual census sample design is a systematic sample of housing units, not a simple random sample without replacement (SRSWOR). Thus, the census sample of persons is a systematic cluster sample, where the cluster is the housing unit. The design factor reflects the variance increment over the variance that would have been obtained if the more simple sampling procedure (i.e., SRSWOR) had been used.

In general, the use of sampling rates between 5 and 15 percent produces estimates with adequate to acceptable reliability for most census tabulation areas and census characteristics. Most census items do not fall into the "rare" category and therefore sampling rates as low as 5 percent would produce estimates with acceptable reliability, particularly for the larger tracts, counties and cities. The above discussion is for illustration purposes.

B. Respondent Burden

The concept of respondent burden is related to the time and effort a respondent has to use to complete a questionnaire for a given sample content, say 1990 census sample content. Time and effort are a function of the length and the nature of the individual items on a questionnaire. This is very important because it is reasonable to expect a somewhat "strong" correlation between respondent burden and quality of the data. The reduction in respondent burden might also have a positive impact on reducing item nonresponse rates, but more importantly on improving mail response rates and population coverage.

We will not make any attempt to qualify or quantify respondent burden by individual item. It is obvious that respondent burden varies from household to household. Factors such as household size and household composition (relationship) determine the real respondent burden. Since all the matrix sampling plans will be applied to the same universe it is expected that any difference in respondent burden should be attributed to either overall sample size or questionnaire composition. We will assume that all persons within a given household will provide responses for all the items asked in a given form and calculate the expected relative respondent burden for each design, relative to the 1990 respondent burden. The statement "the expected relative respondent burden for design 1 is 80 percent" means that for the universe as a whole the respondent burden of design 1 is about 80 percent of the respondent burden of the 1990 census sample design.

The 1990 census sample design respondent burden is calculated by multiplying the number of items (including 100 percent data subjects) by 16.7 percent, the overall 1990 sampling rate. Therefore the overall 1990 census respondent burden for the long form was about (56*.167)9.35. The designated sample size in 1990 was about 18 percent of the population. However, there was a significant sample loss due to item nonresponse. These cases are converted to short forms. To account for these cases, the Census Bureau uses a within class weight adjustment procedure. The initial weight of the respondents (the inverse of the observed sampling rate at the block group level) are adjusted proportionally to account for these cases. The 16.7 percent figure is the approximate observed sampling rate for the 1990 Census. Respondent burden for the matrix sampling plans are calculated in a similar fashion, adding up the individual respondent burden by form to get the overall design respondent burden. Respondent burden comparisons are related to the 1990 Census observed sampling rate. To get respondent burden measures relative to a single long form 20 percent sample, simply multiply the design relative respondent burden by .83.

III. Description of Matrix Sample Designs

A particular sampling plan in matrix sampling is defined by the number of sample forms, the number of items and the sampling rate associated with each form. For example, a matrix sampling plan is defined by the use of three long forms, each containing questions on 20 subjects and applied to 7 percent of the population for an overall sampling rate of 21 percent. Another matrix sampling plan could be defined by the use of 8 long forms systematically assigned to sample the total population, with each housing unit getting a sample questionnaire containing questions about 20 subjects. The content of the questionnaires may overlap, that is, a question may be included on two or more sample forms. Therefore, the sampling rate for an specific data item may be much larger than 12.5 percent.

During the development of the 1990 Census sample design almost all data users indicated a preference to maintain the 1980 small place level of reliability. To do this, all incorporated places with 2500 population or less were sampled at 1-in-2. We would probably want to maintain this feature for the Year 2000 Census. A significant portion of the 1-in-2 universe is in the list/enumerate type of enumeration area. To avoid operational problems inherent to the implementation of a matrix sampling design in L/E areas we might decide to exclude these areas from the matrix sampling universe. If the 1-in-2 sampling rate is maintained and matrix sampling is not implemented to sample small governmental units, the sampling rates associated with the individual questionnaires for each of the 5 matrix sampling plans will be proportionally lower.

The first and most difficult challenge was to allocate the 1990 content into reasonable sets of data items to define the various sample questionnaires or simply the long forms. For the development of these designs, it was decided to assume the 1990 content with the only modification that marital status (100 percent population item #6) and number of rooms in unit (100 percent housing unit item #3) were considered sample data items. It is quite certain that there will be (perhaps quite a few) content changes for the 2000 census aimed at meeting a demand for new data to satisfy current and future needs. However, it was felt that for this early phase of this work it was better to avoid speculation and start with a content that we all can relate to. The reliability of sample estimates, availability of data in general and cross-tabulations in particular, and respondent burden were among the criteria that guided the formation of data item sets, and decisions about the number of long forms and sampling rates. Note that major goals for using matrix sampling for the 2000 census are to not exceed the 1990 census overall respondent burden and to maximize the availability of data at all levels while keeping the level of reliability comparable (or adequate) to 1990. We want to achieve these goals with a total sample of no more than 20 percent.

A. 1990 Census Content

The 1990 long form asked all the questions to collect the data usually referred to as 100 percent data, and in addition asked more specific questions on socio-economic subjects. The long form contained 25 housing questions and 32 population questions. For content purposes P16 is not counted as an item, leaving only 31 population data items and a total of 56 housing and population questions. Of these, there were 19 housing and 25 population questions asked exclusively in the long form. The long form population data items are classified into two major groups, Social and Economic data items. Each of these groups consists of 13 data items (moving marital status to the sample).

The designs were developed giving special consideration to cross-tabulation of data items and sample size. Items that are required to be cross-tabulated were almost always placed on the same supplemental data set, design 5 is an exception. For example, Place of Work and Journey to Work are always in the same set. Another example is Industry and Occupation.

B. Matrix Sampling Designs

The core sampling rate of each of the first four designs is 20 percent. The fifth design stipulates that each housing unit in the universe gets a modified shorter version of the long form. Each data item is referred to as core or supplemental. Core data items, including the 1990 census 100 percent data items except for marital status and number of rooms in unit, are included in all long forms. The supplemental data items are the basis for the design of the various long forms for each matrix sampling plan. Three out of the five designs have a comprehensive form. A comprehensive long form asks all the questions of a sample large enough (2 - 5 percent of the population) to produce reliable estimates of cross-tabulations for medium to large size areas, such as cities or counties. Areas with 10000 population or more are under the large category. An estimate of a 10 percent characteristic for an area in this size category sampled at the smaller rate (2 percent) is acceptable according to our CV criteria.

Design 1 - ECONOMIC CORE

This design is referred to as the ECONOMIC CORE since all three forms include the economic data items except for **place of work** and **journey to work.** The sets of supplemental data items are denoted by Soc I (A), Soc II (B), Economic Core (C) and Housing (D).

For example, form 1 contains three modules or data item sets, Soc I, Soc II, and Economic Core. Module A consists of 8 data items. Form 1 contains 36 questions, including ten 100-percent questions. For instance, CITIZENSHIP (module A), will be collected from about 13.3 percent of the population. The economic data items will be asked of 20 percent of the population. For each of the items, labor force, journey to work, and disability, there are two questions asked in the long form.

Sample estimates of cross-tabulations AC, BC, and CD will be based on a 13.3 percent sample while cross-tabulations AB, AD, and BD will be based on a 6.7 percent sample. Cross-tabulations ABC, ACD, and BCD will also be based on a 6.7 percent sample. Note that these data are required for large places or counties for which a sample of that magnitude provides adequate reliability. For example, consider an area with 20000 population, the coefficient of variation for an estimate of a 10 percent characteristic based on a 6 percent sample is under 9 percent. Recall that a CV of less than 10 percent is considered adequate for most uses of census data. Assuming every person within a household will provide responses for all questions the expected relative respondent burden for this design is about (.067(36+48+48)/9.35) 95 percent of the 1990 census aggregate burden.

Design 2 - COMPREHENSIVE 2 PERCENT, NO CORE

This design is referred to as the COMPREHENSIVE 2% since one of the 4 forms asks all the questions of a 2 percent sample. There is no core for this design. The three sets of supplemental data items for this design are denoted Soc (A), Economic (B), and Housing (C).

Sample estimates of cross-tabulations AC, AB, BC will each be based on an 8 percent sample. Cross-tabulation ABC will be based only on a 2 percent sample, these data should be tabulated only for the larger areas, say areas with 20000 population or more.

The CV of an estimate of a 10 percent population characteristic for such an area sampled at 2 percent is less than 15 percent, which is considered acceptable. The respondent burden for this design is about ([.06(36+43+43) +.02(56)]/9.35) 90 percent of the 1990 census aggregate burden.

Design 3 - 1970 MATRIX SAMPLING DESIGN, COMPREHENSIVE 5 PERCENT

This design is referred to as 1970 matrix sampling design because a similar sampling plan was used for the 1970 census. It has a comprehensive 5 percent sample. This sample is the base for the production of reliable estimates of data cross-tabulated for large areas (10000 population or more). This design only has two long forms. There is no core for this design. The sets of supplemental data items are classified in 6 major groups; Soc I (A), Soc II (B), Econ I (C), Econ II (D), Hous I (E), and Hous II (F).

Estimates of any cross-tabulation of data will be based on at least a 5 percent sample. As indicated before, this sample size is large enough to produce acceptable estimates for most areas. Cross-tabulations of socio-economic and housing characteristics (A, C, and E) will be acceptable for all areas, regardless of size. The expected relative respondent burden for this design is ([.15(31)+.05(56)]/9.35)about 80 percent compared to the 1990 aggregate respondent burden. Keep in mind that a significant reduction in respondent burden is accompanied by a reduction in the per item sampling fraction, which directly affects the reliability of individual item estimates. For instance, data on disability is collected for 5 percent of the population only. For estimates of counties with 2500 population or less (119 or 3.8 percent), the CV of an estimate of a "rare" characteristic such as disability, (5 percent or less) starts to deteriorate (38 percent or less). This might not be a problem at all, if small areas (less than 2500 population) are sampled at 1-in-2, as discussed before in Section III.A.

Design 4 - COMPREHENSIVE 2 PERCENT, THREE SAMPLES

This matrix sampling plan differs from the previous one in two ways; the number of forms and the sampling rates. However the supplemental data sets are identical. The sets of supplemental data items are classified in 6 major groups; same as for Design 3.

Sample estimates for individual data items for smaller counties , such as disability, are better for this design than for design 3, however not by much. The sampling rate increased only by 1 percent (from 5 to 6 percent). Estimates of cross-tabulations, such as ABC, will have their reliability significantly reduced when compared to design 3. For example, the CV of an estimate of a 10 percent cross-tabulation from a 5 percent sample for a place of 2500 population is about 26 percent and about 42 percent if the population is sampled at 2 percent (see Table 3). The expected relative respondent burden is about ([.14(30) + .04(35) + .02(56)]/9.35) 72 percent compared to the 1990 aggregate burden. The reduction in respondent burden is realized due to a significant reduction of questionnaire's length.

Design 5 - "SHORTER" LONG FORM, 100 PERCENT SAMPLE

This design is unique in the sense that it employs 8 forms (or a maximum of 16 long forms), defines 8 supplemental sets and assign each housing unit in the universe to a 12.5 (or 6.25 percent sample if 16 forms are used) percent sample. Well and carefully designed forms optimize the use of sampling and the reliability of sample estimates. There are 6 sets of supplemental population data items and 2 sets of housing data items.

The item sampling rates of this design are high, perhaps too high considering respondent burden. For instance, the sampling rate for industry or occupation is 50 percent. A major problem with this design is the inability to produce required tabulations. For example, version 1 (8 forms) fails to produce 23 out of 84 required cross-tabulations. Version 2 (16 forms) fails to produce 3 out of the 84 required cross-tabulations. The major drawback of this design is the increase in respondent burden relative to 1990. The expected relative respondent burden for version 1 is ([.125(178)/[.167(56) + .833(12)]=22.25/19.35)about 115 percent while for version 2 is ([.0625(350)/19.35) about 113 percent compared to the 1990 aggregate respondent burden. Note that this comparison is being made relative to the 1990 census total respondent burden, including the short and long forms.

Estimates of cross-tabulations of any two modules, from version 1, are acceptable for most areas. However estimates based on version 2 are not acceptable for the smaller tracts and places, that is, areas with less than 1000 population (refer to population Table 2).

IV. Summary of Basic Results

Adequate estimates for a 1 percent population characteristic for a 2 percent sample are only obtained for small state and very large places and counties (500K population or more). For a 5 percent data item acceptable data are obtained for a small size tract (1250 population) for sampling rates over 5 percent. Sampling rates over 4 percent produce acceptable Cvs for a small size tract for a 10 percent population characteristic.

The table below summarizes the results on respondent burden (relative to 1990 and to a single form 20 percent sample) and reliability calculations for each of the design. Estimation and variance estimation will be more complex than under a one sample scenario. For example, Design 4 stipulates

4 samples. Each of the samples will have to be weighted and 4 sets of design factors will have to be produced.

	Reliability	Respondent Burden	
		1990 Design	20%
Design 1	Acceptable	95%	79%
Design 2	Acceptable	90%	75%
Design 3	Acceptable for most areas	80%	66%
Design 4	Acceptable for most areas	72%	60%
Design 5 Set 1	Adequate	115%	116 %
Design 5 Set 2	Adequate	113%	114 %

The last two figures in the 20 percent respondent burden column are slightly larger than the corresponding relative measures for the 1990design. However, note that the total respondent burden for a single form (10 short form questions only) 20 percent sample design is lower than the total 1990 respondent burden (19.2 vs. 19.35).

V. Additional Work

The results of some important small area estimation research will help to answer some of the concerns of small area data users if matrix sampling is used in 2000. The next three sections summarize research planned for the next few months.

A. Correlation Analysis - The next step in our matrix sampling research is a correlation analysis of the 1990 sample content. The identification of highly correlated items will help determine optimal groupings the 1990 content across an optimum number of forms. Exploratory data analysis methods will be used to identify and assess fundamental relationships between data items. Results from this work will be used to refine our imputation models and in our small area estimation research. For small area estimation, we are planning to investigate procedures proposed by Ericksen (1974) and discussed in Griffin and Navarro (1992). We hope to offset any loss in reliability of small area estimates due to the use of matrix sampling.

B. Simulation of Matrix Sampling - We will select one sampling plan and make estimates using 1990 Census data. This will allow us to assess the loss in accuracy for estimates across several tabulation areas. Ericksen's (1974) procedures for small area estimation will also be implemented using 1990 Census data to produce estimates to be compared to the 1990 sample estimates. It is possible that a "smaller" sample would produce more reliable estimates due to the "borrowed strength" from the empirical Bayes modeling of the census weights. The problem of variance estimation associated with small area estimates will be investigated later. Griffin and Navarro (1992) proposed several variance estimators.

C. Estimation of Cross-tabulation - Concurrently we will start research on estimating cross-tabulations for which no one respondent was asked all the items. We will simulate a model-based procedure to generate the full set of census sample data. Direct estimates of any cross-tabulation are available from 1990. The model-based estimates will be evaluated by comparing them to the 1990 direct estimates.

References

Ericksen, E.P. (1924), "A Regression Method for Estimation Population Changes for Local Areas", Journal of the American Statistical Association,, 69, pp 867-875.

Ericksen, E.P. and Kadane, J.B. (1985), "Estimating Population in a Census Year: 1980 and Beyond", Journal of the American Statistical Association, 80, pp 98-108.

Paass, G., (1989), "Stochastic Generation of a Synthetic Sample from Marginal Information", Census Annual Research Conference Proceedings, pp 431-445.

Griffin, R.A. and Navarro, A., (1992), "Survey Design and Estimation for Small Area Statistics from the Decennial Census Content Sample", International Conference on small Area Statistics and Survey Designs Proceedings.

Griffin, Deborah H. and Christopher L. Moriarity, 1990 Decennial Census Preliminary Research and Evaluation Memorandum Series No. 179, "Characteristics of Census Error." U.S. Department of Commerce, Bureau of the Census. September 15, 1992.