Generations and Gender Survey Sample Design Guidelines

Study Design group

M. Simard, Statistics Canada

An Overview of the Generations and Gender Programme

The purpose of the Generations and Gender Programme is summarized in a March 30, 2002 document written by Miroslav Macura, Chief, Population Activities Unit, United Nations Economic Commission for Europe (PAU, UN/ECE):

The Generations and Gender Programme (GGP) is promoted and co-ordinated by the UN/ECE with the financial backing of the United Nations Population Fund. The goal of GGP is a cross-national, comparative, multi-disciplinary, longitudinal study of the dynamics of the family and family relationships in the contemporary industrialized countries, in particular in Europe and North America. The specific aim is to improve the understanding of factors – including public policy and programme interventions – affecting the evolution of two principal family relationships: child-parent relationships and partner-partner relationships. The GGP will scrutinize the start, development and end of those relationships of men and women of all age groups from adolescence to old age. The implication of this wide age range is that child-parent relationships will be considered irrespective of whether the child or the parent is young, middle-aged or old.

The GGP is being developed around a prospective longitudinal survey, the Generations and Gender Survey (GGS), which, wherever possible, will continue for a decade or longer, making it possible to study the changes of families and family relationships as these occur. A longitudinal survey has been proposed since it is better than a cross-sectional survey at measuring time-dependent variables, such as incomes and opinions.

The survey will have approximately equal numbers of men and women within the age range 18-79 as respondents. The GGS will request information from these individuals about their current child-parent and partner-partner relationships. The survey will also collect data on the individuals' mezzo contexts, which are their households, families and social support networks. The final effective sample size, i.e. after the three waves will ideally be of the order of 8,000[1] or more respondents, who would be interviewed at dates three years apart over a period of nine or more years.

Countries that have rich administrative record data may opt to draw on these data and complement them with smaller GGS surveys, which will mainly provide information required by the Programme that is unavailable in administrative records.

[1] Miroslav's document refers to 10,000 respondents, however, an e-mail sent on April 23, 2003 by Jan Hoem (Max Planck Institute for Demographic Research) to Statistics Canada refers to 8,000 respondents.

1. Proposed GGS Analysis

The sample design guidelines factor in the above requirements plus the type of analysis that is to be performed. GGS will be used to perform both cross-sectional and longitudinal analyses.

Jan Hoem and Andres Vikat of the Max Planck Institute for Demographic Research provided the following descriptions of the longitudinal analyses. Of particular interest is the analysis of individual-level demographic event-history data:

> ‘The central objects of interest in such a survey are risks of event-occurrence (or synonymously hazards or intensities) and the patterns of their dependence on fixed and time-varying covariates, most of which are also collected in the survey, but some of which come from the contextual-data base. Typical events in a demographic survey are births (of various orders), moves out of the parental home, the formation and dissolution of marital and non-marital unions, and the like. Risk dependencies are studied by means of hazard-regression methods…Serious additional consideration needs to be given to the role of multilevel modeling when the object of analysis is a set of event histories rather than a set of independent observations for a linear-regression model.
>
> ‘The specification that each national GGS should have a NET sample of some 8,000 respondents is based (1) on our extensive experience with the analysis of FFS and other data, and (2) on the perception that the GGS cannot do with less than some 2,000 men and 2,000 women at reproductive ages, and with similar numbers at post-reproductive ages. Please note that in most countries the FFS had some 2,500 to 6,000 female respondents (all at reproductive ages), and most had about half as many men (Festy and Prioux, 2002, Table 5). It has been much harder to get sharp results for men than for women in the FFS analyses. The FFS for the Czech Republic only had some 1,700 female respondents (and 721 men), and sharp results are correspondingly hard to come by, at least for life segments after initial adulthood.’[2]

And from Andres:

> ‘The main purpose of the GGS is to understand behavioral processes, and to arrive at estimates like relative birth risks (=intensities) by characteristics like educational level, type of activity, income level (in some broad categories), type of partnership (marriage/ cohabitation/ non-residential partnership/ no partnership), duration of partnership, household type, dwelling type, presence of children from a previous partnership, and numerous others, many of which change over an individual's life course (and therefore constitute time-varying covariates). The corresponding parameters would be estimated in multivariate intensity regression models… The multiple levels inherent in these models just refer to different kinds of episodes within a life course of one individual. Considerations for territorial units in a multilevel analysis of regional variation in demographic behavior come on top of that, and are linked with the way contextual data is going to be collected.’[3]

[2] Letter to Sange de Silva, Statistics Canada, from Jan Hoem, Max Planck Institute, on March 20, 2003.
[3] e-mail to Sarah Franklin, Statistics Canada, from Andres Vikat, Max Planck Institute, on April 29, 2003.

2. Survey Design Guidelines

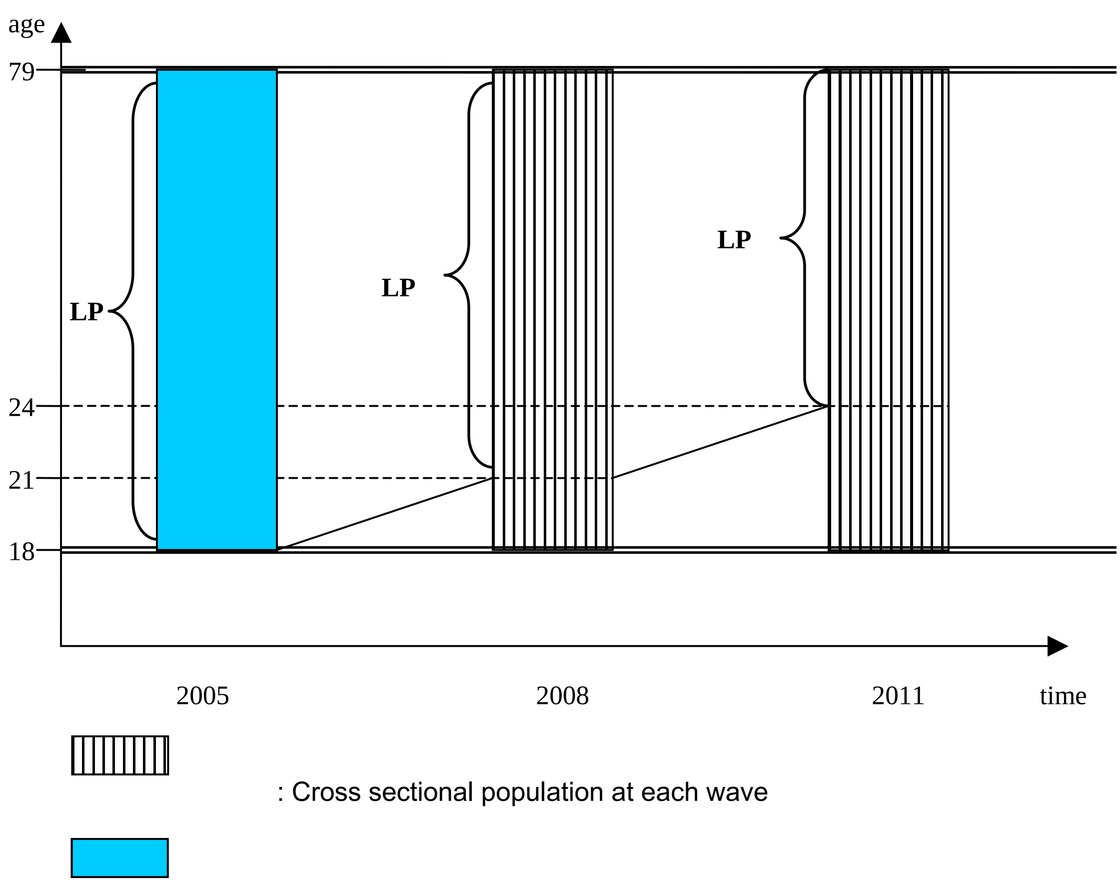
Because GGS is an international survey where international comparisons are to be done and the survey results have the potential to impact on government policy it is important that care be taken in the design of its sample. It is essential that, in each participating country, the survey agency responsible for the sample design and implementation be a reputable firm that is willing and able to conform to the GGS guidelines.

Each country will ultimately determine which sample design is most appropriate for GGS given the most suitable survey frame and the method and cost of survey collection (the GGS questionnaire is designed for face-to-face interviews). STC recommends that the following guidelines be used when selecting a survey frame and sample design:

2.1 GGS target population: the resident non-institutionalized national population aged 18-79

The target population is the population for which information is desired. Note that there is only one longitudinal population -- the specific population established at a given time point and followed through time. Thus, the target longitudinal population is the resident non-institutionalized national population aged 18-79 selected at wave 1 (e.g., in 2005). But there are as many cross-sectional populations as there are waves of the survey, since each cross-sectional population relates to a different point in time. For GGS, at each wave the cross-sectional target population is the resident non-institutionalized national population aged 18-79 for a specific reference year (e.g., 2005, 2008, 2011).

Diagram 1: Conceptual difference between cross-sectional and longitudinal population



: Longitudinal population (LP)

2.2 GGS survey population: may exclude up to 5% of the target population

The survey population is the population actually covered by the frame and surveyed for GGS. Often, exclusions are due to frame limitations or practical constraints – such as eliminating remote regions where survey collection would be prohibitively expensive. In order to facilitate international comparisons, STC recommends that each country minimize as much as possible exclusions from the target population. Any country that excludes more than 5% of the target population must provide valid reasons for the proposed exclusions.

2.3 Survey frame: list versus area frame

The survey frame provides the means of identifying and contacting the units of the survey population. There are two main categories of frames: list and area frames.

List frame

A list frame is a physical list of all units in the survey population (for example, a list generated from a population register). For GGS, in order to satisfy the sample design guidelines, any list frame of residents must include, for each person, the following auxiliary variables: design information: age, sex and place of residence (i.e., geography) and contact or tracing information such as name, phone and/or addresses. Other auxiliary information is also desirable in order to perform nonresponse analysis and weight adjustments (i.e., socio-demographic information such as level of education, income, size of household, etc.).

If administrative data are used to create a list frame, note that the usefulness of the administrative data depends on such criteria as the data's:

- concepts and definitions (they should be consistent with GGS),
- coverage of the target population (at least 95% coverage),
- quality of the data,
- timeliness with which the data are updated,
- reliability of the administrative source,
- privacy issues,
- ease of use of the data.

Area frame

An area frame is a special kind of list frame where the units on the frame are geographical areas. The survey population is located within these geographic areas. Area frames may be used when an adequate list frame is unavailable, in which case the area frame can be used as a vehicle for creating a list frame.

Area frames are usually made up of a hierarchy of geographical units. Frame units at one level can be subdivided to form the units at the next level. Large geographic areas like provinces may be composed of districts or municipalities with each of these further divided into smaller areas, such as city blocks. In the smallest sampled geographical areas, the population may be listed in order to sample units within this area.

Sampling from an area frame is often performed in several stages. For example, suppose that a country does not have a good quality, up-to-date list frame of residents from which to draw the GGS sample. An area frame could be used to create an up-to-date list of households as follows: at the first stage of sampling, geographic areas are sampled, for example, districts. Then, for each selected district, a list frame is built by listing all the households in the sampled district. At the second stage of sampling, a sample of households is then selected. At the third stage of sampling, an individual within a household is selected.

It is important that the geographical units to be sampled on an area frame be uniquely identifiable on a map and that their boundaries be easily identifiable by the interviewers. For this reason city blocks, main roads and rivers are often use to delineate the boundaries of geographical units on an area frame.

Each country will determine the most appropriate survey frame for GGS. If possible, it is recommended that a list frame be used since sampling for GGS from an area frame will be considerably more complicated. However, countries with the appropriate infrastructure already in place can decide to use area frame. Multi-stage sampling will then be used. Note that the control of a targeted sample size of individuals may be more difficult to achieve directly as the sampling unit of the first stage will most likely be geographical area.

When selecting the best frame for GGS, each country should try to minimize the following four types of frame defects:

i. Undercoverage: exclusions from the frame of some units that are part of the target population (e.g., a population register or census data may be out-of-date).

ii. Overcoverage: inclusions on the frame of some units that are not part of the target population. This is often due to a time lag in the processing of frame data (e.g., a population register may include some dead individuals who have not been identified as such).

iii. Duplication: an individual or household may appear several times on the frame.

iv. Misclassification: an individual or household may be misclassified (e.g., a man may be misclassified as a women or a person's age may be incorrect).

2.3.1 Tips and Guidelines

In order to choose and make the best use of the frame, the following tips and guidelines are useful:

i. When deciding which frame to use, assess different possible frames at the planning stage of the survey for their suitability and quality.

ii. Avoid using multiple frames, whenever possible. However when no single existing frame is adequate, consider multiple frame.

iii. Use the same frame for surveys with the same target population. If one country already conducts household surveys and have rotated out panels representative of GGS target population, this may be a very suitable and practical option.

iv. Incorporate procedures to eliminate duplication and to update for births, deaths and out-of scope units and change any other frame information in order to improve and/or maintain the level of quality of the frame.

v. Incorporate frame updates in the timeliness manner possible. Determine and monitor coverage of administrative sources through contact with the source manager.

vi. Emphasize the importance of coverage and implement effective quality assurance procedures on frame-related activities. Monitor the quality of the frame periodically by matching alternate sources and by verifying information during data collection.

vii. Implement map checks for area frames, through field checks or by using other map sources, to ensure clear and non-over-lapping delineation of the geographical area used in the sampling design.

2.4 Perform probability sampling

STC recommends that a probability sample be selected. Probability sampling is a method of sampling that allows inferences to be made about the population based on observations from a sample. In order to be able to make inferences, the sample should not be subject to selection bias. Probability sampling avoids this bias by randomly selecting units from the population (using a computer or table of random numbers). Random means that selection is unbiased -- it is based on chance. With probability sampling, it is never left up to the discretion of the interviewer to subjectively decide who should be sampled.

There are two main criteria for probability sampling: one is that the units be randomly selected, the second is that all units in the survey population have a non-zero inclusion probability in the sample and that these probabilities can be calculated. It is not necessary for all units to have the same inclusion probability, indeed, in most complex surveys; the inclusion probability varies from unit to unit.

There are many different types of probability sample designs. The most basic is simple random sampling and the designs increase in complexity to encompass systematic sampling, probability-proportional-to-size sampling, cluster sampling, stratified sampling, multi-stage sampling, multi-phase sampling and replicated sampling. Each of these sampling techniques is useful in different situations. Again, it is left up to each country as to the probability design selected.

Non-probability sampling, by contrast, is a method of selecting units from a population using a subjective (i.e., non-random) method. An example of non-probability sampling is quota sampling. Since non-probability sampling does not require a complete survey frame, it is a fast, easy and inexpensive way of obtaining data. The problem with non-probability sampling is that it is unclear whether or not it is possible to generalize the results from the sample to the population. The reason for this is that the selection of units from the population for a non-probability sample can result in large biases.

Due to selection bias and (usually) the absence of a frame, an individual’s inclusion probability cannot be calculated for non-probability samples, so there is no way of producing reliable estimates or estimates of their sampling error. In order to make inferences about the population, it is necessary to assume that the sample is representative of the population. This usually requires assuming that the characteristics of the population follow some model or are evenly or randomly distributed over the population. This is often dangerous due to the difficulty of assessing whether or not these assumptions hold. For this reason, STC does not recommend quota sampling or any other form of non-probability sampling.

2.5. Survey designs

The choice of the survey design parameters, namely, stratification, method of selection, sample size determination, sample allocation and actual selection which are the main steps in performing probability sampling depends on the choice of frame. The following section discusses two options: the use of a list frame and the use of an area frame.

If list frame is used: Self-weighted design

A self-weighted design means that each individual in the survey population has the same probability of being selected. STC suggests that countries use a self-weighted design, or as close to a self-weighted design as possible. A list frame will simplify achieving a self-weighted design, as opposed to using an area frame which will be considerably more complicated.

However it is important to note that STC recognizes that the use of a list frame increases survey costs considerably more than the use of an area frame for which the selection of cluster of households reduces operational and interviewing costs. STC recommends that if one country uses an area frame with multi-stage sampling, it will be done using one of the probability sampling methods.

Note: Self-weighted designs simplify analysis. While STC strongly recommends that weights be used at analysis to protect against non-ignorable designs, mis-specified models and non-random attrition patterns (e.g.: a specific group demonstrates higher non-response than other groups which result in non-response bias), we recognize that some of the longitudinal analysis may be model-based and not use the survey weights (for a discussion of design-based versus model-based methods of estimation see the Appendix). Hence the recommendation that a self-weighted design be used.
Examples of self-weighted designs are:

- one stage, unstratified, simple random sample or systematic sample,
- one-stage stratified simple random sample using N-proportional allocation across strata,
- for a two-phase design, self-weighting is achieved by selecting a simple random sample or systematic sample, or a stratified sample with N-proportional allocation at each phase,
- for a multi-stage design, self-weighting is achieved by selecting clusters with probability-proportional-to-size (PPS) at all stages except the final one. At the final stage, a fixed number of units within a cluster is selected (e.g., always pick n=5 at the final stage).

However, if one country uses weights in the production of their estimates, the issue of self-weighted design is less important.

If an area frame is used: Multi-stage sampling

By using a multi-stage sampling, controlling the sample's distribution by age and sex is more complicated. However, some countries have already conducted multi-stage sampling and have experience doing so. The choice of a survey design depends on the experience conducting household surveys in each country. In some country, it would be easier to implement a two-stage design as in other it will be a three-stage sampling design.

The first stage sampling involves the selection of primary sampling units (PSU), which in most cases, are constructed from enumeration area identified and used in a preceding national census of population and housing. The units selected in the second stage are often dwelling or households and the third are typically persons. In multi-stage design, the last stage is the ultimate targeted sampling units, which for GGS is the person.

Note: STC recognizes that a self-weighted design is more difficult to achieve with multi-stage. STC recommends that as long as the design is based on probability sampling and that appropriate estimation technique is used, there is no issue for countries implementing different designs such as unequal probability sampling methods.

If countries conduct a regular rotating household panel national survey [such as a Labour Force survey], the use of the rotated out panels as a sampling frame is one option which will ensure quality national estimates as well as easy top-up sampling procedure and simplify the tracing as well.

Important note on domain of interest: In the choice of the survey design, countries should ensure that the chosen design parameters are driven by the objective of the survey. For GGS, the sampling design should ensure the production of quality estimates for the following main domain of study: Men and Women divided by two age groups: reproductive and non-reproductive age, namely 18-44 and 45-79.

2.5.1 Stratification

If a list frame is used: Stratify the population by sex, age and region

Stratification is recommended for two reasons:

- to ensure that the sample has an adequate representation of men and women of reproductive and non-reproductive ages (see table 1 in item 8.),

- to facilitate regional estimates and link to the metadatabase.

The number of strata should be kept to a minimum in order to avoid dividing the sample into too many, small sub-samples. The following is recommended:

- 2 age categories, dividing the population into reproductive and non-reproductive ages (e.g., 18-44, 45-79),

- as few regions as possible (e.g., aggregate regions wherever possible).

If an area frame is used: First stage sampling: geographical region

Traditionally, when an area frame is used the first stage sampling will be completed with cluster sampling. Cluster sampling is the process of randomly selecting complete groups (clusters) of population units form the survey frame. It is usually a less statistically efficient sampling strategy than simple random sampling and is performed for several reasons. The first reason is that sampling clusters can greatly reduce the cost of collection, particularly if the population is spread out and personal interviews are conducted. The second reason is that it is not always practical to sample individuals units from the population. Sometimes sampling groups of the population units is much easier, such as entire households. Finally it allows the production of estimates for the clusters themselves (e.g. average revenue per household). Different sample designs can be used to select clusters, such as simple random (SRS), systematic (SYS) or probability proportional to size (PPS). A common design uses PPS where sampling is proportional to the size of the cluster.

Each country can decide on which method they prefer to complete their probability sampling. The main criterion is to ensure the minimal number of respondent at the last wave for each of the 4 domains of study.

2.6 Time between waves: 3 years; minimum 3 waves

The sample should be designed for at least three waves: individuals selected for the longitudinal sample in year 1 at wave 1 are followed-up in year 4 (at wave 2) and in year 7 (at wave 3).

2.7 Minimum number of respondents at wave 3 required for the longitudinal sample

The minimum required number of respondents for GGS will vary by country and is driven by the requirement (for longitudinal analysis) that there be at least 2,000 respondent men and women of reproductive and non-reproductive age (see section 2).

3. Sample size determination

This section describes some of the possible methods to derive sample size. The choice of a method will be driven by each country's frame and survey design. Two examples will be presented; one with the use of a list frame and the self-weighted design. In this particular case, a targeted minimal sample size is used to derive the initial sample size. Note that the population size is also required for this method. The second example uses a minimum precision as the target not a minimum sample size. This method requires a target precision for the estimates and a given design effect.

STC recommends that each country have an analysis plan for GGS and ensure that the minimum sample size calculated for either 3.1 or 3.2 of this document meets any precision requirements for estimates.

3.1 Using list frame

There are a fair number of methods available of calculating sample size using a list frame. We present only one example for illustrative purposes.

The following example assumes a list frame and a self-weighted design.

Suppose that a country has 3 million people in the survey population and that it is distributed as follows:

Table 1: Fictitious Distribution of the Survey Population by Age and Sex

		Sex	
		Men	**Women**
Age	**18-44**	N_1=750 000	N_2=840 000
	45-79	N_3=600 000	N_4=810 000

If one country wants to achieve a self-weighted design, the proportion of the sample, a_h, that should fall in cell h is equal to:

$$a_h = \frac{N_h}{\sum_h N_h}$$

where

$$\sum_h a_h = 1$$

Table 2: Distribution of Respondents by Age and Sex

		Sex	
		Men	**Women**
Age	**18-44**	a_1=.25	a_2=.28
	45-79	a_3=.20	a_4=.27

The smallest cell must have 2,000 respondents. From above, that is men aged 45-79. If a_{min} is the smallest proportion, then to determine the sample size in the other cells:

$$n_h = \frac{a_h}{a_{\min}} \times 2000$$

thus,

$$n_1 = \frac{.25}{.20} \times 2000$$
$$n_1 = 2500$$

Similarly, n_2=2800 and n_4=2700. Thus the total minimum number of respondents is 10,000:

$$n = \sum_h n_h$$
$$n = 2500 + 2800 + 2000 + 2700$$
$$n = 10{,}000$$

Note that this achieves a self-weighted design since the probability of selection for all age and sex groups, π_h, are equivalent:

$$\pi_1 = \frac{n_1}{N_1} = \frac{2500}{750000} = .00333$$

$$\pi_2 = \frac{n_2}{N_2} = \frac{2800}{840000} = .00333$$

In general:

$$\pi_1 = \pi_2 = \pi_3 = \pi_4 = \frac{n}{N} = \frac{10000}{3000000} = .00333$$

3.2 Using area frame and estimation precision

Before describing how to do so, we first define sampling error and coefficients of variation.

3.2.1 Sampling error

Sampling error is intrinsic to all sample surveys***.*** Sampling error arises from estimating a population characteristic by measuring only a portion of the population rather than the entire population. A census has no sampling error since all members of the population are enumerated.

The magnitude of the sampling error can be controlled by the sample size (it decreases as the sample size increases), the sample design and the method of estimation.

The most commonly used measure to quantify sampling error is sampling variance. Sampling variance measures the extent to which the estimate of a characteristic from different possible samples of the same size and the same design differ from one another. The standard error of an estimator is the square root of its sampling variance.

Since all sample surveys are subject to sampling error, the statistical agency must give some indication of the extent of that error to the potential users of the survey data. One criterion that is often used to determine whether survey estimates are publishable is the coefficient of variation (CV). The coefficient of variation is the standard error of an estimate expressed as a percentage of that estimate. For example:

$$CV\,(P) = \frac{SE\,(p)}{p}$$

where

p is the true value of some population proportion,

$SE\,(P)$ is the true standard error of that population proportion.

The CV is usually computed as the estimate of the standard error of the survey estimate to the estimate itself, thus for some proportion P:

$$C\hat{V}\,(\hat{P}) = \frac{S\hat{E}\,(\hat{P})}{\hat{P}} = \sqrt{\frac{(1-\hat{P}) \times deff \times (N - n_r)}{n_r \times \hat{P} \times (N-1)}}$$

where

n_r is the number of respondents,
deff is the design effect (explained below).

The design effect (*deff*) is a measure used to quantify the impact of the sample design on the analysis results. Specifically, it is the ratio of the sampling variance of an estimator under a given design to the sampling variance of an estimator under simple random sampling of the same sample size. Therefore, for a simple random sample design, *deff* = 1; for most other designs, typically *deff* > 1. For example, the sampling variance of an estimate from a clustered sample is typically larger than the variance using a sample of the same size not drawn through clusters.

The coefficient of variation is usually expressed as a percentage. It is useful in comparing the precision of sample estimates where their size or scale differs from one another. Statistics

Canada recommends that an estimate with a CV greater than 25% should not be published. An estimate with a CV between 16.5% and 25% may be published but there should be a cautionary note to the user or reader indicating that the estimate has a high sampling variance. An estimate with a CV less than 16.5% may be published without qualification.

For example, to estimate the level of precision we can expect for an estimated proportion of 5% for the smallest stratum (containing 2000 respondents at wave 3), assuming a simple random sample and ignoring the finite population correction factor:

$$\begin{aligned} C\hat{V}(\hat{P}) &= \sqrt{\frac{(1-\hat{P})\times deff}{n_r \times \hat{P}}} \\ &= \sqrt{\frac{(1-.05)\times 1}{2000\times .05}} \\ &= .097 \end{aligned}$$

3.2.2 How to calculate the sample size required to satisfy a given level of precision

For those countries with an analysis plan, it is possible that the minimum sample size calculated in steps 8 and 9 may not be adequately large to produce precise estimates for some domains of interest. For this reason it is recommended that each country also determine the sample size required to meet any precision requirements it might have.

There are standard formulas to calculate the sample size required to precisely estimate a finite population parameter given the design effect for that estimate. For example, to estimate a finite population proportion, *P*, given a targeted level of precision (expressed as a coefficient of variation, CV^2), the following formula could be used (ignoring for now the adjustment for expected nonresponse):

$$n = \frac{deff \times (1-P)\times N}{CV^2 \times P \times N + deff \times (1-P)}$$

Suppose that a country wishes to estimate in each region proportions as low as 5% (i.e., characteristics appearing in only 5% of the population) with a CV of 16.5% and suppose that the design effect in each region is 2 and the size of the population in each region is N=10,000, then:

$$\begin{aligned} n &= \frac{2\times(1-.05)\times 10{,}000}{.165^2 \times .05 \times 10{,}000 + 2\times(1-.05)} \\ n &= \frac{2\times(1-.05)\times 10{,}000}{.165^2 \times .05 \times 10{,}000 + 2\times(1-.05)} \\ n &= 1{,}225 \end{aligned}$$

Thus, if there are 10 such regions, in order to precisely estimate at wave 3 a proportion of 5% in each region, 12,250 respondents are required. Note that this is greater than the 10,000 calculated in steps 8 and 9 earlier.

Note that in order to determine the sample size required at wave 1, the above number of respondents calculated above needs to be adjusted for the expected nonresponse and attrition across the waves as described in chapter.

A country's analysis plan should also determine if cross-sectional estimates are required at each wave. Populations change over time (e.g., due to deaths, immigrants, etc.): the cross-sectional populations at waves 2 and 3 are not the same as the wave 1 longitudinal population. Currently, it is assumed that cross-sectional estimates are not required – that GGS inferences will only be made about the longitudinal population (i.e., the population at wave 1). If a country wishes to produce cross-sectional estimates at each wave, then it may want to add sample at each wave (e.g., of immigrants) in order to ensure cross-sectional representativity.

3.3 Wave 1 sample size must be adjusted for anticipated nonresponse and attrition across the waves

To determine the number of individuals who must be sampled at wave 1 in order to obtain the required number of respondents at wave 3, each country should factor in the expected nonresponse at each wave.

Using the above example, suppose that we need 10,000 respondents at wave 3 and we expect an 80% response rate at each wave and a 10% attrition rate at waves 2 and 3 (attrition refers to individuals who are 'lost' between waves, for example people who move who cannot be traced). Then we must survey 24,113 individuals at wave 1:

$$\begin{aligned} n_{wave\,1} &= \frac{n_{r,wave3}}{rr_{wave1} \times rr_{wave2} \times (1 - att_{wave2}) \times rr_{wave3} \times (1 - att_{wave3})} \\ &= \frac{10000}{.8 \times .8 \times (1 - .1) \times .8 \times (1 - .1)} \\ &= 24113 \end{aligned}$$

where
n_{wave1}=sample size at wave 1
$n_{r,wave3}$=number of respondents at wave 3
r_{rwave1}=response rates at wave 1
att_{wave2}=attrition rate at wave 2

Response rates and attrition rates are likely to vary by sub-groups of population. Different inflation rates can be used in order to ensure the minimal number of respondents. STC recommends that if one country uses different response rates, weights have to be used in order to ensure unbiased estimates of the population.

For example, if one country completed non-response studies showing that young males usually have lower response rates, countries can inflate appropriately.

If a country has an analysis plan, it should ensure that the minimum sample size calculated above meets its analytical needs (for more, see Appendix B).

4. Guidelines for response and attrition rate

It is recommended that the target response rate for GGS be at least 80% at each wave and that the maximum attrition rate be at 10% for each of the 3 waves unless major operational constraints.

Attrition includes lost of units between waves due to not being able to trace them.

STC do not recommend replacing non-respondents with other respondents: each country should make every effort to achieve at least an 80% response rate.

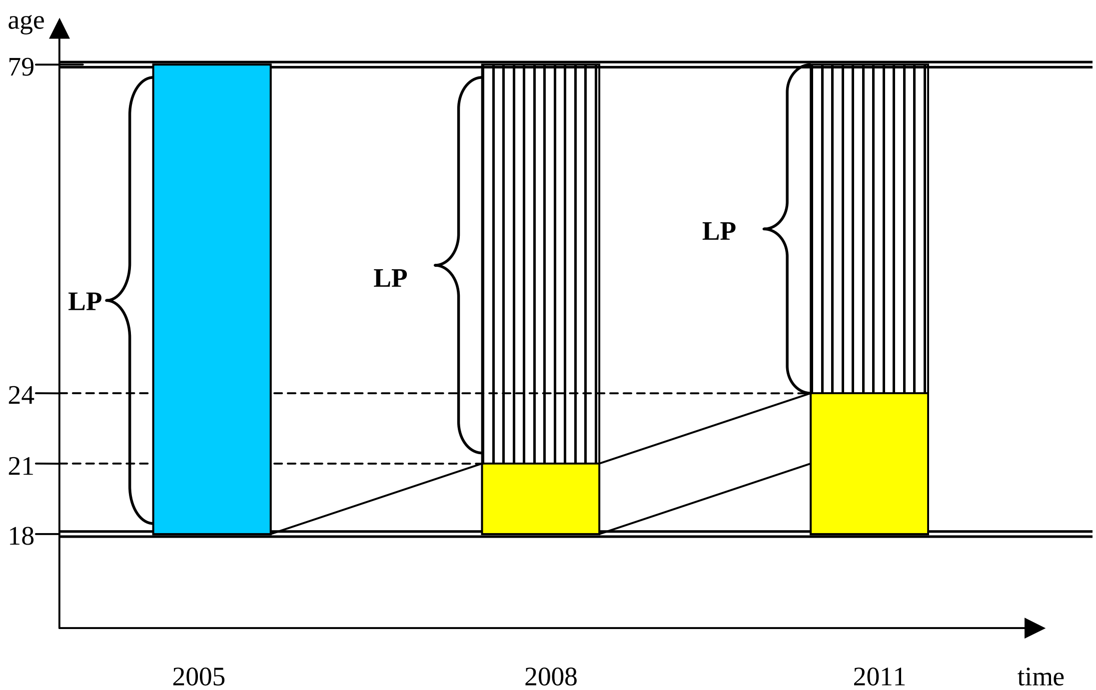
4.1 Over-sampling of subpopulations

Countries may over-sample targeted sub-population. There are several reasons why one country wants to over-sample sub-groups; past studies of non-response or a particular interest in some sub-population. STC recommends that if one country over-sample some sub-population, weights have to be used in order to ensure unbiased estimates of the population.

5. Cross-sectional samples

As seen in diagram 1, the longitudinal population aged from wave to wave, this population is becoming smaller and smaller: the population above 79 is subtracted from the longitudinal population and the youngest at wave 1, the 18 years olds will be respectively 21 and 24 yrs old at each subsequent waves. To ensure the representativity of each cross-sectional population, i.e. from 18 to 79 years for a given wave, the bottom portion (yellow blocks) must be covered by some sampled units as shown in diagram 2.

Diagram 2



.

STC recommends two options:

Option 1: Sample 12-17 year olds in wave 1, but do not conduct interview at wave 1. Trace & interview those who are 18 and over at each subsequent waves.

Option 2: At each wave sample from the cross-sectional population that is not present in the longitudinal sample (thus 18-20 year olds in wave 2 and again only the 18-20 year olds at wave 3 as the 18-20 year olds at wave 2 will be 21-23 year old in wave 3).

As mentioned on page 8, if countries conduct a regular rotating household panel national survey, the use of the rotated out panels as a sampling frame is the simplest option ensuring top-up samples.

6. Other important issues

6.1 Perform non-response weight adjustments

STC recommends that non-response weight adjustments be performed at each wave in order to reduce any nonresponse bias and is strongly recommending the evaluation of any non-response patterns that is non-random. To preserve the self-weighted design, STC recommends that the nonresponse adjustments be performed within the original sample design strata (see also the next point).

6.2 When calculating the final estimation weights, avoid large adjustments to the sample design weights and validate GGS estimates with other sources

The principle behind estimation in a probability survey is that each sample unit represents not only itself, but also several units of the survey population. It is common to call the average number of units in the population that a unit in the sample represents the design weight of the unit. The design weight is the inverse of the probability of selection.

While the design weights can be used for estimation, most surveys produce a set of estimation weights by adjusting the design weights. The two most common reasons for making adjustments are to account for nonresponse and to make use of auxiliary data (e.g., by post-stratification).

The sum of the final estimation weights should equal the country’s total survey population (if a nonresponse weight adjustment is not performed, the sum of the weights will underestimate the total survey population). The final estimation weights should be validated by comparing weighted GGS estimates with other sources (e.g., vital statistics) to verify that the survey’s estimates are accurate.

In order to preserve the self-weighted design, we recommend that the final estimation weights be as close as possible to the original sample design weights (hence the recommendation that the nonresponse adjustments be performed within the sample design strata).

If post-stratification is performed, STC recommends that the post-stratified weight be no greater than 1.5 times the original sample design weight. It is also recommended that if post-stratification is performed, the number of post-strata be kept to a minimum in order to avoid dividing the sample into too many small post-strata which can lead to biased estimators.

6.3 Determine tracing procedures

Attrition can jeopardize the integrity of the sample; high attrition rates in GGS could result in the wave 3 sample no longer being representative of the longitudinal population. It is therefore important that all attempts be made to minimize attrition.

Many things can happen over the course of three years -- the time between GGS waves -- which could make it difficult to contact an individual at subsequent waves. Successful tracing can depend on a large part on the ingenuity and perseverance of those doing the tracing. Some examples of procedures that could be used for GGS include:

- ask the respondent for the name and address of persons close to him/her who are unlikely to move (e.g., parents),

- ask the respondent to notify the survey agency if there is a change of address,

- consider the use of monetary or other incentives to encourage participation and maintain co-operation across waves (e.g., send a survey newsletter once a year),

- send birthday cards every year to remind the individual of the survey,

- institute tracing methods: e.g., telephone directories, motor vehicle registrations, death records for lost persons.

6.4 Define and code non-respondents

At each wave, a person should be considered a non-respondent only if the interviewer has made at least three attempts to contact the person at different times of the day. All non-respondents should be included on the final file along with a nonresponse code in order to be able to calculate nonresponse rates and determine the nonresponse weight adjustments. Every person sampled at wave 1 must appear as a record on the final file along with a final status code. This includes respondents, non-respondents and out-of-scope individuals.

Examples of final status codes include:

- Out-of-scope: The sampled individual does not belong to the survey population. For example, if the survey population is 18-79 and the interviewer discovers that the sampled individual is 16, then this individual is out-of-scope.

- Refusal: The sampled individual refused to participate in the survey or refused to continue before the questionnaire contained enough information to qualify as partially completed.

- No one at home: At least three attempts were made at different times of the day, but no member of the household could be contacted.

- Temporarily absent: The household was contacted but the sampled individual was absent during the entire survey period.

- Unable to trace: All attempts to trace the household or sampled individual were unsuccessful.

- Language difficulties: The interview could not be conducted due to language difficulties.

- Interview prevented due to some disability.

The GGS response rate would then be calculated as:

$$\text{response rate} = \frac{\text{number of responding units (i.e., complete} + \text{partial)}}{\text{resolved in-scope units} + \text{unresolved units}} \times 100$$

For example, suppose a sample of 1,000 units is selected and 800 are resolved (complete, partial, refusal, out-of-scope, etc.) after one week of data collection. Of the resolved units, 700 are in-scope for the survey. Of the in-scope units, 550 respond to the survey (either complete or partial responses). Then, the response rate after the first week of the survey is 550/(700+200)= 61.1%.

6.5 Calculate replicate weights for variance estimation

STC recommend using the collapsed Jackknife or bootstrap to estimate sampling variance of a survey estimate. On requests, more details will be provided on how to calculate these weights.

7. Documentation of the sample design

The following items should be included in the sample design documentation:

- a description of the sampling frame used (including auxiliary variables on the frame and a description of frame defects),
- a definition of the survey population (including percentage under-coverage of the target population; define all exclusions from the target population),
- wave 1 sample size (describe how it was calculated; assumed nonresponse/attrition rates),
- stratification variables,
- sample allocation across strata,
- sample design used (e.g., one-stage stratified SRS, two-stage cluster design, etc. – describe how sampling was conducted at each stage/phase, how clusters were defined, etc.),
- survey response rates observed,
- post-stratification (if performed, explain which variables were used),
- weighting (describe how the sample design and final estimation weights were calculated; describe how nonresponse weight adjustments were performed),
- variance estimation (describe the method used to estimate sampling variance).

Appendix A: Design-Based and Model-Based Methods for Estimating Model Parameters

The GGS sample design guidelines effect to satisfy both the needs of design-based and model-based analysis. The following explanation of design-based versus model-based analysis paraphrases Binder and Roberts (2003) and Pfeffermann (1993).

Sampling weights weigh sample data to correct for the disproportionality of the sample with respect to the target population. When survey data are used to fit a model, one of the first questions an analyst asks is whether and how to account for the survey design at analysis: whether to use sampling weights for the point estimates of the unknown parameters and how to estimate the sampling variance of the estimators required for hypothesis testing and for deriving confidence intervals.

The pure model-based approach would normally result in ignoring the sample design, unless the sample design is an inherent part of the model, such as allowing for different parameter values in different strata for a stratified sample design. In the model-based approach, inferences are made about a conceptual infinite superpopulation. In the traditional model-based approach, the randomization mechanism used to select the units in the sample may be ignored for making model-based inferences.

In the design-based approach, inferences are made about the finite population from which the sample was drawn. The randomization mechanism by which the *n* observations in the sample are selected from the finite population are dictated by the sample design and inferences require that the sample design be accounted for.

Even for model-based analysis, STC strongly recommends that the sampling weights be used since they can protect against non-ignorable designs and mis-specified models (see for example: Chambers and Skinner, 2003, Pfeffermann, 1993, Skinner, Holt and Smith, 1989).

The ignorability of the sampling design depends not only on the design and the available design information but also on the model and the parameters of interest. When the sample is selected by simple random sampling, the model holding for the sample data is the same as the model holding in the population before sampling. With complex sampling designs, the two models can be very different however and failure to account for the sample selection process might bias the inference.

In principle, when all the design features are known, the analyst could verify if the sample design is ignorable. Frequently, however, the analyst has only limited knowledge about the actual sampling process. It is here where the sampling weights come into play.

Even theorists who otherwise oppose the use of sampling weights in a modeling context agree that the sampling weights can play a vital role in situations where the ignorability of the sampling design is at stake. The following is a quote from Fienberg (1989) -- 'The one exception in which the use of weights may be appropriate is outcome-based sampling where the sampling plan may be informative for the model of interest...' See also Hoem (1989). Notice also that even if all the relevant design variables are known, incorporating them in the model may become a major undertaking. As argued by Alexander (1987) - 'no model will include all the relevant variables and few analysts will wish to include in the model all the geographic and operational variables which determine sampling rates. The theoretical and empirical tasks of deriving, fitting and validating such models seem formidable for many complex national demographic surveys.’

Table 1: Model Based versus Design Based Methods

	Assumed model is valid	**Model is mis-specified**
Model –based	• Consistent • Efficient • Valid variance estimates • Valid inferences • May be best	• May be inconsistent if sample design is non-ignorable or if model means and/or variances are incorrectly specified • Variance estimates may be invalid • Inferences may be invalid
Design -based	• Consistent • May be inefficient • Valid variance estimates • Valid inferences	• Consistent for model parameter defined by the mean of the linearized pseudo-variables • Valid conditional variance estimates if model means are correctly specified • Valid estimates of total variance • Valid inferences

References

Alexander, C.H. 1987. A Model Based Justification for Survey Weights, *Proceedings of the Section on Survey Research Methods*, American Statistical Association, p. 183-188.

Binder, D.A., and G.R. Roberts. 2003. Chapter 3: Design-based and Model-based Methods for Estimating Model Parameters, in *Analysis of Survey Data*, edited by R.L. Chambers and C.J. Skinner, John Wiley and Sons, 2003.

Chambers, R.L. and C.J. Skinner. 2003. *Analysis of Survey Data.* John Wiley and Sons,

Fienberg ,S.E.1989. Modeling Considerations: Discussion from a Modeling Perspective in *Panel Surveys*. Kasprzyk, D., Duncan, G., Kalton, G. and Singh, M.P. eds. 1989. New York: John Wiley.

Hoem, J.M. 1989. The Issue of Weights in Panel Surveys of Individual Behaviour, in *Panel Surveys*. Kasprzyk, D., Duncan, G., Kalton, G. and Singh, M.P. eds. 1989. New York: John Wiley.

Kasprzyk, D., G. Duncan, G. Kalton, and M.P. Singh, eds. 1989. *Panel Surveys*. New York: John Wiley.

Macura, M. 2002. *Executive Summary: The Generations and Gender Programme: A Study of the Dynamics of Families and Family Relationships, Advancing knowledge for policy-making in low-fertility, aging societies*, Population Activities Unit, United Nations Economic Commission for Europe, March 30, 2002.

Pfeffermann, D. 1993. The Role of Sampling Weights when Modeling Survey Data, *International Statistical Review*, 61:2, p.317-337.

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