International Labour Office
Geneva

ILO school-to-work transition survey: A methodological guide

Module 3

Sampling methodology

ILO school-to-work transition survey:

A methodological guide

Module 3

Sampling methodology

Sara Elder

Youth Employment Programme
International Labour Office

First published 2009

ILO school-to-work transition survey : A methodological guide

International Labour Office. - Geneva: ILO, 2009.

ISBN 978-92-2-121418-2 (print)

ISBN 978-92-2-121419-9 (web pdf)

ILO publications can be obtained through major booksellers or ILO local offices in many countries, or direct from ILO Publications, International Labour Office, CH-1211 Geneva 22, Switzerland. Catalogues or lists of new publications are available free of charge from the above address, or by email: pubvente@ilo.org.

Visit our website: www.ilo.org/publns

Photocomposed in Switzerland SCR

Printed in Switzerland NOU

Contents

Tables

Foreword

The transition into adulthood is an exciting time for many young women and men, bringing with it the prospect of social and economic independence. For some youth, however, the challenge of finding employment, let alone satisfying employment, can be daunting, and the inability to prove oneself as a productive member of society can overshadow all else to create a sense of frustration and negativity during a time that is meant to be full of hope. It is hard to feel positive about one's future prospects when one becomes blocked at the entry level.

But we should not paint too gloomy a portrait. Many transitions to adulthood, from school to work, are successful from the point of view of the young person and from the point of view of society as beneficiary of their productive contribution. Certainly there must be value in defining a "successful" transition and making the distinction between these and "difficult" ones, i.e. situations in which the young person has little option but to take up unproductive, low-paid and insecure work or drop out of the labour force to wait for rosier times. At least such was the assumption behind the creation of the ILO school-to-work transition survey (SWTS) and the analytical framework, as described in the pages of this Guide, which has been developed around it.

The Guide in itself does not answer the big questions: for instance, why are some transitions difficult and some not; are lengthy transitions an inevitable consequence of a national environment characterized by low levels of job creation; and perhaps most importantly, can governments, working with employers', workers' and youth organizations, do anything about it? What it does do, however, is introduce a holistic methodology that can guide countries to find the answers for themselves. It sets the framework for gathering information on the characteristics and labour market attachments of young people as well as on the enterprises that could absorb them. As stated in Module 1, the data in itself is not unique. What are unique as outcomes from the implementation of the SWTS are the development of indicators that define the stages of transition and the quality of transition and the application of "decent work" as a concept to be integrated into the analytical framework built around the SWTS.

The ILO is committed to helping governments and social partners identify main employment issues and design and implement integrated policy responses. This Guide toward implementation of the SWTS is offered to our constituents as a means to enhance the capacity of national and local-level institutions to undertake evidence-based analysis that feeds social dialogue and the policy-making process. There is a sense of urgency in many countries in addressing the challenge of youth employment.

We recognize that large shares of youth in transition have yet to attain decent employment. Unless they succeed yet another generation of productive potential will remain underutilized and a cycle of poverty and decent work deficits will continue. Clearly, there is still substantial room for action to bring about progress toward the goal to "achieve decent work and productive employment for all, including women and young people" (Millennium Development Goal 1B). It is our sincere hope that this Guide can aid the process and help countries to broaden the information base from which they can address the specific challenges associated with enabling their young populations to reach their full potential.

José Manuel Salazar-Xirinachs
Executive Director
Employment Sector

Acknowledgments

The author wishes to thank all those, past and present, who have had a hand in the development and implementation of the ILO school-to-work transition survey. The surveys have always been resolutely supported by the ILO Youth Employment Programme and for this the author is particularly grateful. Special thanks go to Farhad Mehran, who pioneered the basic concepts of the transition, and Makiko Matsumoto, whose feedback based on the applied tabulation of the raw data has greatly improved the relevant sections of this product. Kee Beom Kim and Mariela Buonomo Zabaleta provided thoughtful comments on the draft. All modules were written by the author with the exception of Module 3 on sampling methodologies, which is a product of El Zanaty and Associates, Egypt.

Purpose and structure of the guide

This guide is intended to help anyone interested in running an ILO school-to-work transition survey. The modular design means that it can be used in full – as a package to advise users from the point of conception of the project through production of an analytical report and workshop to present the findings – or in part, according to the specific needs of the user. Certain modules can be used outside the scope of the school-to-work survey; the module on sampling (Module 3), for example, can guide the design of sampling frameworks of any survey. Likewise, the module on tabulation (Module 4) contains instructions for producing standard labour market statistics, such as the unemployment rate, that researchers can use for any number of purposes, including training on labour market information and analysis.

The guide is organized as follows:

Module 1 Basic concepts, roles and implementation process

Module 2 SWTS questionnaires

Module 3 Sampling methodology

Module 4 Key indicators of youth labour markets: Concepts, definitions and tabulations

Module 5 Disseminating survey results

3.1

Introduction

This module is concerned with sampling issues in the design and implementation of the SWTS. The importance of this module comes from the fact that no matter how well a study is done in other ways, if the sample has not been properly found, the survey results cannot be regarded as correct. Though this module may be more difficult than the others, it is perhaps the most important module in this toolkit. The module is intended to guide a survey team that has existing knowledge on the subject matter of sampling methodology and does not, therefore, spend too much time on basic concepts. The focus here is on specific sampling issues of the SWTS comprising sample design, implementation and measuring sampling errors. This module also explores and provides detailed guidelines for sampling frameworks when they are readily available from other surveys. It also presents different scenarios and methods of dealing with limitations in countries where the SWTS will be implemented, including cases when sampling frameworks from which to build on do not exist. Finally, the module offers guidelines for a standardized sampling framework. The general objectives of the module are to:

- Assist countries in conducting SWTS surveys through following the sampling designs presented herein so as to ensure accuracy and integrity of the sample.

- Provide guidance on sampling methodology that results in unified measurement across countries, thus allowing for both local and cross-country comparisons.

- Transfer knowledge and expertise related to sampling to the local implementing organization.

The rest of this module is organized as follows. Section 3.2 describes technical features of the most commonly used sampling designs that could be adapted and used in the SWTS. Section 3.3 goes into greater detail on the overall framework for generating an SWTS sample, combining several of the methods introduced in section 3.2. Section 3.4 discusses reasons for weighting and how to weight the sample data, while section 3.5 presents methods for computation and analysis of information on sampling errors and data quality. Finally, section 3.6 introduces in brief the methodological techniques applied to obtain a SWTS sample in Egypt as an illustrative example.

3.2
General introduction to sampling

The following definitions and technical terms are relevant to the design and selection of the SWTS sample (as outlined in section 3.3):

3.2.1 SWTS target population and sampling frame

Based on the objectives of the SWTS, the target population of the survey is all young people in the country aged from 15-29.

The sampling frame for the SWTS can be one of two types. The first type is a list of all members of the target population, while the second type is a method of selecting any member of this population. Sampling frames for the general population can be electoral rolls, street directories, telephone directories and customer lists from utilities which are used by almost all households, such as water, electricity, sewerage, and so on. It is preferable to use a list that is the most accurate, complete and up to date. The nature of this list is expected to differ from country to country. Some countries use a list of households, while other countries use a list of people. For the SWTS, a list of households is more useful than a list of people. This list allows the analysis to link transition to household characteristics, such as household income, for example.

National statistical offices are typically the best source of the information needed for a sampling frame since they collect socio-economic and demographic data through periodic household-based sample surveys, such as labour force surveys, living standards measurement surveys, household budget and expenditure surveys, and demographic and health surveys. These surveys may not produce specific data on youth employment, but they do yield information that is useful for obtaining a reliable sampling frame. Working with the statistical agencies is therefore the easiest route for obtaining the sampling frame for use in the SWTS. Other possible sources for obtaining the SWTS sampling frame are labour ministries and public and private research institutions who conduct related researches.

3.2.2 SWTS sample

The SWTS is based on a sample, as opposed to collecting data from the entire target population. This sample is selected according to established principles. These are:

1. the sample must be representative of the entire population; and
2. one must be able to extrapolate inferences, within known and acceptable margins of error, from the sample to the wider population.

The success of the SWTS depends on how the sample is selected and how the different population components are represented in the sample. The SWTS sample can be drawn in several different ways, such as probability sample, quota sample, purposive sample, and volunteer sample, but is most likely to be designed using a combination of methods as the entire population is whittled down to the ultimate units (i.e. youth aged 15-29) in a step-by-step approach (see "multi-stage sampling" and the staged approach outlined in section 3.3). This subsection introduces the most widely-used sample selection methods that can be used in the SWTS. One method may be appropriate for one stage and a different method for the next.

3.2.2.1 Probability samples

Probability samples are sometimes known as random samples. They are the most accurate of the sample selection methods. Any survey aimed at generalizing results drawn from a sample to the whole population of interest must be based on probability sampling. When using a probability sample, each element in the population has a known and non-zero chance of being selected into the sample. Usually, each member of the population has the same chance of being included in the probability sample. With a probability sample, the first step is usually to try to find a sampling frame. Using this frame, individuals or households are numbered, and some numbers are chosen at random to determine who is surveyed. If no frame is available, other methods are used to ensure that every population member has an equal, or known, chance of inclusion in the survey.

To show how random samples based on a sampling frame can be selected, consider the following example. Imagine a list of N households in a certain geographical area, numbered from 1 up to N. Assume further that you want to survey n of them. To draw a *simple random sample*, choose n different random numbers, between 1 and N. Any household whose number is chosen will be surveyed. If the same number comes up twice, the second occurrence is ignored. If there is a chance of repeated numbers equal to r, about $n(1+r)$ selections should be made to get n households. *Systematic sampling* is another type of random sampling that is commonly used in surveys. This type of sampling ensures that no sampling unit will come up twice. In this sampling procedure, no matter how many households or individual you will interview, you need only one random number to draw a systematic sample. The procedure for selecting a sample systematically from a list can be described as follows. Suppose that an equal probability sample of n units is required from a population of N units. From the list of units numbered sequentially from 1 to N, one unit is selected from every $I = N/n$ units in the list. If I happens to be an integer and a random number r between 1 and I is chosen, then, starting with r, every I-th unit is selected. In other words, the sequence numbers selected are r, $r + I$, $r + 2I$, …, $r + (n\text{-}1)I$. The general case when $I = N/n$ is not an integer can be dealt with as follows. Select a real (not necessarily an integer) random number r in the range $0 < r \leq I$. Then the sequence as defined above is constructed. Each term of this sequence, rounded up to the nearest integer, identifies a selected unit.

Probability sampling is the basic methodology of household-based SWTS, which aims to generalize from the sample characteristics to the population or area covered. The types of non-probability sampling (such as quota and purposive sampling) described later in this section are applicable more when the current sample frame is not available (which in the case in many developing countries that have weak or no national statistical offices).

Conditions for using this type of sampling:	Recent and reliable sampling frame of households exists.
Advantage:	Sample is representative of population of interest and results drawn from the sample can be generalized to the entire population.
Disadvantage:	Developed (and recent) list of household presumes a developed statistical agency and tools; often not possible in developing countries for this reason.

3.2.2.2 Quota samples

Quota sampling refers to selection with controls, ensuring that specified numbers (quotas) are obtained from each specified population subgroup (e.g. households or persons classified by relevant characteristics), but with essentially no randomization of unit selection within the subgroups. No population list is used, but a quota, usually based on census data, is drawn up. For example, suppose that in the youth population aged 15-29 years, 50 per cent are known to be males and half of each sex in this target group is in the age group 22-29. If each interviewer had to obtain 20 interviews, this interviewer would be assigned to interview ten males and ten females, five of each aged 15-21, and five of each aged 22-29. It is usually the interviewers who decide how and where they find the respondents. In this case, age and sex are referred to as control variables. This method is usually justified in terms of its convenience, speed and economy. It assumes that the main variability lies across, rather than within, the chosen subgroups, so that, once sufficiently small and homogeneous groups have been defined and properly represented, it is not important which particular individual units within any groups is interviewed.

One of the problems with quota samples is the difficulty to include some respondents rather than others. For instance, in the previous example, the interviewer may quickly find ten females, and five males in the age group 15-21, but it may take a lot of time finding males in the age group 22-29. Another problem is, if too many control variables are used, interviewers will waste a lot of time trying to find respondents to fit particular categories. For example, the interviews may be specified not only based on sex and age, but also based on occupation and household size. Thus, it is very important with quota sampling to use appropriate control variables.

Even when using a probability sampling, sometimes partial quota sampling can be useful. A common example is when choosing one respondent from a household. The probability method begins by finding out how many eligible individuals there are in the household, then selecting an interviewee purely at random. There are practical problems with this approach. For example, appropriate inclusion of males in the sample might not be attainable due to practical considerations. Thus when a household has been randomly selected, it might be beneficial to use quota sampling to choose the person to be interviewed.

Another potential application of quota sampling is that in many (more developed) countries the majority of youth in the age group 15-29 will be in school and there might be only a small percentage that are working, especially among young women. For research purposes, however, it is very important to have a sizable number of working respondents since a main concern of the SWTS is to study the working conditions of youth within the country. A solution to this problem is attainable if the available sampling frame includes detailed information on the youth in the age group 15-29, such as whether they are working or still in school. In this case, a quota sample may be selected and the control variable is whether or not the youth is working.

Conditions for using this type of sampling:	Statistics of subgroups of the population are already known (control variables) and are used to determine interviewees; only a few appropriate control variables are used in the selection of the sample.
Advantage:	Quicker and cheaper than starting from full population lists; no sampling frame is required.
Disadvantage:	Difficulty to locate sufficient respondents of particular characteristics; extreme care to be taken in use of control variables.

3.2.2.3 Purposive samples

A purposive sample refers to selection of units based on personal judgement rather than randomization. This judgemental sampling is in some way "representative" of the population of interest without sampling at random. One of the commonest uses of purposive sampling is in studies based on very small numbers of areas or sites. In these studies, variability with random selection is expected to be excessively large and, hence, potentially more damaging than the bias inherent in selection by judgement. The areas included may be determined on the basis of judgement although, within each area included, the selection of ultimate units may be randomized. Generally speaking, if the budget is small and only a small number of towns and cities can be included, you may choose these in a purposive way, perhaps ensuring that different types of town were included.

A *maximum variation sample*, which is sometimes called a maximum diversity sample, is a special type of purposive sample. The main objective of a maximum variation sampling technique is to select a sample that, in most cases, is more representative than a random sample. Note that a random sample is not always the most representative, especially when the sample size is small. The basic idea of a maximum variation sampling technique can be described as follow. Instead of seeking representativeness through randomness, including a wide range of extremes would guarantee to a large extent representativeness. The usefulness of this procedure is very clear when one is selecting a multi-stage sample, which is explained in more detail later. The first stage is usually to draw a sample of districts or governorates in the whole country. If the number of governorates is very small (for example, less than 30), it is likely that a random sample will be unrepresentative in some ways.

When using maximum variation sampling, you should try to include all the extremes in the population. This method is usually used to choose no more than 30 units. For example, you might decide to select eight governorates in your sample. Then you may select: 1) a rural governorate located in the west; 2) an urban governorate located in the west; 3) a rural governorate located in the east; 4) an urban governorate located in the east; 5) a rural governorate located in the north; 6) an urban governorate located in the north; 7) a rural governorate located in the south; 8) an urban governorate located in the south.

The logic behind this procedure is that if you deliberately select very different areas, the aggregate answers obtained from respondents in these areas will be close to the average.

Conditions for using this type of sampling:	Only a small number of towns or cities can be included in the first stage of sampling as described later in the multi-stage sampling technique.
Advantage:	More representativeness than would be produced by a random sampling type when very small numbers of areas or sites are to be selected.
Disadvantage:	Based on personal judgement rather than randomization and therefore includes a bias in the sample selection.

3.2.2.4 Snowball samples

In some communities (especially those in developing countries), the only feasible way to find its members is by asking other members. The first step in this procedure is to find a few members of the population using any method. This step is denoted as the first round. Then you ask each of these first-round members if they know of any others. The names given will form the second round. Then you go to each of those second-round people, and ask them for more names. This process is repeated for several more rounds. The process is stopped when you start hearing about the same people over and over again. The methodology used to stop the procedure can be described as follow. For each round, count the number of names you get and the number of new names obtained in this round. Then calculate the percentage of new names to the total number of names. For example, if the second round gives you 100 names, but 30 of them are for people who were mentioned in the first round, then the percentage of new names for that round is 70 per cent. This percentage of new names is high at first, but then drops sharply. When the percentage of new names drops to around 10 per cent, then stop. This often happens at the fourth or fifth round. After performing this, something close to a list of the whole population is available, and many of the population members will know that you are planning some research. Using that list, you can draw a random sample. Snowball sampling works well when members of a population know each other. The problem with snowball sampling is that isolated people will not be included in the study.

Conditions for using this type of sampling:	Not very large population and no population list is available; members of the population knows each other.
Advantage:	Easiest way to produce something close to a list of the whole population when no existing list is available.
Disadvantage:	Requires a lot of work when the population is large; isolated people will not be included in the study, which introduces some bias.

3.2.2.5 Volunteer samples

A volunteer sampling procedure might be used when the above procedures are not possible. In general, samples of volunteers should be treated with caution. However, since all survey research involves some degree of volunteering, there is no fixed line between a volunteer sample and a probability sample. The main difference between a pure volunteer sample and a probability sample of volunteers is that, in the former case, volunteers make all the effort; no sampling frame is used.

Conditions for using this type of sampling:	A volunteer sample can produce accurate results only if a high proportion of the population voluntarily returns questionnaires.
Advantage:	No sampling frame is required; less effort in distributing questionnaires to particular individuals and convincing them that participation is worthwhile.
Disadvantage:	With volunteer samples, the proportion who volunteer is the main source of problems. If too few of the population volunteer for the survey, then the research team should investigate what was so special about them. There is usually no way of finding out how those who volunteered are different from those who did not. On the other hand, if the whole population volunteer to take part in the survey, there is no problem.

3.2.2.6 Stratification

A more representative sample can be selected using the stratification procedure. The basic idea here is to divide the target population into strata (groups) based on characteristics that you think are important. Stratification leads to reduced sampling error because it can ensure that all relevant portions of the population are included in the sample. Stratification is easy to do, and it should be used whenever possible for optimal coverage purposes. But for it to be possible, you need to have a) census data about smaller parts of the whole survey area, and b) some way of selecting the sample within each small area. The principle of stratification is simply that, if an area has X per cent of the population, it should also have X per cent of the interviews. The following example shows how stratification can be done and how many respondents should be selected from each cluster. Suppose that a certain area consists of k zones (e.g. localities, cities, etc.). Suppose further that each zone includes n_i households ($i = 1, 2, \ldots, k$) and the total number of households in this $n = \sum_{i=1}^{k} n_i$. Also assume that you need to select S strata from this area. Then the number of strata that should be selected from each zone, S_i, can be calculated using the following:

$$S_i = \frac{n_i}{n} S.$$

The number of strata for each zone should be rounded to the nearest whole number. Sometimes, because of rounding, the designated total number of strata is one more or less than the total you planned for. One of the proposed solutions to fix this is to change the final number of strata, adding or subtracting one.

Generally speaking, the strata represent relatively homogeneous groupings of units. Therefore, the resulting sample is made more efficient by ensuring that units from each population group are appropriately represented in a controlled way.

Conditions for using this type of sampling:	Census data about smaller parts of the whole survey area must be available.
Advantage:	Reduced sampling error; permits separate control over design and selection of the sample within each stratum; more representativeness of the population characteristics; reduced travel and other costs of data collection.
Disadvantage:	Requires subjective choices in determining the defining criteria, number and boundaries of strata.

3.2.2.7 Combined sampling methods: Multi-stage sampling

In national-based surveys such as the SWTS, sampling is done in several steps (as outlined in the following section). The first step is usually to choose a purposive sample of governorates (provinces or administrative regions) from the total number of governorates in the country. In the second stage, a stratified sample of districts (or localities) within each governorate is selected. This procedure continues until the sample of targeted individuals is determined. The multi-stage sampling adapted in SWTS presents several advantages:

- By concentrating on the units to be enumerated into “clusters”, it reduces travel and other costs of data collection.
- For the same reason, it can improve the coverage, supervision, control, follow-up and other aspects that have a bearing on the quality of data collected.
- Administrative convenience of survey implementation increases.
- Selecting the sample in several stages reduces the work and cost involved in the preparation and maintenance of the sampling frame.

The effort involved in sample selection can also be reduced, since more information is usually available for classifying and stratifying larger units.

3.3

SWTS sample design and implementation

In order to fully capture the transitions of different target groups of youth, the SWTS survey addresses the supply and demand factors. As such, the SWTS comprises two surveys – one addressed at youth themselves and another targeted at the employers of youth. The respondents in the SWTS are composed of in-school youth, job-seekers, young employees, self-employed and own-account workers, employers and managers who hire young workers. Thus, the sample of the SWTS includes two main samples: the youth sample and the employers sample.

The following subsections describe in detail the design for both samples.

3.3.1. Youth sample

The youth sample is a sample of eligible persons that is selected in a series of sampling stages, i.e. multi-stage sample:

First stage: In the first stage, the whole country may be divided into administrative regions, such as governorates or provinces. Then a sample of these regions is selected, preferably using a purposive sampling technique to guarantee representativeness. A maximum variation technique, which is described earlier, can be used in the sample selection. Financial, accessibility and time constraints should be taken into consideration in the selection of the first-stage sample.

Second stage: In this stage, each administrative region selected in the first stage may be divided into localities or census enumeration areas (EAs), and a sample of these areas is selected using a stratified technique. The units selected at this stage are usually called primary sampling units (PSUs). At this stage, a frame of PSUs is needed which a) lists the units covering the entire population in each selected administrative region exhaustively and without overlaps, and b) provides information for the selection of units efficiently, such as maps and good household listings. This frame is usually called the primary sampling frame (PSF). A self-weighted stratified systematic sampling technique is recommended in the selection of the PSUs. *Self-weighted* means that the number of PSUs selected from each administrative region should be proportionate to the population size in this region. In this stage, good maps and descriptions for identification and demarcation for each PSU are needed, together with up-to-date information on their size and characteristics. The PSUs should cover the entire governorates selected in the first stage exhaustively and without overlaps. Appropriate size should be selected for each PSU, which conforms to the organization and cost structure of the survey data collection operation. The units serving as PSUs should not be too large or too small. The number of PSUs should be kept limited to control travel and supervision costs.

Third stage: The third stage may consist of dividing each of the PSUs selected in the second stage into smaller areas such as blocks, and then selecting one or more of these third-stage units (TSUs) from each selected PSU. This process may continue until a sample of sufficiently small ultimate area units (UAUs) is obtained. Again, self-weighted stratified systematic sampling techniques are recommended in the selection of the UAUs. The choice of the type of area units to be used in the survey, and the number of such units to be selected for the sample, are very important issues since the type of units chosen to serve as the PSUs and other higher-stage units can greatly affect survey quality, cost and operation. Here we present some general advice in the choice of such units. Firstly, it is not necessary to use units of the same type or size as PSUs in all governorates. Secondly, the survey team should not confuse the formal administrative label with the actual type of units involved.

It should be noted that the appropriate type and size of units depend upon survey circumstances and objectives. Generally speaking, the appropriate size for units serving as the PSUs and higher-stage units depends on survey circumstances and objectives. It is worth noting that many surveys in developing countries use census EAs as PSUs, which typically consist of 100-300 households. The range of the number of selected households per ultimate area in most SWTS conducted so far is 10-50 households. The choice of such a number may reflect differing national circumstances and differences in the type of units involved. However, it is likely that much of this variation is not based on real statistical or cost differences.

Fourth stage: At this stage, which is the last stage, in each selected sample area (or UAU) individual households may be listed and a sample selected with households as the ultimate sampling units (USUs). In the survey, information are collected and analysed for the USUs themselves including youth in the target age group, or just individual youth within sample households. A systematic sampling technique is recommended in the selection of the households in this stage if a list of all households in the UAU is available. If the list of households is not available, one can proceed as follows:

1. Using a quota sampling technique, as described earlier, the interviewer selects a sample of households from the UAU to which s/he is assigned. The control variables that should be used in the selection of the sample are different from one country to another, based on the youth characteristics in the country. However, in most cases, sex and age are important in this selection. Another characteristic of possible interest is ethnic grouping especially if ethnic discrimination is a common problem. It is advised that not too many control variables are used, since too many of them would make the selection process very difficult.

2. Another solution is to prepare a list of households in the UAU using a snowball technique, as described earlier. After preparing the list, a systematic random sample may be chosen.

3. If the above two procedures are not feasible, then a sample of volunteers from the UAU may be considered.

Two different approaches may be applied in the selection of the youth to be enumerated. ***The first approach*** is to enumerate all youth in the target age group within a selected household. This approach is usually preferable when convenience, speed and economy are of interest. In this approach, the sample size of households is expected to be less than the ultimate number of youth in the required sample, especially in developed countries.

The second approach is to select randomly one youth in the target age group from a selected household. In this approach, to guarantee wider coverage of the sample and male/female representation, each UAU may be divided into two segments – one for males and one for females. Then from each selected household in the male segment, a male youth is selected randomly and interviewed and, from each household in the female segment, a female youth is selected randomly and interviewed. The selection of this approach is usually justified in terms of representativeness and wider coverage. To select a youth randomly from the household using this approach, the Kish grid methodology may be applied. This grid is a table of numbers, named after the statistician who invented it. The first step in the Kish grid selection method is to find out how many people living in the household are eligible to be interviewed, including people who reside there, but are not there when you visit. The youngest youth in the household is listed as number 1, the second youngest is number 2, and so on. The last digit in the number of the household where you do an interview is then recorded. In the Kish grid, circle the row containing the last digit in the household number, and the column for the number of eligible youth.

The number in the cell where the column and row meet is the person to interview. For example, if the last digit in the household number is 2 and there are 3 eligible youth in the household, then interview the first youngest youth. If that person is not there when you call, arrange to come back later. See table 3.1 for an illustration.

Table 3.1	**Kish grid**							
Last digit in the HH questionnaire number	Number of eligible youth in household							
	1	2	3	4	5	6	7	8
1	1	1	1	1	1	1	1	1
2	1	2	1	2	1	2	1	2
3	1	2	3	1	2	3	1	2
4	1	2	3	4	1	2	3	4
5	1	2	3	4	5	3	4	5
6	1	2	3	4	5	6	3	6
7	1	2	3	4	5	6	7	4
8	1	2	3	4	5	6	7	8
9	1	2	3	4	5	6	7	8

3.3.1.1 Youth sample design

The youth sample design requires deciding on the number of stages to use and, at each stage, the type of units, method of selection and number of units to be selected. For simplicity, and without loss of generality, we will assume that the PSU = UAU. The ideas described here can be extended to the general case when each PSU is divided into smaller areas.

Suppose that g administrative regions (governorates) are selected from the country in the first stage and suppose further that the population size in each of these governorates is π_i (this is the number of households, or the number of youth in the age group 15-29), $i = 1, 2, \ldots, g$. Suppose further that the total number of UAU to be selected is θ. Then the number of UAUs to be selected from each of the selected governorates is $a_i = \theta \frac{\pi_i}{\sum_{i=1}^{g} \pi_i}$.

To select the UAUs from each governorate with probability proportional to the size of its population, suppose that p_j is the population size of unit j (number of households, or number of youth in the age group 15-29), P_j the accumulation of these sizes for all units 1 to j, ordered in some meaningful way, P this sum over all units in the governorate (i.e. the overall population size in the governorate), then a systematic probability proportional to size sample can be selected as follows. First, the sampling interval to be applied to the cumulative population size of the governorate is $I = P/a_i$. Then, a number r in the range $0 < r \leq I$ is chosen randomly. The first unit in the list of all units in the governorate whose size equals or exceeds r, i.e. the unit sequence number j satisfying the relationship $P_{j-1} < r \leq P_j$, is to be selected as the first unit. Then, starting with r, the selection point is increased each time by I, giving a sequence such as $r' = r + I, r + 2I, \ldots, r + (a_i - 1)I$, and the unit with cumulative population size satisfying the relationship $P_{j-1} < r' \leq P_j$ is selected at each stage.

After selecting the UAUs, a sample of *b* ultimate units (e.g. households) from each UAU is selected using the systematic sampling technique described in the previous section if a list of households exits, where *b* is a constant. If this list is not available, then the other techniques described above may be used in the selection of *b* ultimate units. The overall sample size of households is $n = b * \theta$. It can be shown that the overall selection probability of an ultimate unit using this procedure is equal to $f = \frac{n}{\sum_{i=1}^{g} \pi_i}$, which is a constant.

In the ideal situation, the above design leads to control over sample size, fixed workload *b* per sample area, and a uniform overall sampling probability *f* for each ultimate unit (e.g. household, youth). In practice, however, it is unlikely that all of the conditions required for this design are satisfied exactly. Some common variants of this basic design might take place. For example, the UAU size (p_j) determining selection probabilities of area units is usually based on past information collected at the time the area frame is compiled. These may differ from the actual sizes (p'_j) at the survey time. Inaccuracy of the population size is a common and, hence, important problem. In this case, a self-weighting design of ultimate units is preferable over a fixed-take design. In the self-weighting design, the ultimate units all receive the same overall probability of selection, and the number of these units selected is allowed to vary to the extent the population sizes used in the selection of the UAUs differ from their actual sizes. To avoid extreme variations both in the overall selection probabilities and in the sample-takes per UAU, one can use the following equation to determine the number of units to be selected from each UAU:

$$b_j = b * \sqrt{\frac{p'_j}{p_j}}$$

In this case, the overall sample size is $n = \sum b_j$, where the summation here is over all the UAUs in the country.

In the above design, if very large or very small areas exist, one can use the following suggestions. A very large UAU means that the size of the area p_j is greater than the sampling interval *I*, so that the probability of selecting this area is greater than 1, which is not possible. One suggestion in this case is to divide large units in the sampling frame into smaller segments (divided into smaller areas) such that no segment exceeds *I* in size. The segments then form the appropriate sampling units. On the other hand, a very small UAU means that the size of the area p_j is smaller than *b*, so that the sampling rate in this area is greater than one, which is not possible. In this situation it is recommended that small units in the sampling frame be grouped together (merged to form larger areas) such that no group is smaller than *b*. The groups then form the appropriate sampling units.

3.3.1.2 Choosing the youth sample size

In the SWTS survey, the choice of the youth sample size is the most basic and important issue in sampling design. This issue, however, remains a difficult one that eludes any purely scientific answer. Generally speaking, the choice of the youth sample size must incorporate several factors, including practical considerations in conducting the survey (i.e. availability of trained personnel, deadlines for completion, computing facilities), the desired precision of the estimates and survey costs. This subsection presents some guidelines on choosing a sample size that balances some of these factors.

Suppose that a simple random sample of size *n* is to be selected from the entire youth population in the age group 15-29 to estimate the proportion of transited youth. Suppose further that the actual proportion of transited youth in the population is *P*. To determine a suitable value for n, several important questions need to be addressed. The first question is how many youth are in the population that the sample is supposed to represent. Statistically, this only becomes an issue if the population size is limited (for example, below 10,000 individuals). But if the population of youth is large (10,000 or more), then the sample size does not depend at all on the size of the population to be sampled. Here, we assume that the population size is large (at least 10,000 youth). The second question is how much potential margin of error and confidence level you want in the survey. It is well known that the sampling error in estimating the proportion of transited youth when conducting the SWTS survey is: $SE = \sqrt{\frac{P(1-P)}{n}}$. To be 100*(1-α) per cent confident that the true proportion of the transited youth will be within $\pm Z_{\alpha/2} SE$ points from the sample estimate of this proportion, you need to sample:

$$n = \frac{Z^2_{\alpha/2} P(1-p)}{SE^2} \qquad (1)$$

where $Z_{\alpha/2}$ represents the equivalent value corresponding the a (1-α) per cent confidence level from the standard normal table. Usually (1-α) per cent is chosen as 95 per cent and, in this case, $Z_{\alpha/2}$ = 1.96. This means that if the same survey was done 100 times, then 95 of those times the interval $\hat{P} \pm Z_{\alpha/2} SE$ would contain the actual proportion of transited youth in the population, where $\hat{p}$ is the sample estimate of this proportion. For example, you guess that maybe a quarter of all youth in the country are transited, so *P* is 25 per cent. You want the figure to be correct within ±3 per cent with confidence level 95 per cent, so the required sample size is *n* = 1.96 x 25 x 75 / (3 x 3) = 408.

Equation (1) requires knowing roughly how many youth are transited, and also how large a standard error you can tolerate and the confidence level you feel comfortable with. As noted above, most surveys use a confidence level of 95 per cent and a tolerance sampling error of approximately 1 to 3 per cent. The value of *P* in equation (1) is usually determined based on personal beliefs and experience. Scientifically, one might perform a pilot study (a pre-study) to get a rough estimate for *P*. On the other hand, if the value of P cannot be roughly determined prior to the survey, usually it is assumed that $P = 0.5$. It can be shown that the maximum value of the sample size n is satisfied when $P = 0.5$. Thus, using this value guarantees greater precision (smaller standard error) of the sample estimates.

You would calculate the sample size for a survey as shown above if you use a simple random sampling methodology design, and cost is not taken into account. However, the SWTS survey is done using a multi-stage sampling design and is limited by a budget. Thus, your starting point in applying the SWTS may not be how much error you can tolerate, but rather how little error you can get for a given cost and a different design methodology.

In general, sampling precision is not only determined by the size of the sample but also by its design, i.e. its efficiency or "design effect". The design effect is a comprehensive summary measure of the effect on sampling error of various design complexities. It is the ratio of actual to simple random sample standard error. The design effect measure removes the effect of factors common to both, such as size of the estimate and scale of measurement, population variance and overall sample size. Suppose that D is a measure of design effect, then the youth sample size n_y is calculated using: $n_y = nD^2$ where n is as defined in equation (1). The value of D is usually estimated by information obtained from related surveys, such as the labour force survey. However, roughly, a value between 1.5 and 2 is usually used in surveys based on multi-stage sampling designs. In the above example, if a value of 2 is used as a rough estimate for D, then the youth sample is $n_y = 1633$. On the other hand, if information from previous surveys is available, the design effect measure can be calculated using the following equation:

$$D^2 = 1 + (\bar{b} - 1)\rho$$

where $\bar{b} = \frac{\sum b_j}{\theta}$ and ρ is a measure of the intra-cluster correlation (the correlation between areas).

Other considerations that might be taken into account when selecting a sample size are expected response rate and cost of data collection. If it is expected that only R per cent of the selected youth would respond, then the sample selected has to be larger by the factor $\frac{1}{R}$. Also, suppose that C is the cost of covering an area, and c is the cost of collecting data per ultimate unit in the sample. Then the sample selected has to be larger by the factor $F_c = 1 + \frac{C}{b * c}$. Briefly, the youth sample size required for performing the SWTS survey may be calculated using:

$$n_y = \frac{nD^2 F_c}{R}$$

For instance, in the above example, suppose that R = 90 per cent, C = US$ 50, c = US$ 5 and b = 30, then the required youth sample size is n_y = 2,420.

The number of households to be interviewed (n_h) depends on both n_y and the approach that will be applied in the selection of the youth to be enumerated. As mentioned previously, two approaches may be used. In the second approach, which is interviewing one youth per household, the number of households $n_h = \frac{n_y}{K}$, where K is the percentage of households in the country with at least one youth in the age group 15-29. In the above example, if K = 70 per cent, then the number of households to be visited is n_h = 3,456.

In the first approach, which is based on interviewing all the youth in the household, the number of households to be visited is approximately $n_h = \frac{n_y}{K\bar{y}}$, where $\bar{y}$ is the average number of youth in the age group 15-29 per household in the country. Assume that $\bar{y}$ = 2, then using the first approach you need to visit n_h = 1,728 households.

Apart from this, the SWTS conducted in several countries suggested a target youth sample size in the range 1,300 to 4,000. A target sample below 1,300 is useful only if you have a very low budget, and little or no information on what proportion of the population engages in work activity, or if the entire population is not much larger than that. A target sample size over 4,000 is most probably a waste of time and money, unless there are subgroups of the population that must be studied in detail. If you do not really need a large sample size, and have more funds than you need, do not spend it on increasing the sample size beyond the normal level. Instead, it is better to spend it on improving the quality of the survey, e.g. getting more interviewer training, more detailed supervision, more verification and more pre-testing.

3.3.2 Employers sample

It is recommended that the employers sample size be based on the youth sample size n_y. If it is known that the average youth in the age group 15-29 hired by employers (within the formal and informal economy) in the country is v, then it is recommended that the target sample of employers be with size $n_E = \frac{n_y}{v}$ enterprises. If this average is unknown, then roughly use an employer sample $n_E = \frac{n_y}{10}$ size of enterprises. The employer sample should be divided into two samples; a sample from formal enterprises and a sample from informal enterprises. The distinction between the formal and informal enterprises usually depends on the compliance of the enterprise to certain rules that imply formality. These rules are different from one country to another, even from one area to another within the same country. However, the majority of developed countries use the availability of a licence, the commercial or industrial register, if required, and keeping regular accounts as indications of formality. So, if these conditions are all satisfied, the enterprise is of a formal nature, while if at least one of them is not satisfied, the enterprise is considered to be informal. The size of the formal and informal samples may be based on the percentage of youth in the age group 15-29 hired by each type of enterprises. If, for example, the percenatge of youth in the age group 15-29 hired in the formal economy is 70 per cent, then the target sample of formal enterprises may be equal to $0.7 * n_E$ and, consequently, the target sample of informal enterprises be equal to $0.3 * n_E$. If this percentage is unknown, then the employer sample may be split equally between formal and informal enterprises.

The formal enterprises should be selected from the same governorates selected for the youth sample unless it is known that enterprises that hire youth are located in other governorates. For example, in Egypt, the majority of enterprises hiring youth in the Cairo governorate are located in the Giza governorate. Thus, enterprises in the Giza governorate should be targeted in the employer sample if Cairo governorate is considered for the youth sample, regardless of whether or not the Giza governorate is considered for the youth sample.

To select the formal employers' sample, a complete frame is needed. Usually, an updated list of formal enterprises is available in most countries. A systematic sampling methodology may be applied to select the formal enterprises' sample if this list is available. If not, a qouta or maximum variation sampling technique may be applied. The enterprises may be classified by type of industry in the different governorates to estimate the number of enterprises needed to be selected from each governorate and each type.

The informal enterprises' sample may be chosen from lists of enterprises prepared by the data collection teams during fieldwork procedures. Each team may be assigned to list all the enterprises doing business in the same area unit visited by the team, and then an enterprise may be chosen randomly from the list.

3.4

Weighting of sample data

There is a variety of reasons for weighting the sample data, including adjustment for probability of selection, non-response, stratification, etc. Weighting is a reasonable way of approximation to adjust an existing sample for known biases. In this section, we discuss in some detail the reasons for weighting and how to weight the sample data.

3.4.1 Reasons for weighting

Weighting for non-response: No matter how hard we try, we cannot control who decides to participate in our survey and who decides not to. In some situations, it is possible to compare respondents and non-respondents on the basis of some attributes, such as gender. For instance, if census data give a good estimate of the true proportion of males and females, and after collecting the data we find that the observed distribution of gender does not conform to the true population, one may wish to weight responses to adjust accordingly. For example, suppose that too few women have responded, then it is better to weight their responses more than the male responses. If the true proportion by gender is 50-50, and if one got 20 females and 80 males, then one could weight each female response by 4. This, in effect, gives 80 females and 80 males. However, to avoid artificially increasing sample size from 100 to 160, one needs further weighting to scale back to 100. This could be achieved by further weighting both females and males by five-eighths.

Weighting for post-stratification adjustment: The same logic and same weighting strategy discussed above for non-response applies if the under-representation of a given stratum (e.g. urban males aged 20-25) is due to non-response or due to disproportionate stratified sampling. The objective here is to weight the existing cases in a way which increases the representation in the adjusted sample of the strata that are under-represented in the raw data.

Weighting to account for the probability of selection: The selection procedure described in the previous section assumes that each household has an equal chance of being selected. However, individuals within households with more people have a lower chance of being selected. This means they are under-represented and should be weighted more. Assuming one survey per household and the average number of eligible individuals per household is y, the weight for any surveyed individual in the sample is then the number of eligible people in that household divided by this average. For instance, if a given household had six eligible individuals and the average is $y = 2$, the weight for that case would be 6/2 = 3.

The final weight of each respondent is the product of all weights calculated for this individual to account for imbalances in sampling rates.

3.4.2 Advantages of weighting

By weighting data to compensate for imbalances between the proportions of targeted participants among subgroups in the population and the proportions in those subgroups who choose to respond, we ensure that the estimates are adjusted to provide a better fit to what we believe to be the true characteristics of the target group. Generally speaking, weighting provides us with more accurate population estimates. When the sample data are to be weighted, it is highly recommended to attach to each individual case or record its weight as a variable in the data file. Most of the required population estimates, such as proportions, means, ratios and rates, can then be produced easily, without the need to refer to the structure of the sample.

3.4.3 Effects of weighting on variance and bias

The effect of weighting is generally to increase the variance over that in a corresponding self-weighting sample. A close approximation to the factor by which variance is increased is the following:

$$E_w = \frac{n\sum w_i^2}{(\sum w_i)^2}$$

where w_i is the weight of the *i*-th individual. This factor can be then used to calculate the sampling error and confidence intervals of the population parameters. The importance of this equation is that it gives the magnitude of the effect by which all variances for different survey estimates are inflated more or less uniformly as a result of weighting. The bias resulting from ignoring weights depends on the difference in both the mean values and sizes of the groups with different weights, and is not the same for different types of statistics. Hence, its relative magnitude in relation to the effect of weighting on variance can vary depending on the type of statistic considered.

3.5
Data quality and sampling errors

The quality of any survey is questionable, and needs to be guaranteed and measured. In order to control the quality of data collected, two issues need to be emphasized. First, the quality of the collected data has to evaluated, and then the error in the calculated indicators has to be measured. Accordingly, two important topics need to be discussed in the following sections: data quality and sampling errors.

3.5.1 Data quality

In any survey, some tables could be produced to look at the quality of the collected data. Those tables are usually calculated for variables that could be compared with existing data from different sources (such as census data); for example, distribution of age by single years for the household members (in the case of selecting a household sample) or for interviewed youth. This distribution table will show the heaping of the age data. If the distribution of members over the different single years is similar to the census data, this means that the survey data quality is good. Also, Mayer's index could be measured to look at the accuracy of age data.

3.5.2 Sampling errors

The estimates from a sample survey are affected by two types of errors: 1) non-sampling errors, and 2) sampling errors. Non-sampling errors are the results of mistakes made in implementing data collection and data processing, such as failure to locate and interview the correct household, misunderstanding of the questions on the part of either the interviewer or the respondent, and data entry errors. To minimize this type of error, numerous efforts have to be made during the implementation of the survey. These efforts include high quality and intensive training, good field supervision, double entry of the data to minimize data entry errors. However, non-sampling errors are impossible to avoid and difficult to evaluate statistically.

Sampling errors, on the other hand, can be evaluated statistically. The sample of respondents selected in the SWTS survey is one of many samples that could have been selected from the same population, using the same design and expected size. Each of these samples would yield results that differ somewhat from the results of the actual sample selected. Sampling errors are a measure of the variability between all possible samples. Although the degree of variability is not known exactly, it can be estimated from the survey results.

A sampling error is usually measured in terms of the *standard error* for a particular statistic (mean, percentage, etc.), which is the square root of the variance. The standard error can be used to calculate confidence intervals within which the true value for the population can reasonably be assumed to fall. For example, for any given statistic calculated from a sample survey, the value of that statistic will fall within a range of plus or minus two times the standard error of that statistic in 95 per cent of all possible samples of identical size and design.

If the sample of youth had been selected as a simple random sample, it would have been possible to use straightforward equations for calculating sampling errors. However, a multi-stage stratified design needs to use more complex equations. To calculate sampling errors for the SWTS, if it is a multi-stage sample, one has to use the Taylor linearization method of variance estimation for survey estimates that are means or proportions.

3.5.2.1 Practical calculation of sampling errors

Practical procedures for estimating sampling errors must consider the actual sample design, but need to be flexible enough to apply to diverse designs.

Two methods are commonly used in practice for surveys based on complex multi-stage designs:

1. Computation from comparisons among certain aggregates for "primary selections" within each stratum of the sample.
2. Computation from comparisons among estimates for "replications" of the sample. The term "primary selection" refers to the set of ultimate units obtained by applying a certain specified sub-sampling procedure to each selection of a primary sampling unit. The term "replications" refers to parts of the sample, each of which reflects the structure (e.g. clustering, stratification, allocation) of the full sample, and differs from it only in size.

Method 1 (estimating variance from comparison among primary selections)

This method is based on the comparison among estimates for independent primary selections within each stratum of a multi-stage design.

This is perhaps the simplest approach for computing sampling errors of common statistics such as proportions, means, rates and other ratios, and the method can be easily extended to more complex functions of ratios such as differences or ratios, double ratios and indices. The basic equations are as follows.

Consider a population total Y obtained by summing up individual values Y_{hij} for elements j over PSU i, and then over all PSUs and strata h in the population:

$$Y = \sum_h Y_h = \sum_h \sum_i Y_{hi} = \sum_h \sum_i \sum_j Y_{hij}$$

The above is estimated by summing appropriately weighted values over the units in the sample:

$$y = \sum_h y_h = \sum_h \sum_i y_{hi} = \sum_h \sum_i \sum_j w_{hi} y_{hij}$$

For the combined ratio estimator of two aggregates *y* and *x*:

$$r = \frac{y}{x} = \frac{\sum_h y_h}{\sum_h x_h} = \frac{\sum_h \sum_i y_{hi}}{\sum_h \sum_i x_{hi}} = \frac{\sum_h \sum_i \sum_j w_{hij} y_{hij}}{\sum_h \sum_i \sum_j w_{hij} x_{hij}}$$

the general expression for variance is:

$$Var(r) = \sum_h \left[(1 - f_h) \frac{a_h}{a_h - 1} \sum (z_{hi} - \frac{z_h}{a_h})^2 \right]$$

where a_h is the number of primary selections in stratum *h*, f_h the sampling rate in it, and the computational variable *Z* defined as:

This approach is based on the following assumptions about the sample design:

1. The sample selection is independent between strata.
2. Two or more primary selections are drawn from each stratum ($a_h > 1$).
3. These primary selections are drawn at random, independently and with replacement.
4. The number of primary selections is large enough for valid use of the ratio estimator and the approximation involved in the expression for its variance.
5. The quantities x_{hi} in the denominator (which often correspond to sample sizes per PSU) are reasonably uniform in size within strata.

The above variance estimation equations are simple, despite the complexity of the design, being based only on weighted aggregations for the primary selections, and identification of the strata. The complexity of sampling within PSUs does not appear to complicate the estimation procedure. No separate computation of variance components is required. This provides great flexibility in handling diverse sampling designs, which is one of the major strengths of this method, and the reason for its widespread use in survey work.

Method 2 (comparison among replications of the full sample)

These procedures are more complex and computer intensive, but can be applied to statistics of any complexity. The basic idea is that of "repeated re-sampling". This approach refers to the class of procedures for computing sampling errors for complex designs and statistics in which the replications to be compared are generated through repeated re-sampling of the same parent sample.

Each replication is designed to reflect the full complexity of the parent sample. The replications in themselves are not independent replications (in fact they overlap), and special procedures are needed to control the bias in the variance estimates generated from comparison among such replications.

Compared to method 1, repeated re-sampling methods have the disadvantage of greater complexity and more computational work. They also tend to be less flexible in the sample designs handled.

They have the advantage, however, of not requiring an explicit expression for the variance of each particular statistic. They are also more encompassing: by repeating the entire estimation procedure independently for each replication, the effect of various complexities – such as each step of a complex weighting procedure – can be incorporated into the variance estimates produced.

The various re-sampling procedures available differ in the manner in which replications are generated from the parent sample and the corresponding variance estimation equations evoked. We examine here three general procedures: the "jack-knife repeated replication" (JRR), the "balanced repeated replication" (BRR), and the "bootstrap exit", though the last mentioned is not yet established for general use in complex designs. The JRR generally provides the most versatile and convenient method.

Jack-knife repeated replication (JRR):
With JRR, a replication is formed by dropping a small part of the total sample, such as a single PSU in one stratum. Consequently, each replication measures the contribution of a small part such as a single stratum.

The basis of this method is as follows. Consider a replication formed by dropping a particular PSU *i* in stratum *h* and, to compensate for the missing PSU, appropriately increasing the weight of the remaining (a_h-1) PSUs in that stratum. The estimate for a simple aggregate (total) for this replication is:

$$y_{(hi)} = \sum_{k \neq h} y_k + \frac{a_h}{a_h - 1}(y_h - y_{hi}) = y - \frac{a_h}{a_h - 1}(y_{hi} - \frac{y_h}{a_h})$$

with the average of estimates over the stratum,

$$y_{(h)} = \frac{\sum_i y_{(hi)}}{a_h}$$

and the average over all $a = \sum_h a_h$ replications,

$$\tilde{y} = \frac{\sum_h \sum_i y_{(hi)}}{\sum_h a_h}$$

the expression for variance of any statistic can be written in various statistically equivalent forms:

$$Var_1(y) = \sum_h \left[(1-f_h)\frac{a_h - 1}{a_h} \sum_i (y_{(hi)} - y_{(h)})^2 \right]$$

$$Var_2(y) = \sum_h \left[(1-f_h)\frac{a_h - 1}{a_h} \sum_i (y_{(hi)} - \tilde{y})^2 \right]$$

$$Var_3(y) = \sum_h \left[(1-f_h)\frac{a_h - 1}{a_h} \sum_i (y_{(hi)} - y)^2 \right]$$

In the JRR method, the standard variance form for simple aggregate *y*:

$$Var(y) = \sum_h \left[(1-f_h)\frac{a_h}{a_h - 1} \sum (y_{hi} - \frac{y_h}{a_h})^2 \right]$$

is replaced by one of the above three expressions (usually the last of the three, as it is more conservative). Based, as they are, on nearly the full sample, estimates such as $y_{(hi)}$, $y_{(h)}$ and even more so, their overall average $\tilde{y}$ are expected to be close to the full sample estimate *y*, even for complex statistics. Hence their variance, expressed by any of the three forms, provides a measure of variance of *y* as well. This applies to statistic *y* of any complexity, not only to a simple aggregate.

In many cases the basic assumptions regarding the structure of the sample are not met exactly. Accordingly, a practical solution can be found in most situations. For example, in some cases, the minimum of two primary sampling units per stratum could not be achieved. The practical solution in this case is to pair adjacent units as one stratum.

3.6

A country example: Egypt SWTS (2005)

This section presents descriptions of the 2005 SWTS sample in Egypt to illustrate some of the features and principles of sample design that are highlighted in this module.

3.6.1 Youth sample

The first thing that was needed for the survey was a nationally representative sample frame. The CAPMAS, which is the state body governed as the official source for data and statistics did not have a comprehensive sample frame (i.e. no household listing from which to draw the sample). In addition, the CAPMAS divides the country into enumeration areas (EAs) which is based on the number of dwelling units. However, EAs were not defined in terms of easily identifiable boundaries, and also sketch maps of these areas were not available. Consequently, a decision was made to use the Interim Egypt Demographic and Health Survey 2003 (EIDHS, 2003) which is a master sample that was developed based on the CAPMAS frame of PSUs. This sample guaranteed an updated frame and it also used *shiakhas* and villages as the basic sampling units. In each governorate, the list of *shiakhas* and *medinas* constituted the initial sample frame for the urban area and, similarly, the list of villages for the rural area.

The following are the steps taken in order to select the youth sample:

1. A target sample of around 4,000 youth was needed for the analysis. Taking into consideration that there was no youth listing from which to draw the youth sample immediately, the proportion of youth in households was calculated from the EIDHS 2003. Analysis showed that there is, on average, 0.80 youth (aged 15-29 years) per household. Thus, it was expected to find 4,000 youth in around 5,000 households (4,000/0.80). Taking into consideration 10 per cent non-response rate, a total of 5,520 households were needed to be selected.

2. *Selection of ten governorates:* Ten governorates were selected for the survey. Those governorates were randomly selected; four from lower Egypt and four from upper Egypt, while two of them represent the urban governorates (i.e. half the governorates within each region).

   For drawing the governorates within each region: interval = the total number of governorates / the targeted number of governorates; then, using a spreadsheet, a random number-generating function was used to develop the first governorate to be selected. Then the interval was added to get the second one, and so on. See results in table 3.2.

The sample was designed to be self-weighted, since there was no need to have governorate-level estimates. This made the analysis easier. So, we needed to allocate the 5,520 households to be selected from each selected governorate depending on the size of the governorate relative to the total population. The census estimates of the governorate population were used to get the proportions (i.e. population size was available from the census data or calculated based on projections of it).

3. *Selection of the PSUs:* A random systematic sample was used. Two set of calculations were needed:

   a) Number of PSUs within each governorate: Looking at the EIDHS sample design, to have a wide representation on average 30 households should be drawn from each PSU. Thus, for the SWTS, the same design was implemented (i.e. an average of 30 completed interviews of households per PSU). So, for each governorate, the number of PSUs was calculated (i.e. governorate targeted households/30).

   b) Procedure of PSU selection: The same procedure for governorate selection was applied. We calculated the interval as follows: PSUs in the frame/targeted PSUs. Thereafter, using a spreadsheet, a random number-generating function was used to get the initial PSU to start from and then follow on with the interval. It is also worth noting that rounding was used (rounding or truncation could be used, but the same rules should be applied for both the PSU and household selection).

4. *Selection of households:* A random systematic sample selection was implemented. The interval was calculated: Total number of households in PSU/targeted households for PSU. Thereafter, using a spreadsheet, a random number-generating function was used to create the first household to start with.

Table 3.2 Illustrative examples of governorates selection

			Urban	Rural	Total	Selection	
Urban governorates	1	Cairo	7629909		7629909	*	Total of U.G. 4
	2	Alexandria	3755902		3755902		# of selected G. 2
	3	Port Said	529661		529661	*	Interval 2
	4	Suez	478554		478554		RAND. 2.4334 VALUE 1.1432 TRUNC. 1
Lower Egypt	1	Damietta	312642	718316	1030958		
	2	Dakahlia	1368027	3471262	4839289	*	Total of L. Egy. 9
	3	El Sharkia	1143538	3866111	5009649		# of selected G. 4
	4	Kalubia	1553828	2250336	3804164	*	Interval 2.25
	5	Kafr El-Sheikh	592033	1949130	2541163		RAND. 2.1097
	6	Gharbia	1211887	2647421	3859308	*	VALUE 2.3563
	7	Menoufia	645269	2525754	3171023		TRUNC. 2
	8	Behera	925496	3677178	4602674		No.2
	9	Ismailia	422049	422039	844088	*	
Upper Egypt	1	Giza	3290929	2247920	5538849		
	2	Beni Suef	518078	1690376	2208454	*	Total of U. Egy. 8
	3	Fayoum	528901	1842877	2371778		# of selected G. 4
	4	El Minia	758499	3202126	3960625	*	Interval 2
	5	Assuit	905059	2445982	3351041		RAND. 1.3177
	6	Souhag	800317	2930547	3730864	*	VALUE 2.8564
	7	Qena	801239	2487552	3288791		TRUNC. 2
	8	Aswan	465377	637198	1102575	*	

3.6.2 Employers sample

The target sample of employers was 300 enterprises. The employer sample was divided into two samples; one sample from formal enterprises and one sample from informal enterprises. The distinction between the formal and informal enterprises in Egypt usually depends on the compliance of the enterprise to certain rules that imply formality. These rules are the availability of a licence, the commercial or industrial register, if required, and keeping regular accounts. So, if these conditions were all satisfied, the enterprise would be of formal nature, while if at least one of them was not satisfied the enterprise would be considered to be informal. To find the list of all enterprises (frame) was not easy. Also, no informal enterprise frame was available.

To select the employer sample, a complete frame was needed. After reviewing the databases available in Egypt about enterprises, including Kompass, GOFI and the Commercial Ahram Guidebook, 2004, a decision was made to use the latter as it is the most recent and reliable source of economic activities conducted in Egypt as a frame of employers. The enlisted enterprises were classified by type of industry in the different governorates to estimate the number of enterprises needed to be selected from each governorate and each type. The number of enterprises selected from the guidebook was 171, and 167 interviews were completed.

The informal enterprise sample was chosen from lists of enterprises prepared by the data collection teams. Each team was assigned to list all the enterprises doing business in the same sample unit visited by the team, and then an enterprise was chosen randomly from the list. A total of 184 were selected from the informal sector. The number of employer interviews completed was 347.

3.6.3 Example for calculating the sampling errors

A question in the questionnaire asks if the interviewee “Ever worked while attending school”. This was a yes/no question. The sample statistic rate of individuals who answered yes to this question was 5.8 per cent. We need to estimate the 95 per cent confidence interval for the population parameter corresponding to the above sample statistic.

The required data for the analysis is the following four variables:

1. Cluster identification variable (the PSU/segment number: HPSU).
2. The strata number (stratum).
3. Sample weight (if the sample is self-weighted, we create a weight variable with the value 1 for all cases).
4. The variable corresponding to a question, for example, question 207 (variable is Y207).

Data preparation: The variable Y207 is recorded such that the answer "yes" is coded "1", and the answer "no" is coded "0". This makes the mean of the variable Y207 equal to the probability of answer "yes".

Software: There are many software packages that are capable of calculating the sampling errors. The most common ones are SPSS, SAS, Stata and R.

If using SPSS (at least version 12) is required, with the complex samples module installed. The first step to calculate sampling error is to run the analysis preparation wizard. The syntax is given below:

```
CSPLAN ANALYSIS
/PLAN FILE= 'C:\Surveys\Sampling errors\Samperr.csaplan'
/PLANVARS ANALYSISWEIGHT=weight
/PRINT PLAN
/DESIGN  STRATA= stratum  CLUSTER= HPSU
/ESTIMATOR TYPE=WR.
```

The second step is to run the complex samples description command. The syntax is given below:

```
CSDESCRIPTIVES
/PLAN FILE = 'C:\Surveys\Sampling errors\Samperr.csaplan'
/SUMMARY VARIABLES =Y207
/MEAN
/STATISTICS SE COUNT CIN (95)
/MISSING SCOPE = ANALYSIS CLASSMISSING = EXCLUDE.
```

Results: The sampling error of the sample statistic rate of question Y207 is given in the table below. This means that the variable estimates range from 4.627 to 6.953 per cent with a 95 per cent confidence interval.

Univariate statistics

		Estimate	Standard error	95% Confidence interval		Unweighted count
				Lower	Upper	
Mean	Y207 Ever worked during school	.05790	.00590	.04627	.06953	3057