Bullion

Volume 36
Number 3 *Vol. 36 No 3 - Vol 37 No. 1 July 2012 - March 2013*

Article 11

3-2013

Design, implications and analysis of household survey: practical issues to note

O. O. Akanji
Central Bank of Nigeria

DESIGN, IMPLICATIONS AND ANALYSIS OF HOUSEHOLD SURVEYS: PRACTICAL ISSUES TO NOTE

DR. (MRS.) O. O. AKANJI
Central Bank of Nigeria

1.0 INTRODUCTION

Household surveys are conducted using complex sample designs and these sample designs are generated from enumerated units. In the analysis of the household surveys, it is necessary to build the weights, establish variances of the survey estimates in a manner that will reflect the complex sample design.

Household survey utilizes complex sample designs to control survey costs. In addition, complete sampling frames that list all individuals or households in the enumerating area are usually not available. Even when population registries are available the cost of implementing a household interview survey based on a simple random sample design would be prohibitively high.

A typical household survey design structure starts with the features of stratification, and then stages of sampling units, up to the observational units. These features have their implications which range from standard errors of estimates and/or the disproportionate sampling to the need to impose more than one analytic process for special purpose analyses (Table 1).

Most sample designs for household surveys use complex sample designs involving stratification, multi-stage sampling, and unequal sampling rates. Consequently, weights are needed in the analysis to compensate for unequal sampling rates and adjustments for non-response which could lead to more unequal weighting. The complex sample design needs to be taken into account in estimating the precision of survey estimates.

After the sample design, the development of weights is the next stage for the production of simple "descriptive" estimates such as the totals, means and proportions/percentages that are widely presented in survey reports.

TABLE 1

A Typical Household Survey Design Structure

Features	Possible Definition	Implications
Strata	States or Community Type (Urban versus Rural)	May reduce standard errors of estimates. Control distribution of sample may lead to disproportionate sampling
First-stage sampling units	Census enumeration areas or similar geographical areas. Villages in rural strata	Facilitate clustering of the sample to control costs. Facilitate development of complete frames of housing unit addresses only in sampled areas. Selected with probability proportional to size.
Second- stage sampling units	Housing unit addresses	May contain none, one, or more than one household or unrelated person. Selected with equal probability within first-stage sampling units
Third-stage sampling units (when not all household members are automatically included in the sample)	Household members	Sample selected from roster of household members obtained from a responsible adult household member. May lead to unequal weighting in order to account for household size.
Observational units	Households. Household members. Agricultural or business enterprises operated by the household members. Special files for sub-groups e.g., adults in the work force. Events or episodes pertaining to household members. Repeated measures over time (panel surveys)	May require more than one analytic file for special purpose analysis

Source: Federal Office of Statistics (Now National Bureau of Statistics) Household Survey Framework

This paper outlines the development of weights and their use in computing survey estimates and provides a general discussion of variance estimation for survey data. The paper deals with what are termed "descriptive" estimates, such as the totals, means, and proportions which are widely used in survey reports. The paper further discusses the three forms of "analytic" uses of survey data that can be used to examine relationship between survey variables, namely multiple linear regression models, logistic regression models and multi-level models. These models form a set of valuable tools for analyzing the relationships between a key response variable and a number of other factors.

This paper is to enhance the analytic processes of survey data as generated in the established household survey framework for economic policy management and decision making. The paper is divided into four sections. Section 1 is the introduction while Section 2 is the descriptive statistics. Section 3 discusses analytic statistics emphasizing the models that bring out the relationship between the surveys variables. Section 4 concludes and proffers recommendations.

2.0 DESCRIPTIVE STATISTICS: WEIGHTS AND VARIANCE ESTIMATION

Household surveys are commonly designed to produce estimates of population totals, population means, or simple ratios of totals of means. Example of totals might be total population, men in the work force, women in the work force, or the number of children ten years old or younger. Example of means might be average income for persons in the work force, average income of women in the work force, and average income of men in the work force. Ratio estimates might be required to estimate the proportion of households with total income below the poverty level or the average households whose principal wage earner is a female. Household surveys are designed to produce national estimates. However, the survey could be designed to give estimates for geo-political regions or for cross-sectional domains. Household surveys may be repeated at regular intervals to obtain periodic estimates (e.g., annual or five year estimates). The statistics obtained from these surveys are "descriptive" statistics. Descriptive statistics include the estimates themselves as well as some measure of the precision of these estimates.

Descriptive reports may include standard errors of estimates or interval estimates based on those standard errors. Estimation of the standard errors requires an analysis that takes account of the household survey sample design. These types of fairly simple descriptive statistics constitute the majority of the official statistics that describes the household surveys. Example is the National Bureau of Statistics (NBS) National Household Survey used in the computation of Consumer Price Index.

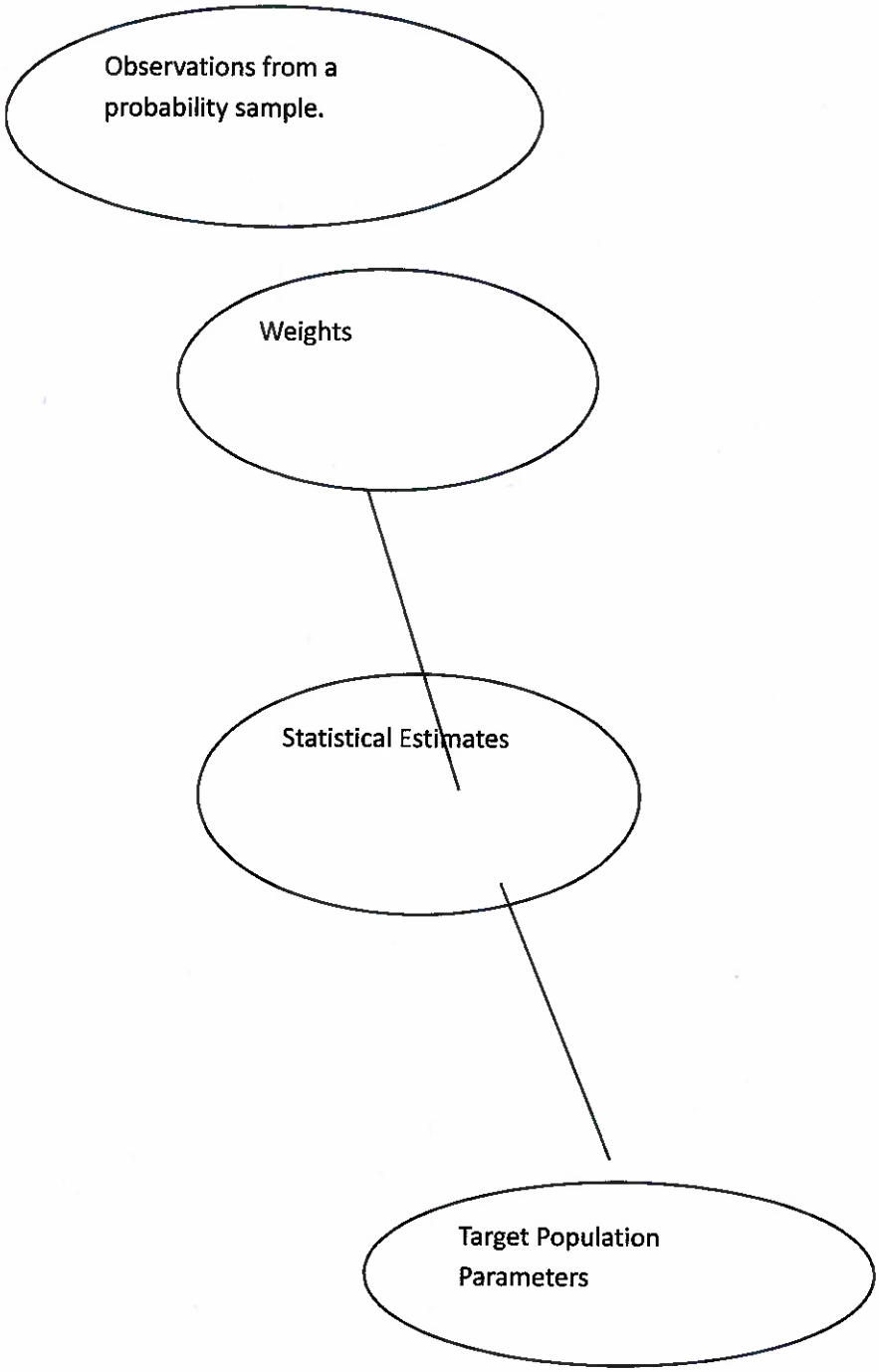
In conducting Household surveys, survey weights are developed. The weights are designed-based weight that will assist in the adjustment for non-responses, post-stratification problems/errors, field data, and collection errors. The term "survey weights" is used to differentiate them from strict "design-based weights". These weights provide the link between the observations from a probability sample of households and summary measures or population parameters about the household population.

The figure 1 below illustrates the link. The population of all households is sometimes called the target population or the universe. Without the application of both probability sampling and weighting, there is no supporting statistical theory to provide a link between the sample observations and the target population parameters.

Consequently, any analysis that ignores the sample design and the weights must be based on assumptions. If the sample is designed to generate equal probability sample, then the weights for estimating means, rates or relationships among variables may be safely ignored. Kish (1965) called these designs "epsem" designs and noted that even complex multi-stage samples can be designed to be "epsem" for sampling units at the final or near final stage of the design. It should be noted that adjustments for non-response may create unequal weights even if the design was initially "epsem". Also, if post-stratification or multi-dimensional calibration is applied to the data through adjustments to the weights, these processes will always create unequal weights adjustments and therefore, unequal weights.

Some analysts are, however, willing to make the assumptions that would allow analysis of household survey data without weights or with equal weights. These assumptions are most tenable when applying models to the data to study relationships between a dependent variable and a number of independent explanatory variables, Horvitz and Thomas (1952) indicate that for the theoretical case of surveys with complete response from all sample members, the use of designed-based weights computed as the inverse of each observational unit's probability of selection provides for unbiased estimates of population totals and other linear statistics. However, in practice, household surveys always encounter some non-response, which can lead to bias estimates. Techniques have been developed which attempt to reduce the bias due to non-response.

Fig 1 Application of Weights and Statistical Estimation



The simplest approach involves partitioning the sample into weighting classes so that within these classes the differences between the population parameters for respondents and non-respondents are believed to be much smaller or to be ignorable (Rubin 1987). Ratio adjustments to the weights are then performed within the weighting classes so that each class is represented in the adjusted estimates in the same proportion as it would have been represented in the selected sample.

The process of probability sampling does not necessarily guarantee that the selected sample's distribution on known characteristics will be identical to that of the total population. Stratification before sample selection can ensure this condition hold for some characteristics, but may not be possible for others if the classification variable is not available on the frame used to select the sample. Rather than conducting complex ratio adjustments for each estimates produced from the household survey data, post- stratification is

often incorporated as a one- time weight adjustment, which then automatically applies to all estimates produced using the adjusted weights. The simplest approach to post-stratification adjustment uses a partitioning of the sample similar to that used for weighting class non-response adjustment.

The final weights attached to an analytic file produced from household survey may contain the following factors:

- The design-based weight computed as the reciprocal of the overall probability of selection;
- A non-response adjustment factor;
- A post-stratification adjustment factor;
- A weight-trimming factor.

These factors should be documented so that any analyst can review them. The adjustment factors applied to the initial design-based weights involve some subjective and sometime arbitrary judgments in the definition of weighting classes, the selection of control totals for post-stratification adjustment, and in the extent of weight trimming applied to control the design effect. When unexpected results or apparent anomalies occur in the survey estimates, it is not uncommon to thoroughly review the weighting process as well as all other aspects of the total survey design and implementation.

In general the analytic uses of household survey data provide special challenges due to complex survey designs which include the use of weights and design structure (Skinner, Holt and Smith, 1989; Korn and Graubard, 2003; and Chamber and Skinner, 2003). Even simple statistics such as means become non-linear in complex surveys. To estimate a population mean from a complex survey, it is necessary to estimate a population total for the variable interest, say family income, and to estimate the size of the population, say total number of families. The mean is then estimated as the ratio of the two estimates. Mean family income would be estimated as:

$$\text{Estimate of mean family income} = \frac{\text{Estimate of total family income}}{\text{Estimate of total number of families}}$$

This estimated mean turns out to be a non-linear function (a ratio) of two linear statistics. In complex surveys, the sample size (number of observations of a particular type) is itself a random variable. These types of no-linear estimates are not unbiased for small samples, but are consistent in the trivial sense that if the sample size were increased to the finite population size, the non-linear estimate would exactly equal the comparable finite population value (Cochran, 1977).

3.0 ANALYTICAL STATISTICS

Having discussed the simple descriptive statistics, this section discusses the analytic statistics. This is the statistics that examine the relationships among variables. In fact, the moment data users wish to compare estimates among domains, the nature of the required statistics becomes "analytic". Simple analytic statistics may be based on differences among domains, e.g., a comparison of the proportion of households with total income below the poverty level in two geo-political subdivisions or a comparison of crop production over the last two years. Sometimes the estimates in a simple comparison are independent of one another so that the standard error of the difference can be determined strictly from the standard error of the individual estimates. Under these circumstances, the standard error of the estimated difference between two domain means can be derived as:

$$se(\bar{y}_1' - \bar{y}_2) = \sqrt{\{se(\bar{y}_1)\}^2 + \{se(\bar{y}_2)\}^2}$$

This formula for the standard error of a difference assumes that the two estimates are independent and, as a result, their estimates are uncorrelated. This form of the standard error of differences is convenient for data users, because they can derive the standard error of a difference from published standard errors of the individual estimates. However, with complex sample designs, domain estimates are often correlated. The variance of the difference of two domain estimates then includes a covariance term:

$$Se(\bar{y}_1 - \bar{y}) = \sqrt{\{Se(\bar{y}_1)\}^2 + \{se(\bar{y}_2)\}^2 - 2\,\mathrm{cov}(\bar{y}_1, \bar{y}_2)}$$

The covariance term is generally positive, and hence it leads to a lower standard error of the difference estimate than the independent case discussed above. Household surveys can be designed to take advantage of the covariance term in the standard errors of estimates of differences; longitudinal panel surveys achieve a high positive covariance among annual estimates by utilizing a common, continuing sample of individuals or households. Because the standard error of the difference cannot be derived from the published standard errors of the individual estimates, it becomes necessary to anticipate what comparisons are of greatest interest and to publish their standard errors also.

For strictly descriptive statistics about finite populations, the standard error of descriptive estimates is correctly reduced by the application of a finite population correction factor. In the simplest case of simple random sampling, the finite population correction factor is:

$$fpc = 1 - n/N$$

Where n is the sample size and N is the population size. If the purpose of the analysis is analytic, then, even in the simplest case of statistical significance of the observed difference between two domain means, the use of the finite population correction factor is inappropriate (Cochran, 1977, pp. 34-35). This is because the form of the statistical significance test requires one to hypothesize whether both domain populations could have arisen from a common infinite hypothetical population (a single super-population).

3.1 Linear regression models.

For the purposes of discussing linear and logistic regression models, it is convenient to assume that sampling is "with replacement" at the first stage. We can further assume that the analytic file of observation data includes index variables for strata, designated by h, and for primary sampling units (PSUs), designated by i. Additional structure variables do not need to be identified when we are willing to use the "with replacement" design assumption at the first stage of sample selection as discussed in section II above. The full implications of using a complex household sample design are incorporated into the estimates of model coefficients and their standard errors only if we use a statistical package that properly accounts for the household survey design including the analytic weights and the design structure (strata and PSUs). When we discuss multi-level models, the focus will change to incorporating the design structure into the model and the analysis will permit estimation of effects related to the structure variables.

A linear regression model that involves one continuous explanatory variable and one categorical explanatory variable can be expressed as:

$$y_{hij} = \alpha x_0 + \beta x_{1hij} + \sum_{d=1}^{D} \gamma_d x_{2dhij} + \varepsilon_{hij}$$

In this model, observations are represented by the observed dependent variable, Yhij; which is an intercept variable, x0, always set to 1; an observed continuous explanatory variable X1hij, and a set of indicator variables X2dhij, defining D levels of categorical variable. The regression model parameter α, β, and γd (d= 1, 2,—— D) are termed regression coefficients and are estimated by analysis. The final term in the model is the error term and measures the deviation from the model associated with the j-th observation associated with the i-th PSU of the h-th stratum. This is a main effects model, since it contains no interaction effects

3.2 Logistic regression models

When the dependent variable is categorical, linear regression approaches do not apply. Although multinomial modeling procedures are available, we will be discussing only the binary (two-level) categorical variables that can be analyzed using logistic regression models. In this sense, logistic regression is a special simpler case of multinomial regression.

For a two-category or binary dependent variable coded as 0 and 1, linear regression approaches will work but can produce predicted values outside the range of 0 to 1. Linear regression might be used as a preliminary step with a binary dependent variable to identify explanatory variables that are good predictors of the dependent variable. A logistic regression model that involves one continuous explanatory variable and one categorical explanatory variable can be expressed as:

$$\text{Log}\left[\frac{p(x_{hij})}{1-p(x_{hij})}\right] = \alpha x_0 + \beta_{1x1hij} + \sum_{d=1}^{D} \gamma_d x_{2dhij} + \varepsilon_{hij}$$

Except for the dependent variable, the terms in the logistic model are defined the same way as in the linear regression model. To understand the logistic transformation, consider an example where $p(x_{hij})$ is a function of the explanatory variable; designate it by p for convenience. Further assume that p is the probability that a household with a given set of values for the explanatory variables has an income level below the established poverty level. Then p/(1-p) is called the odds of being in poverty, and log(p/1-p)) is the log odds of p, sometimes called logit(p). The logistic regression model tries to relate the log odds of p to the x's. The observations are single households where we observe not the probability of being in poverty, but the actual current status; in poverty or not in poverty. Also, since the dependent variable is a log odds of p, each parameter [α,β1, and γd(d = 1,——, D) is also on the log odds of p scale; furthermore, the relationship between the log odds of p and the x's is assumed to be linear.

3.3 Use of multi-level models

We now discuss multi-level modeling. In this case we must recognize the survey data structure especially structure imposed by surveys that are designed to be multi-stage. For example agro-ecological regions in a country may form strata, and from each, a number of administrative units may be selected. The latter will form the primary sampling units. Secondary units are then selected from each primary unit; subsequent units are selected from the secondary units and so on. This leads to a hierarchical data structure. It can involve the use of stratification variables at one or more of the levels.

For example, a survey concerning farming household in a region may involve using the administrative divisions of the

region as primary units, then choosing villages from each division and then selecting households from each village, perhaps ensuring that different wealth categories of households are included. It is pertinent to pay attention to the different sources of variability in the data collected at the household level. It incorporates variation between the administrative divisions, variations between villages, and variation between households within villages. Often data are also collected at each level of the hierarchy, here at the household level, at the village level and at the administrative level. It is the important to recognize and note which variables are measured at the village level (e.g., existence of an extension officer, government subsidies for fertilizer), and which are measured at the household level (e.g., socio-economic characteristics of the household).

Multi-level modeling is the key statistical technique of relevance in designing household surveys. This modeling approach (Goldstein, 2003, 1993; Snijder and Bosker, 1999; Kreft and Leeuw, 1998) is desirable because it allows relationships across and within hierarchical levels of a multi-stage design to be explored, taking account of the variability at different levels. Inter-correlations between variables at the same level are also taken into account. It is worth highlighting briefly at this junction, the consequences of ignoring the hierarchical structure. This happens when the data are aggregated to a higher level or disaggregated to a lower level. If the analysis is relevant and it is required only at one level, there is no problem. However, care must then be taken that any inferences are made only at that level. It will not be possible to make references about one particular level of the hierarchy from data analysed at another level. Thus an analysis ignoring the hierarchy will not permit cross-level effects to be explored. Another difficulty arises if data is analysed at its lowest level by regarding the higher-level units as a factor in the analysis. This is inefficient because it does not allow conclusions to be generalized to all higher-level units in the population- they will only apply to the sampled units.

We use the CBN/FOS/ NISER Informal Study of 2001 to illustrate a scenario where the use of multi-level modeling can be beneficial in exploring relationships. The study looked at factors contributing to informal sector operations in 36 states of the Federation across 5 major activities chosen and 200 interviews conducted using Household survey named NISH with min-focus groups comprising 5 to 10 respondents from different households. Within the 36 states enumerating areas were created such that these areas had influences on informal activities. The study also collected 'perceptions of what constituted informal activities' using Logit analysis to obtain the odd and cross referencing analysis.

It is important however to note that this survey used non-probability sampling, it may therefore be argued that any analytical conclusions may not be generalizable to any clearly defined target population. However, the data obtained at the focus group level were analysed using multi- level model. Therefore to enable results to be generalized across all the states, a random variable that was derived using the conceptual definition of the informal sector. The Focus groups within enumerated areas also entered the model as a random effect- thus bringing the flavor and the essence of multi-level modeling. Such models also allow interactions among states level variables and variables at the focus group level to be engaged.

3.4 Modeling to support survey processes.

Even when a household survey is used strictly to provide descriptive statistics, there may be need for modeling to support other survey processes. Adjustments for non-response are often based directly on statistical models; Groves et al (2002, pp. 197-443) discuss a variety of methods for accounting for non-response all of which must assume some statistical model. Logistic regression models may be used to develop predicted response propensities for the purpose of non-response adjustment or to identify weighting classes based on similar response propensities. Predictive statistical models may also be used as part of the procedure for imputing missing data (e.g., Singh, Grau and Folsom, 2002). Finally, statistical models can be used to evaluate methodological experiments embedded in surveys (e.g. Hughes et al. 2002).

4.0 CONCLUSIONS

The aim in doing this paper is to discuss issues that are involved in the analysis of survey data. These issues include the use of survey weights and of appropriate variance estimation methods with both descriptive and analytic uses of survey data. The paper also provides an overview of practical situations where modeling techniques have a role to play in survey data analysis. They are useful tools but their application requires careful thought and attention to their underlying assumptions.

In this paper, we discussed the role of survey weights and recognition of the sample structure in developing both descriptive and analytic statistics from survey data. Survey data analysis software that use survey weights and take account of sample structure may be used to estimate the parameters of both linear and logistic regression models based on survey data. The estimates based on the sample are estimates of what would be obtained from fitting the models to the entire finite population.

Furthermore, standard errors of the estimates can also be obtained. The explanatory variables in regression models applied to survey data are almost always observed as they exist in the population rather randomly assigned according to some experimental design. Analyst need to be clear that regression coefficients based on survey data simply reflect relationships that exist between the dependent variable and the explanatory variables in the population and do not necessarily imply causation. We also discussed how the parameters of regression and logistic regression models relate to simple descriptive statistics and how they may be interpreted for some relatively simple models.

Multi-level modeling as discussed in this paper is an "advanced" technique which is best carried out in consultation with a statistician familiar with the use and limitations of the technique. At present multi-level models appear to be rarely used in analyzing surveys except the use of National Integrated Survey of the Household (NISH), but their use is highly desirable for the insights they can provide concerning inter-relationships between variables at different levels and their ability to take account of variability amongst sampling units at different levels in a multi-stage design.

The paper has shown that the formulation of multi-level models is not too difficult for those who are familiar with the application of general linear models (GLM), but again, we should note that there are assumptions associated with the models that must be checked by carrying out residual analyses as in the case of GLMs. The multi-level modeling approach can also be undertaken and care needed in deciding which effects are random and which are fixed and this model specification will help in answering specific survey objectives

In conclusion, all the statistical techniques do have their various limitations; and there is the need to acquire the software that can recognize the sampling designs. The paper offered some modeling techniques that can serve as useful tools for survey data analysis. It is recommended for survey analysts and researchers as it helps to obtain more information from expensive survey for data collection especially through household surveys.

References

CBN/FOS/NISER, 2001. A Study of Nigeria's Informal Sector. Volume 1, Statistics on Nigeria's Informal Sector.

Chambers, R. L., and C. J. Skinnet, 2003. Analysis of Survey Data. Wiley, Chichester, UK

Cochran, W. G., 1977. Sampling Technique. Third Edition. New York: John Wiley & Sons

Deville, J.C. and C.E. Sarnal, 1992. "Calibration Estimating in Survey Sampling." Journal of the American Statistical Association 87: 376-382.

Folsom, Ralph E., Jr., 1991. "Exponential and Logistic Weight Adjustments for Sampling and Non-response Error Reduction." Pp. 376-382 in Proceedings of the Social Statistics Section, American Statistical Association.

Folsom, Ralph E. and Michael B. Witt, 1994. "Testing a New Attrition Non-response Adjustment Method for SIPP." Pp 428-433 in American Statistical Association, Section on Survey Research Methods.

Folsom, R. E. and A.C. Singh, 2000. "The General Exponential Model for Sampling Weight Calibration for Extreme Values, Non-response, and Post-stratification." In Proceedings of the Survey Research Methods Section, American Statistical Association. Indianapolis, Indiana.

Goldstein, H. 2003. Multi-level Statistical Models. 3rd Edition. Arnold, London

Goldstein, H., J. Rasbash, M. Yang, G. Woodhouse, H. Pan, D. Nuttall and S. Thomas, 1993. A multi-level analysis of school examination results. Oxford review of education, 19, 425-33

Graubard, B. I. and E. L., Korn, 2002. "Inferences for super-population....." Statistical Science 17: 73-96

Groves, Robert M., Don A. Dillman, John. I., Eltinge and Roderick J.A. Little, 2002. "Survey Non-response" New York, NY: John Wiley & Sons, Inc.

Horvitz, D. G., and D. J. Thompson, 1952." A generalization of sampling without replacement from a finite universe." The Journal of the American Statistical Association" 47: 663-685.

Kish, Leslie, 1965. Survey Sampling, New York: John Wiley & Sons.

Kreft, I. and J. de Leeuw, 1998. Introducing Multi-level Modeling. Sage, London.

Korn, E. L., and B.I. Graubard, 2003. Estimating variance components by using survey data. Journal of the Royal Statistical Society B, 66, 175-190.

Redesigning an ongoing National Household Survey: Methodological Issues. DHHS Publication No SMA 03-3768, edited by Gfroerer, J., Eyerman, J. and Chromy, J., Rockville, Maryland, USA.

Pfeffermann, D., C, J. Skinner, D. J Holmes, H. Goldstein and Rasbash, 1998. "Weighting for unequal selection probabilities in multi-level models. Journal of the Royal Statistical Society B, 80, 23-40

Rubin, Donald. B., 1987> Multiple Imputation for Non-response in Surveys. New York, NY: John Wiley & Sons

Singh, Avinash et al 2002 in DHHS Publication No SMA 03-3768

Skinner, C.J., D. Holt and T. M. F. Smith. Editors 1989. Analysis of Complex Surveys. Wiley, New York.

Snijder, T. A. B. and R.J. Bosker, R.J. 1999. Multi-level Analysis: An Introduction to Basic and Advanced Multi-level Modeling. Sage, London.

Woodruff, R.S., 1971. "A simple method for approximating the variance of a complicated sample" Journal of the American Statistical Association 66: 411-414.