WHO Multi-country Survey Study on Health and Responsiveness 2000-2001

T. Bedirhan Üstün, Somnath Chatterji,
Maria Villanueva, Lydia Bendib
Can Çelik, Ritu Sadana,
Nicole Valentine, Juan Ortiz,
Ajay Tandon, Joshua Salomon,
Yang Cao, Xie Wan Jun,
Emre Özaltin, Colin Mathers
and Christopher J.L. Murray

GPE Discussion Paper 37

30 November 2001

TABLE OF CONTENTS

SUMMARY **1**

1. INTRODUCTION **3**

2. OVERALL GOALS OF THE SURVEY STUDY **6**

2.1 Specific Aims of the WHO Multi-country Survey Study *7*
2.2 The Specific Aim of Cross-population Comparability *7*

3. DEVELOPMENT AND CONTENT OF SURVEY INSTRUMENT **11**

3.1 Selection of Health Domains *11*
3.2 Formulation of Questions *12*
3.3 Operationalization of Cross-population Criteria *13*
3.4 Health State Descriptions *14*
3.5 Health State Valuations *14*
3.6 Development of the Responsiveness Module *15*
3.7 Content of the Responsiveness module *16*
3.8 Mental Health (Depression and Alcohol Use) *17*
3.9 Chronic Health Conditions *17*
3.10 Adult Mortality *17*
3.11 Health Related Areas: Environmental Factors *17*
3.12 Health Financing *17*

4. MODES USED IN THE WHO MULTI-COUNTRY SURVEY STUDY **19**

4.1 Household Individual Interviews *19*
4.2 Household Brief Face-to-Face Interviews *19*
4.3 Computer Assisted Telephone Interview (CATI) *19*
4.4 Postal Self-Administered Surveys *20*

5. INITIAL TESTING OF THE QUESTIONNAIRE **21**

5.1 Translation and Cognitive Interviewing *21*
5.2 Translation and Back-Translation *21*
5.3 Pilot Tests *21*

6. SITES **23**

6.1 Selection of Participating Countries *23*
6.2 Sampling Plans *23*
6.3 Training of Interviewers and Trainers *25*
6.4 Implementation of Survey in the Field *25*
6.5 Quality Assurance *26*
6.6 Feedback During Data Collection *26*
6.7 Ethical Aspects *27*

7. DATA COLLECTION AND MANAGEMENT 28

7.1 Data Coding 28
7.2 Data Entry Program 28
7.3 Data Transfer 28
7.4 Data Quality Checks 29
7.5 Data Cleaning and Feedback 29
7.6 Quality Assurance Steps for Data 29

8. DATA ANALYSIS STRATEGIES 30

8.1 Basic Descriptive Analysis 30
8.2 Specific Analyses for Health State Valuations 31
8.3 Specific Analyses for Health System Responsiveness 31

9. PRELIMINARY RESULTS 32

9.1 Participating Countries 34
9.2 Results by Survey Modes 34
9.3 Representativeness of Samples 36
9.4 Survey Length 37
9.5 Reliability of the Survey Instrument 38
9.6 Implementation of Measured Calibration Tests 40
9.7 Vignette Implementation 41
9.8 Overall Survey Metrics 42
9.9 Comparison of Survey Mode Effects 47
9.10 Comparison to Other Multi-country Surveys 51
9.11 Costs 54
9.12 Selected Results on Health Module 55
9.13 Selected Results on Responsiveness Module 56
9.14 Selected Results on Adult Mortality 57

10. DISCUSSION AND CONCLUSIONS 58

10.1 Feasibility and Utility 58
10.2 Use of Data 58
10.3 Approaches for Cross-Population Comparability 58
10.4 Survey Costs 59
10.5 Plans for the World Health Survey (2001-2002) 59

11. REFERENCES 63

APPENDICES 67

Appendix 1: List of Collaborating Institutions 71
Appendix 2: Detailed Sampling Plans for Participating Countries : 73
Appendix 3: Deviation of Study Sample from UN Population Database 140

LIST OF FIGURES

Figure 1: Response Category Shift – Different Rulers 8
Figure 2: Quality Assurance Steps for Data 29
Figure 3: Different Survey Modes in WHO Multi-Country Survey Study 32
Figure 4: WHO Multi-country Survey Study, 2000-2001 33
Figure 5: Example of a Sample Population Deviation Index 37
Figure 6: Sample-Population Representativeness by Survey Mode 46
Figure 7a: Comparison of Household vs. Postal Surveys for Representativeness 47

Figure 7b: Sample-Population Deviation Summary Index Comparison of Postal vs. Bftf Surveys 48
Figure 8a: Summary Quality Assessment of Household Surveys by Reliability and Missing Value 49
Figure 8b: Summary Quality Assessment of Household Surveys by Reliability and Representativeness 49
Figure 8c: Summary Quality Assessment of Household Surveys by Representativeness and Missing Value 49
Figure 8d: Summary Quality Assessment of WHO Surveys by Representativeness and Missing Value 50
Figure 8e: Summary Quality Assessment of WHO Surveys by Response Rate and Missing Value 50
Figure 8f: Summary Quality Assessment of WHO Surveys by Representativeness and Response Rate 50
Figure 9a: Representativeness of DHS Surveys by Country 52
Figure 9b: Relationship between Missing Data and Representativeness for DHS Surveys 52
Figure 10: Comparison of Representativeness across Different Surveys 53
Figure 11: Comparison of Cost by Mode 54
Figure 12: Adjusted and Unadjusted Mobility by Age 55
Figure 13: Adjusted and Unadjusted Self-care by Age 56
Figure 14: Adjusted and Unadjusted Dignity by Age 56
Figure 15: Plot of Weighted and Smoothed Log of Death Rates from Survey, Compared with Corresponding WHO 2000 Estimates for Females in Nigeria 57

LIST OF TABLES

Table 1: Assessment Instrument Domains 12
Table 2: Distribution of Completed Survey Countries by WHO Regions 33
Table 3: Countries Participating in the Survey Study 34
Table 4: Summary Survey Diagnostics for Different Modes 35
Table 5: Sex Ratio (Male/Female) in the UN Population Database and in the Household Survey Country Samples 36
Table 6: Timing of Individual Sections of Full-length Household Interviews from Pilot Phase 38
Table 7: Mean Interview Duration (Minutes) of Full-length Household Interviews – Main Phase 38
Table 8a: Reliability Statistics for Major Sections of the Health Component of the Survey 39
Table 8b: Reliability Statistics for Major Sections of the Responsiveness Component of the Survey 40
Table 9: Summary Results of Calibration Tests 41
Table 10a: Survey Metrics for the Household Survey –Health & Responsiveness Components 43
Table 10b: Survey Metrics for the Brief Face-to-Face Survey Health & Responsiveness Components 44
Table 10c: Survey Metrics for the Postal Survey – Health & Responsiveness Components 45
Table 11: Average Cost Per Completed Interview 54

SUMMARY

In order to develop various methods of comparable data collection on health and health system responsiveness WHO started a scientific survey study in 2000-2001. This study has used a common survey instrument in nationally representative populations with modular structure for assessing health of indviduals in various domains, health system responsiveness, household health care expenditures, and additional modules in other areas such as adult mortality and health state valuations.

The health module of the survey instrument was based on selected domains of the International Classification of Functioning, Disability and Health (ICF) and was developed after a rigorous scientific review of various existing assessment instruments. The responsiveness module has been the result of ongoing work over the last 2 years that has involved international consultations with experts and key informants and has been informed by the scientific literature and pilot studies. Questions on household expenditure and proportionate expenditure on health have been borrowed from existing surveys. The survey instrument has been developed in multiple languages using cognitive interviews and cultural applicability tests, stringent psychometric tests for reliability (i.e. test-retest reliability to demonstrate the stability of application) and most importantly, utilizing novel psychometric techniques for cross-population comparability.

The study was carried out in 61 countries completing 71 surveys because two different modes were intentionally used for comparison purposes in 10 countries. Surveys were conducted in different modes of in- person household 90 minute interviews in 14 countries; brief face-to-face interviews in 27 countries and computerized telephone interviews in 2 countries; and postal surveys in 28 countries. All samples were selected from nationally representative sampling frames with a known probability so as to make estimates based on general population parameters.

The survey study tested novel techniques to control the reporting bias between different groups of people in different cultures or demographic groups (i.e. differential item functioning) so as to produce comparable estimates across cultures and groups. To achieve comparability, the self-reports of individuals of their own health were calibrated against well-known performance tests (i.e. self-report vision was measured against standard Snellen's visual acuity test) or against short descriptions in vignettes that marked known anchor points of difficulty (e.g. people with different levels of mobility such as a paraplegic person or an athlete who runs 4 km each day) so as to adjust the responses for comparability . The same method was also used for self-reports of individuals assessing responsiveness of their health systems where vignettes on different responsiveness domains describing different levels of responsiveness were used to calibrate the individual responses.

This data are useful in their own right to standardize indicators for different domains of health (such as cognition, mobility, self care, affect, usual activities, pain, social participation, etc.) but also provide a better measurement basis for assessing health of the populations in a comparable manner. The data from the surveys can be fed into composite measures such as "Healthy Life Expectancy" and improve the empirical data input for health information systems in different regions of the world. Data from the surveys were also useful to improve the measurement of the responsiveness of different health systems to the legitimate expectations of the population.

1. INTRODUCTION

Countries need timely information on critical outcomes to evaluate their health policies, manage their health systems and monitor progress. Routine health information systems are meant to provide affordable and timely information. In most developing countries, these have long focused on civil registration systems for vital events and registries of services delivered through publicly owned facilities. Developed and developing countries, however, have increasingly recognized the role of periodic household surveys to fill critical information gaps in the data provided by the health information system. The large number of general and specialized household surveys fielded in countries at all levels of development used to measure child mortality, utilization, health financing, mental health, disease-specific outcomes etc. are an indication of the potential role of surveys in filling critical information gaps. In addition, information on aspects such as responsiveness of a health system is unavailable from health information systems and needs to be collected through surveys designed for this purpose.

Data from surveys in different countries, however, often have a serious problem of *comparability*. For example, responses of individuals vary by country or by population sub-groups due not only to real differences in the quantity of interest but also to differences in norms and expectations, or cognitive processing of survey questions. The general health reporting in EUROSTAT[1] surveys in 12 European countries has revealed a six-fold difference between the proportion of people reporting good or very good health in Denmark and Portugal. Similarly there is a four-fold difference between the proportion reporting bad or very bad health. This fact is not congruent with other health correlates such as mortality or health service use and may create serious problems if such data are used as a basis for comparison across countries or population groups (1, 2, 3, 4, 5). Such examples are manifold indicating the need to improve the comparability of self-reported health data.

To obtain comparable data, it is essential to pay great attention to questionnaire development. This requires clarity in what is meant by the concept under measurement (e.g. what domains should be included) and its operationalization in a survey instrument (e.g. question wording, response categories, the meaning of responses, use of a comparator against which individuals report their experiences, translation protocols and classical techniques for psychometric equivalence) are all important. In addition, there is the need to control for possible "differential item functioning" which involves a shift in the response category cut-points between populations or subgroups. This occurs when people at similar levels of health give different answers to describe their health (6).

In order to develop methods to gather comparable data across populations, WHO launched a survey study, the WHO Multi-country Survey Study, in 2000-2001 through a series of carefully designed steps that attempt to deal with the shortcomings in existing methods and arrive at common instrument modules and techniques suited to multiple user needs to measure health system performance outcomes. The first step in the process was to review existing survey instruments and cultural comparison techniques. More than 300 international tools used in more than 50 countries were systematically reviewed to identify their items, their utility in survey conditions and psychometric properties (7, 8). These included whether the questions were clear and unambiguous, translated in a meaningful way, have an identically interpreted response scale, good test-retest

[1] Statistical Office of the European Communities, European Commission.

reliability and have validity (e.g. concurrent validity with known reference tests or construct validity to predict other impacts). Where possible we examined the calibration properties of different instruments using methods derived from "Item Response Theory" which indicated whether different populations use similar cut-points in their rating of responses (9). We then took into account well known sources of bias in questionnaires including "social desirability", "central tendency" (i.e. aversion to end points), and other framing effects such as the "Halo effect", "Carry-over effect", and "Positive bias" (e.g. answers to questions are affected by other questions or by interview style) (10)

It is clear that measurements of health and health-related parameters need to be applicable cross-culturally, reliable, calibrated for relevant response categories and valid[2] (11). However, these characteristics are not sufficient to ensure cross-population comparability. In addition to the classical psychometric criteria, to make meaningful international and cross-population comparisons, an instrument should have a common metric in different populations, i.e. the same response level should correspond to the same level of health (or responsiveness) in a given domain. Evidence of equivalent metric properties should be shown by external calibration tests and other possible mechanisms. Comparability of results adds a new dimension to international survey instrument development. The difference between comparability on the one hand and validity and reliability on the other hand can be illustrated with two thermometers, one of which is Celsius, the other is Fahrenheit. Both thermometer measures give valid and reliable measurements of temperature. However, 26 degrees on one thermometer is not comparable to 26 degrees on the other thermometer.

Comparability is fundamental to the use of survey results for benchmarking and evaluation but has been under-emphasized in instrument development. WHO is trying to increase attention on the importance of comparability in instrument development. An example of a monitoring task that puts a premium on comparability and the gaps in existing health information systems is the comparative reporting by WHO of healthy life expectancy and responsiveness in the World Health Report 2000 (12). Such comparative reporting highlights the need for valid, reliable and comparable survey instruments to measure these outcomes. The list of outcomes for which valid, reliable and comparable survey instruments are needed is not restricted to health state and responsiveness but must also include coverage of critical health interventions, utilization of services, risk factors such as tobacco or alcohol use and disease specific outcomes. Notably, instruments with established

[2] *Validity* is the extent to which a survey instrument measures what is intended to measure. It describes whether the instrument as to how actually the instrument is able to capture the real nature of what is measured. This can be measured for an item, a series of items or overall instrument level. *Reliability* is the extent to which repeated use of the instrument gives the same result. Reliability is the consistency of the measurement, or the degree to which an instrument measures the same way each time it is used under the same condition with the same subjects. In short, it is the repeatability of the measurement. A measure is considered reliable if a person's score on the same test given twice is similar. One can think of reliability as measurement invariance or conversely the extent to which a measurement is subject to random measurement error. Validity is the extent to which a measurement correcting for random measurement error is correlated to the true level. In other words, validity can be understood as unbiasedness. Validity is the strength of our conclusions, inferences or propositions. It is the best available approximation to the truth or falsity of a given inference, proposition or conclusion. Construct validity is the degree to which inferences we have made from our study can be generalized to the underlying concepts in the first place. For example, if we are measuring mobility as an outcome, can our definition (operationalization) of that term in our study be generalized to the rest of the world's concept of mobility?

validity, reliability and comparability (i.e. the generalizability of findings across diverse populations) are not widely available.

Though WHO has been, as part of its mandate, collecting systematic information with regard to causes of death and morbidity in its member states around the world, more recently there has been a clear recognition that the impact of health conditions is perhaps better understood when non-fatal health outcomes are taken into account over and above mortality. Health is an abstract and complex concept, yet there is an expectation that we all have an intuitive universal understanding of health. Moreover, our notion of well-being encompasses areas that go beyond health and health is a fundamental human capacity that interacts with other areas of well being. The challenge is to separate the constituents of a health experience that are intrinsically health from those that are non-health or health related in clear recognition of the fact that health experiences are not context free (13).

It is now recognized that building a scientific base to inform policies, strategies and programs is essential. In this context, WHO is operationalizing a framework for the measurement of health and the impact of all such actions whose primary intent is to improve, restore or maintain health. To catalyze the development of valid, reliable and comparable survey instruments to measure key outcomes, WHO launched the multi-country national household survey study in 2000-2001. The purpose of this paper is to report on the objectives, design, instrument development and execution of this multi-country study.

2. OVERALL GOALS OF THE SURVEY STUDY

The WHO Multi-country Survey Study was a research exercise to develop instruments that would allow the measurement of health, responsiveness and other health-related parameters in a comparable manner and would provide useful information to refine this methodology. The Study focused on the way populations report their health and value different health states, the reported responsiveness of health systems and the modes and extent of payment for health encounters through a nationally representative general population-based survey.

The WHO Multi-country Survey Study had as its **first objective** the **assessment of health** in different domains using self-reports by people in the general population. The International Classification of Diseases (**ICD)** was used as the framework to gather information on mortality in the households and self reported morbidity (such as the diagnosis of depression, alcohol-related problems and other chronic health conditions) (14) and the revised International Classification of Functioning, Disability and Health (**ICF)** was used to describe the essential elements of non-fatal health outcomes (15). The survey also included vignettes and some measured tests on selected domains that were intended to calibrate the way respondents categorized their own health. This part of the survey allowed for direct comparisons of the health of different populations across countries.

A related objective of the WHO survey was to measure the value that individuals assign to descriptions of health states and to test if these varied across settings. These health states were described as decrements in major domains of body functions and activities. Valuations involve the construction of summary measures of population health taking into account both fatal and non-fatal health outcomes.

The **second overall objective** of WHO Multi-country Survey Study was to test instruments to measure the **responsiveness of health systems.** The concept of responsiveness is different from that of people's satisfaction with the care they receive, in that it examines what actually happens when the system comes in contact with an individual. It includes 2 major categories: respect for persons, which includes the respect for the dignity of the individual, confidentiality, communication and autonomy; and client orientation that consists of prompt attention, amenities of adequate quality, access to social support networks and the choice of institution and care provider.

Additionally the WHO Study aimed to test instruments in areas such as health expenditures, adult mortality, birth history, various risk factors and assessment of main chronic health conditions with additional modules.

2.1 Specific Aims of the WHO Multi-country Survey Study

More specifically the aims of the WHO Multi-country Survey Study were to:

1. Develop valid, reliable and comparable instruments to describe individual health states and health system responsiveness on a core set of domains and to test these in household surveys.

2. Test the validity of different modes of survey implementation including long-form household, short-form household, self-administered postal and computer-assisted telephone interview.

3. Contribute to the development of WHO's and Member States' capacity to field surveys with quality control, appropriate sampling, and data management strategies as well as to build capacity to analyse data from complex surveys.

4. Address critical methodological questions related to: identifying the most parsimonious set of questions that would suffice to adequately measure the health of a population in an efficient and cost-effective manner; and maintaining cross-population comparability.

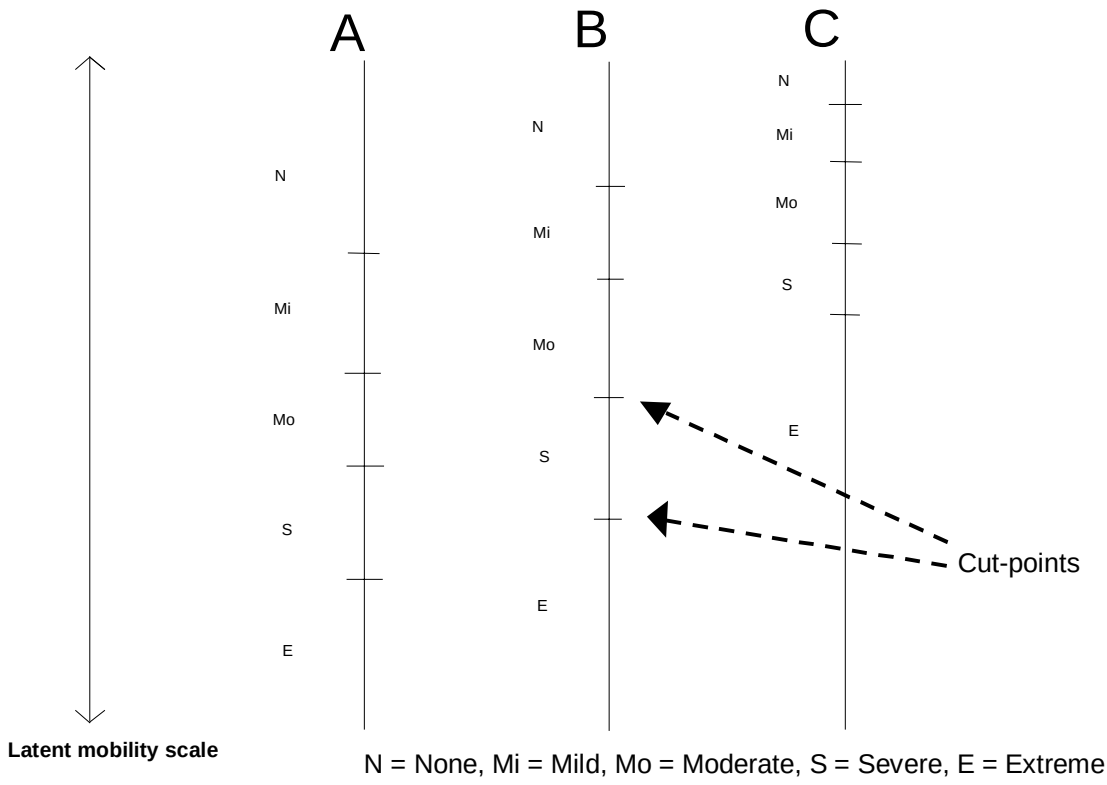
5. Test other candidate modules including adult mortality and health expenditure.

The focus of the current survey study has been to collect data in different modes across the world on a core set of health domains as well as on domains related to responsiveness. This is superior to analyzing existing data sets that have been collected in different parts of the world using a range of methods (4). The survey study achieves a better representation of multiple dimensions of health and health systems and obtains comparable data across countries. One basic objective is to systematically address issues related to the most parsimonious set of questions that would suffice to adequately measure the health of a population and the responsiveness of a health system in an efficient and cost-effective manner and maintain cross-population comparability.

2.2 The Specific Aim of Cross-population Comparability

There is sufficient evidence indicating that self-reports on health (or responsiveness or other quantities of interest) may not be comparable across countries or even across different socioeconomic sub-groups within a population (6). The problem of comparability can be conceptualized in terms of response category cut-point shifts. Even when reliability and within population validity have been well established, the meaning that different groups attach to the labels used for each of the response categories in self-reported questions can vary greatly (see Figure 1).

Figure 1: Response Category Shift – Different Rulers



For each domain, there is some "latent scale" for that domain that represents the true level, which by definition is unobserved. For example, let us assume that there is a underlying health domain for "mobility" and we attempt to assess mobility by asking respondents how much difficulty they have climbing stairs. For this self-reported survey question, the response categories are labeled by order of difficulty as "*no difficulty*", "*mild difficulty*", "*moderate difficulty*", "*severe difficulty*", and "*extreme/cannot do*", and A, B and C represent three different populations. When different individuals respond as "mild difficulty" in climbing stairs, the response could indeed map to different levels of mobility in different populations (e.g. from climbing a few steps to a few flights of stairs depending on their norms). The survey results would be reliable and valid within each population but the results cannot be compared across populations without adjustment.

We hypothesize that cut-points vary between populations because of different cultural or other expectations. Cut-points are also likely to vary within cultural, socio-demographic or other groups within populations: the cut points for older individuals may shift as their expectations for a health domain diminish with age; men may be more likely to deny declines in health so that their cut-points may be systematically shifted as compared to women; contact with health services may influence expectations for a domain and thus shift cut-points.

Comparable measurement of individual health and population health, or any other such quantity that is assessed through self-reports (e.g. responsiveness), therefore, requires explicit strategies to measure the response category cut-points of each item in different populations and socio-demographic groups. Methods to establish cut-points fall into two basic categories:

(a) a scale that is strictly comparable across individuals and populations is established and measurements on this comparable scale are used to establish the response category cut-points for each survey item; or

(b) categorical responses are obtained from different groups for a fixed level on the latent scale. If the level is fixed, variation in the responses provides information on the differences in cut-points across individuals and populations.

A comparable scale of measurement can be achieved in two different ways:

(a) when a domain is measured using multiple items, the underlying factor in the data may be, under certain assumptions, comparable;

(b) when a measured test – such as a Snellen's eye chart for vision or the posturo-locomotion-manual (PLM) test for mobility – can be used to establish a comparable scale. However, for some domains of health such as pain, reliable and valid measured tests may not exist or be affordable, or may even be unethical.

An alternative strategy for establishing cross-population comparability is to fix the level on a given domain and assess variation in the response categories across individuals, groups, and populations. In other words, if the level of mobility is fixed but one group gives the response category of "no difficulty" while another gives the response category of "some difficulty", that information can be used to assess the response category cut-points. The same strategy can be used to assess notions such as dignity or promptness of attention where the responses may range from "very good" to "very bad". Two ways of evaluating a fixed level comparisons and thus variation in cut-points, are available: **vignettes** and assessment of **comparable homogeneous groups**.

a) A *vignette* is a description of a concrete level of ability on a given domain that individuals are asked to evaluate. To assess the response category cut-points, each respondent is asked to assess the level for a hypothetical case described in a vignette. The vignette fixes a given level in the domain of interest such that variation in the response categories is attributable to variation in the response category cut-points. Vignettes fix the level on a domain and only cut-points need to be estimated.

b) *Comparable homogeneous groups* in different populations are used for comparing responses to an item. For example for health states, recent acute changes in health from injuries such as fractures might be used to identify reasonably comparable groups. Alternatively some lifestyle or occupational characteristic might be used to identify these groups such as a group of elite athletes. Similarly for responsiveness, all attendees at a given facility can be asked about their experience, for example, of prompt attention or dignity with an assumption that the facility treats all users belonging to a homogenous population (as defined by ethnicity, income, etc.) more or less the same.

The comparable homogenous group approach has, however, its limitations for two reasons. First, identification of the groups needs to be independent of any measurement of health status. Even when groups are identified through some factor such as an injury, doubts can always be raised as to the true comparability of groups. In the same way one can well imagine that a given facility indeed treats people belonging to different categories such as income or ethnic groups differently. It may be difficult to persuade people that apparent differences in the responses are due to cut-point shifts as

opposed to true differences in that domain between groups. Second, to be able to assess variation in response category cut-points for all response categories, a series of homogeneous groups must be used. Analytically, each homogeneous group is like one vignette. This means that the comparable group strategy can only work if several comparable groups are studied. Despite these limitations, it may be worthwhile assessing comparable groups as an adjunct to other methods.

The current set of surveys uses the vignette and measured test approach to calibrate self-report on health domains and the vignette approach alone for the responsiveness domains in order to make cross-population comparable analyses (6).

3. DEVELOPMENT AND CONTENT OF SURVEY INSTRUMENT

A. Health Module

3.1. Selection of Health Domains

In preparation for the module on health in the WHO Multi-country Survey Study an extensive review of existing instruments was carried out. This review was closely synchronized with the revision of the International Classification of Functioning, Disability and Health (ICF) which is a classificatory framework for components of health. ICF describes health and health-related domains (15). An item pool was constructed and the published psychometric properties of each question documented. Qualitative research identified the core constructs in different countries (8). In addition, the health section of the interview was presented to a WHO Committee of Experts that met in Geneva in August and September 2000. This list of domains was then presented to a group of experts in the measurement of health from all WHO regions at a UN/OECD meeting in Ottawa in October 2000. From the above item pool and based on the qualitative research, the health domain questions were selected according to the following criteria:

- they should be linked to a conceptual framework of the ICF;
- they should have face and construct validity, i.e. they must be linked to the intuitive, clinical and epidemiological concepts of health; and together the questions should be comprehensive-enough to reflect major health conditions;
- they should be amenable to self report;
- they should build on the existing knowledge base of common questionnaires;
- they should be cross-population comparable; and
- it should be possible for some domains to be linked to a calibration test.

The domains included in the WHO Survey Instrument are listed in Table 1.

Table 1: Assessment Instrument Domains

Health Domains	Health Related Domains	
• Vision	• Self-care:	*daily activities including eating*
• Hearing	• Usual activities:	*household activities; work or school activities*
• Speech	• Social functioning:	*interpersonal relations*
• Digestion	• Participation:	*societal participation including discrimination and stigma*
• Bodily excretion		
• Fertility		
• Sexual activity		
• Skin & disfigurement		
• Breathing		
• Pain		
• Affect		
• Sleep		
• Energy / vitality		
• Cognition		
• Communication		
• Mobility and Dexterity		

3.2 Formulation of Questions

Based on the selected health domains the questions were selected with reference to existing survey research instruments mainly from the work based on the Who Disability Assessment Schedule (WHO-DAS II) (7). The following guidelines were used in the construction of the health module for achieving reliable and valid assessment of health in different populations:

a) clear and unambiguous questions to allow for similar cognitive processing by the respondents, translation across different languages and application in different cultures

The language of the questions in the survey was intentionally chosen to reflect 'extent of difficulty' actually experienced by the individual in carrying out the tasks or actions. Further, the questions were framed so that the response categories were uniform and referents used were as concrete as possible (e.g. how much difficulty do you have in seeing and recognizing a person across the road, i.e. from a distance of about 20 meters?).

b) reliable recall period used as timeframe

For the purposes of this survey a period of one month was chosen, as recall is known to rapidly deteriorate beyond this period.

c) basic test-retest reliability as proof of consistent application

Questions that have earlier been demonstrated to have good reliability were preferred.

d) basic concurrent validity with known reference tests (or in-depth expert evaluations)

There is limited data with regard to this area. However, wherever possible, questions that have been demonstrated to have good correlation with measured performance on tests for that domain were chosen.

e) construct validity to predict other impacts, consequences or determinants (e.g. such as outcomes, service use, costs or other known variables).

f) invariance in the measurement properties across different populations

Modern psychometric techniques such as Item Response Theory (IRT) provide tools to detect relative shifts in cut-points (9). Each item in a questionnaire can be examined by its difficulty and the respondents' ability. Such analysis may display the items that work similarly across groups and those that demonstrate differential item functioning (DIF). However, if all items suffer DIF in a systematic and correlated manner usual IRT methods fail to identify this problem, because these models assume that cut-points on a scale are independent of each other and not a systematic function of the characteristics of individual respondents. Therefore newer techniques using item response modeling approaches need to be developed to deal with response category cut-point shifts to make meaningful comparisons of data.[3]

3.3 Operationalization of Cross-population Criteria

The criteria above (a - e) are necessary yet not sufficient for cross-population comparability. When systematic reporting biases occur (e.g. when poor people report their health to be better than the 'truth' since they may have lower expectations, or in certain cultures where people report their health to be worse than the 'truth' since reporting good health is thought to bring bad luck) some external mechanism must be found to get additional information that allows for making adjustments for these reporting biases. To achieve cross-population calibration properties in different populations some specific approaches were used in the WHO Multi-country Survey Study:

(a) *Calibration tests* for some of the key domains that summarize overall health such as the domains of vision, cognition and mobility were selected. Tests for these domains were chosen as they are relatively easy to carry out in a large-scale survey setting, do not require very specialized interviewer training, are not dependent on equipment infrastructure and have been used in other surveys comparing self report with performance. For near and distant vision, standard vision charts that use symbols were chosen instead of letters in order to ensure applicability in illiterate populations. For cognition, standard tests of verbal fluency, immediate and delayed recall of a word list, and a cancellation task for attention were chosen. Finally, the posturo-locomotion-manual task was chosen for mobility as a composite task involving different aspects of moving. These tests were discussed with international experts such as WHO's prevention of blindness programme, neuropsychologists, and movement experts from the NINDS, USA (16, 17, 18, 19, 20).

[3] The methods developed by WHO such as HOPIT and CHOPIT can be conceptualized as modified IRT models (21)

Calibration tests are an objective measurement of what the survey questions basically intend to measure in different cultures and can be used as a closer approximation of "truth". Performance on these allows one to adjust for biases in self-report. For example, the responses to the question on how well a person can see an object at arm's length can be calibrated against standard near vision tests.

(b) *Standard case vignettes* are well-described case stories with precise concrete levels of health status. They can be applied in different cultures and calibrations can be obtained. Each vignette is a description of a specific level of ability on a given domain that individuals are asked to evaluate. By using vignettes, the level on a domain is fixed and only the cut-points need to be estimated. Vignettes were developed for the seven major domains of health: mobility, self care, pain, affect, cognition, usual activities and vision. The vignettes spanned the breadth of the scale (*no difficulty or problem* to *extreme difficulty or problem*) and were discussed with international experts to ensure cross-cultural applicability and understandability.

3.4 Health State Descriptions

The health domains and questions were developed based on the experience in the revision of the International Classification of Functioning, Disability and Health (ICF) (15) and the development of assessment instruments linked to ICF such as the WHO Disability Assessment Schedule (WHODAS II) (7). The WHODAS II was conceived as a general health state assessment measure capable of being used for multiple purposes including epidemiological surveys. It covers the major domains of cognition, self care, mobility, interpersonal relationships, daily activities at work and in the household, and social participation and impact. Prior to the WHO Multi-country Survey Study, during 1997-2000 the WHODAS II was *separately* piloted in 21 centres across 19 countries in 1,431 subjects that spanned the adult age group (≥18 years of age). The analysis examined reliability, convergent validity with other assessment instruments as well as other performance measures, sensitivity to change following intervention and relationships to valuation of health states. Findings from the field trials show a stable factor structure that is replicable across countries and population groups, unidimensionality of domains, good test-retest reliability with a kappa 0.65 to 0.91 (7). Twenty-three items from the final version of the WHODAS II, as well as items that were used for the WHODAS II development item pool during early phases of testing, were used in the survey study questionnaire in the domains of pain, vision, hearing, cognition, self-care, mobility, usual activities, interpersonal relations and social participation.

3.5 Health State Valuations

Health state valuations, (also known as preferences or disability weights) representing overall assessments of the levels of health associated with different states. They provide the critical link between the non-fatal health experience and mortality in summary measures of population health (such as healthy life expectancy or disability-adjusted life years, among others).

Measurement of health state valuations depends not only on assessments of the health levels in different states, but also on other values such as attitudes toward risk and uncertainty, distributional concerns, or preferences for immediate rather than future outcomes (referred to as time preference).

A related issue is that measurement methods are highly abstract and cognitively demanding and have demonstrated poor reliability and validity in the general community.

In order to address these problems, the WHO multi-country survey study, besides obtaining actual self-reports on different domains of health, asked respondents to value their own, as well as a set of other, health conditions. A brief description of each health condition was provided and respondents were asked to rate each of these on the seven core domains. Respondents were then asked to rank the different health conditions and then rate them on a visual analogue scale where 0 is death and 100 is perfect health.

In addition, more detailed surveys using multiple methods for valuation, such as Time Trade Off (TTO), Person Trade Off (PTO) and Standard Gamble (SG), were carried out among respondents with high levels of educational attainment in the same sites. The primary objective of the population-based surveys was to collect information on health state valuations in the general population in order to better understand differences across countries and within countries by age, sex, education, income and other variables. The multi-method study was designed to allow empirical adjustment of the valuations obtained in the general population surveys to account for the scaling properties of the simple measurement instrument used in these surveys. Because the multi-method exercise involves valuation methods that are abstract and cognitively demanding, it was implemented among individuals who were educated and were willing to undertake such a task and included students, journalists and policy makers, care providers and health care professionals. Some individuals with physical (such as diabetes or fractures) and mental health conditions (such as depression or substance use disorders) were also able to undertake these exercises with adequate explanations and visual props.

The simple visual analog scale used in the general population surveys provides a relatively simple measurement tool for assessing the valuation of health levels associated with the hypothetical states. The scaling properties of the visual analog scale have been challenged: mild states may be overvalued due to scale distortions. Nevertheless, once the nature of the distortion has been defined as a function, it allows for empirical adjustment of valuations. The study provided information on how the visual analog scale relates to other valuation techniques in order to estimate the underlying health state valuations that inform responses to all different measurement methods. By formalizing the relationships between the different valuation techniques and the underlying quantity of interest based on previous theoretical and empirical findings, statistical methods were used to recover these underlying valuations and to simultaneously characterize the nature of the scale distortion in visual analog scale responses (as well as to quantify other values such as risk aversion, distributional concerns and time preference). The product of this analysis is a function that may be used to adjust visual analog responses to the appropriate scale for valuations (22, 23).

B. Responsiveness

3.6 Development of the Responsiveness Module

Responsiveness of health systems to the legitimate expectations of populations is recognized as an important goal. To operationalize this concept and measure it meaningfully in different settings a

survey instrument was developed for the long household survey, postal and brief household surveys. The content of the household and postal instruments did not differ much, but there were fewer items in the postal and brief household questionnaires. [4]

The questions in the Responsiveness module were field-tested and comments on the questionnaire taken from a number of experts. As part of the development of the existing questionnaire, a key informant survey was run initially in 35 countries across 1791 individuals. In addition, 3 pilot household surveys were conducted in Tanzania, Colombia and the Philippines (about 150 individuals per country). Based on this experience, and in consultation with several international experts, a new questionnaire was developed.

The questions in the current household and postal survey were field tested prior to their finalization as part of the pre-testing of the entire survey study instrument (24).

3.7 Content of the Responsiveness Module

Within the responsiveness section of the survey subjects were asked if they have had an outpatient, home care or inpatient contact with the health system. They were asked the name of the last place of care and to identify whether this was their usual place of care. They were then asked to rate their experiences over the past 12 months. They were also asked about their utilization of health services over the last 30 days. The questions on responsiveness covered 8 domains, all relevant to inpatient visits but only 7 used for outpatient visits. Social support was the domain asked only to inpatients. The 8 domains and the corresponding number of questions per outpatient and inpatient domain were: prompt attention (4 outpatient, 1 inpatient), dignity (4 outpatient, 1 inpatient), communication (4 outpatient, 1 inpatient), autonomy (4 outpatient, 1 inpatient), confidentiality (2 outpatient, 1 inpatient), choice of institution and care provider (3 outpatient, 1 inpatient), and basic amenities of acceptable quality (3 outpatient, 1 inpatient).

All domains included a summary rating question (scaled 1 to 5, very good to very bad). In addition, several domains included report questions on how often a particular experience had occurred during encounters with the health system (scaled 1 to 4, always to never).

All questionnaires on responsiveness included vignettes, descriptions of hypothetical scenarios, which respondents are asked to rate using the same rating scale used in the responsiveness questions ("very good" to "very bad"). These vignettes were pre-tested on a small group of people.

[4] In addition, a "Key Informant Survey" with similar questions given to selected key informants (e.g. providers, consumers, policy makers, media workers etc) was also developed to test possible concurrence of its results with the personal report of responsiveness construct in the same set of countries. Key informants gave their opinions of their health system responsiveness of the public and private sectors, the extent of unequal treatment and experiences for different population groups within their country, how they measure and value different states of inequality in responsiveness, and how they value the importance of the different responsiveness domains within the overall construct. The Key Informant Survey is an explorative validation study, which is currently under way, and its results will be reported separately (25).

C. Other Modules

3.8 Mental Health (Depression and Alcohol Use)

The survey questionnaire, based on WHO Composite International Diagnostic Interview (CIDI)'s Depression and Alcohol Sections, specifically screened for depression and alcohol use disorders using questions that have been used extensively in many international studies (26). Since depression and alcohol related disorders are a major cause of health burden world wide it was felt necessary to measure these in a comparable manner across countries. Further, since these conditions are associated with stigma in all societies and produce substantial restrictions in participation, it was necessary to also measure health state valuations associated with them.

3.9 Chronic Health Conditions

Though there is conflicting evidence on the reliability of self-reported morbidity in different surveys, information was gathered on this area in the current survey using a checklist of chronic health conditions (27). In addition, to improve the validity of the information, respondents were asked whether the diagnosis was made by a doctor, what investigations were carried out, whether any specific treatment was received and what impact the health condition had on the person's life. This allowed for the relationship between self-reported health status and morbidity to be estimated across different populations.

3.10 Adult Mortality

The Multi-country Survey Study included a module that asked questions related to adult mortality in the past two years in each household. This was to supplement information on overall rates of adult mortality in countries where these data were not available from vital registration systems, and to explore if it was possible to obtain reliable information on some of the causes. Information from the head of the household was collated with information from other sources such as medical records wherever available (28). Attempts were made to compare the information thus obtained with the life-tables that are available for each country to determine the validity of this mode of data collection on mortality.

3.11 Health Related Areas: Environmental Factors

The survey questionnaire also collected some basic information related to environmental risk factors in the form of mode and place of cooking and the kinds of fuel used for cooking. In addition to providing minimal information about the risks to health depending on the cooking environment, it was expected that this information would also be correlated with socioeconomic status (29).

3.12 Health Financing

The WHO Multi-country Survey Study included questions on health expenditure and financing within the context of the interview. Respondents were also asked to provide information on the relative proportion of household income that is spent on health as compared to accommodation and

food (30). Since respondents answer the questions within a health survey, and therefore in the context of health, it is estimated that the information obtained is more relevant to the health experience than other expenditure and income surveys.

4. MODES USED IN THE WHO MULTI-COUNTRY SURVEY STUDY

The WHO Multi-country Survey Study provided survey content in different modes with different possible sampling strategies. The modes were in-person individual interviews, telephone surveys, and mail surveys with a view to test the mode effect on the parameters in question as well as the comparison of efficiency and cost of various applications. The basic modes that were used are described below and shown in the summary in Figure 3 (see page 32).

4.1 Household Individual Interviews

Interviews for the household survey were conducted face-to-face using paper and pencil questionnaires. In each household a single adult individual (>18 years) was selected by a random process (i.e. Kish Table which identifies a predefined individual in the household with a known probability) after completing a full household roster. The survey protocol specified that all interviews should be conducted in privacy. Where members of the household, neighbours or friends were present, the interviewers requested privacy and, where necessary, steps were taken to ensure that interviewers were the same sex as the respondent.

4.2 Household Brief Face-to-Face Interviews

In view of the costs of carrying out a full face-to-face interview (lasting around 90 minutes) and the need to carry out the survey in as many countries as possible, a briefer version of the questionnaire (around 30 minutes) was carried out in a face-to-face interview in several countries. This version focuses on selected key domains of health and responsiveness.

4.3 Computer Assisted Telephone Interview (CATI)

In two countries, where telephone coverage is extensive the brief survey (30 min), described above, was administered using this format. The telephone interviews use computer technology to automatically sequence and branch questions, which eliminates interviewer error in failing to ask questions. They can achieve a better sampling coverage because of the known sampling frame and random digit dialing.

4.4 Postal Self-Administered Surveys

Since it is relatively inexpensive to carry out a postal survey in countries where literacy levels are high and the reach of the postal system is good, the brief survey questionnaire was used in a mail format in many countries. In some countries (i.e. Turkey and Egypt) the survey was hand couriered to the respondents and collected back from them.

The survey was carried out in some countries using more than one mode. This has allowed the data from the different modes to be compared in order to estimate the effect of the mode of the survey.

5. INITIAL TESTING OF THE QUESTIONNAIRE

5.1 Translation and Cognitive Interviewing

The survey questionnaire was translated using a translation and linguistic analysis protocol developed by WHO. Forward translation was carried out by health experts and a bilingual group examined the accuracy and appropriateness of the translation. A report on linguistic analysis and translation of all key terms used in the questionnaire was requested from all sites. Back-translations were made by an independent group of linguistic experts. A team at WHO reviewed the reports and finalized the questionnaires.

Following translation, the instrument was pilot tested on 100 subjects at each site. During this test the respondents were asked specific debriefing questions in order to determine if the survey questions were understood and if the intent of the question was accurately conveyed. Respondents were also asked to elaborate on the reasons why a particular response category was chosen for a question. This information was analyzed to see if the instrument was being used in the same manner across sites and if it was feasible to obtain information in a consistent manner.

5.2 Translation and Back-Translation

Sites were asked to translate and back-translate a list of 145 items from the questionnaire. The aim of this process was to achieve different language versions of the English questionnaire that are conceptually equivalent in each of the countries. The focus was on cross-cultural and conceptual, rather than on literal/linguistic equivalence.

Some words or phrases were somehow problematic and did not convey the concept addressed by the original item. In some countries (e.g. Indonesia), scales of time (some of the time, a good bit of the time, most of the time) were back-translated as scales of frequency (occasionally, often, every time). Other terms gave a similar, yet different, meaning. For example “distress” was back-translated as “pain, anguish, stress or difficult/dangerous situation”. Some items had no equivalent in the local language and were difficult to translate. For example, in Nigeria, the term “bipolar disorder”, when back-translated gave the equivalent of “mental problem”.

5.3 Pilot Tests

The instrument was piloted in 100 respondents at each household survey site (ten countries) and 50 respondents were also re-tested to determine reliability. Feedback was given to sites based on the qualitative experiences as well as the quantitative results. In countries which applied other modes (brief face-to-face, telephone and postal surveys) a limited number of pilot tests were completed (between 10-50 interviews) to test the adequacy of translations and feasibility of applications.

The pilot study was very informative and provided valuable information for modification of substantial portions of the interview. The results of the pilots indicated that the test-retest reliability of self-report, vignettes and the calibration tests were high though the test-retest for calibration tests varied by country. In view of the findings of the pilot work the vignettes were modified and the

manner of implementation of the performance tests were discussed with each site. Further quality assurance steps were introduced including a detailed video for training; supervising tests during site visits; and standardizing these tests in each country with self-report and vignettes in controlled conditions.

The pilot data also served to reduce the questionnaire length by about 30%. Items that had very poor reliability and were not providing additional information on a domain were either deleted or combined. Questions that were unclear were rephrased. The pilot highlighted problems related to the length of the overall questionnaire and hence in the main study respondents were allocated to different rotations in the modules of the questionnaire (e.g. half the respondents were asked the valuation module and the other half were administered the calibration tests.) The pilot clearly demonstrated the value of vignettes and calibration tests: vignettes were added for multiple domains of health and responsiveness and modified to reflect more clearly a level of function in the specific domain.

6. SITES

6.1 Selection of Participating Countries

Identifying potential survey partners was based on interest in the study, previous survey experience, capability for conducting high-quality surveys at national level, available resources, and acceptable budget proposals and timeframe. An extensive search was done among universities, research institutions and other international agencies. Potential collaborators were asked to provide details on the surveys they had carried out, submit a sampling frame, a detailed budget breakdown and a plan of action for conducting the survey. Geographic distribution and country development was also taken into account. Overall, more than four hundred contacts were made in over 100 countries and this process took around four months. On average, three proposals per country were received. Criteria for site selection included:

- Former experience in health surveys in various modes.
- Quality of the proposal in terms of its response to technical specifications.
- Ability of the institution to carry out the survey in a timely manner.
- Survey Costs.
- Access to representative general population samples.

Based on evaluation of the proposals, consultations with technical experts, and feedback from WHO Regional Offices, Country Representatives and consultants, decisions were made with regard to which sites to select. Seventy-one surveys were conducted in sixty-one countries.

6.2 Sampling Plans

To develop sampling plans, an up-to-date registry of all persons residing in the country was preferred as the sampling frame. When such listings were not available, registries providing postal coverage or post office listings were considered.

The sampling plans were implemented in different survey modes as follows. Post stratification corrections were made based on the population data available from the UN population databases.

6.2.1 Household Survey

Generally sample sizes of between 5,000 to 10,000 male and female adults aged above 18 years, non-institutionalized and living in private households, were obtained (with the exception of Slovakia where 1183 respondents were sampled) as a multistage stratified probability sample that was reasonably representative of the country's population. In each sampling stage of the design, probability methods were implemented to ensure that a representative sample of the target population was obtained. Each site was asked to develop a sampling plan together with WHO and submit the final sample description and calculation of weights according to the technical specifications. Some countries already had a sampling frame in place used at the Government level, e.g. Central Agency for Public Mobilization and Statistics (CAPMAS) in Egypt. One respondent per

household was randomly selected using the Kish tables (31) that allocates equal probability of selection to each eligible individual within the household. Quota sampling methods or other respondent selection procedures that are not probabilistic were considered unacceptable.

In three countries (China, India, Nigeria) survey samples were not representative of the whole country because of the size of the countries and language barriers. In China, the study was carried out in the 3 provinces of Gansu, Henan and Shandong; in India the study was carried out in the state of Andhra Pradesh; in Nigeria it was carried out in the Yoruba speaking regions of Ibadan, Iseyin, Ido and Ogo of Oyo state.

6.2.2 Brief Face-to-Face Survey

Male and female adults older than 18 years, non-institutionalized and living in private households were selected as a multi-stage random cluster sampling. The number of respondents ranged from 489 in Iceland to 3000 in Croatia. The sampling points represented the whole territory of the country surveyed and were selected proportionally to the distribution of the population. Metropolitan, urban and rural areas were covered with the exception of a few countries, where only urban areas were covered as they represented most of the country and where rural areas were remote and difficult to access. Sample weights based on probability of sampling for each individual were estimated.

6.2.3 Postal Survey

Male and female adults older than 18 years, non-institutionalized and living in private households, were taken into the sampling frame from existing address lists and mailed the questionnaire. In those cases where an acceptable sampling frame of individuals was unavailable, households were selected. In that case, the recipient of the mailed survey was then asked to select household members older than 18 years and ask the person with the closest birthday to complete the questionnaire. The questionnaire included a question on the number of people aged 18+ residing in the household to help develop weights for the data. Around 5000 questionnaires were mailed in each country (except Canada where around 1500 questionnaires were sent). The sample sizes ranged from 705 (out of 3000 mailed questionnaires) in the Republic of Korea to 4524 (out of 5000 mailed questionnaires) in Turkey. In some countries (i.e. Turkey and Egypt) the survey was hand couriered to the respondents and personally collected from them (32).

6.2.4 Computer Assisted Telephone Interview (CATI)

One thousand male and female adults aged over 18 years, non-institutionalised and living in private households were taken into the sampling frame. Two variations of the random digit dialing (RDD) were used in both countries in which CATI was implemented.

Random digit dialing was used in Luxembourg to include new and unlisted numbers. This procedure is designed to overcome problems of sampling from telephone directories, which is the usual sampling frame for telephone surveys. Directories are often inaccurate and out-of-date. They are also incomplete because of unlisted numbers (33).

In Canada telephone numbers were selected from the most recently published telephone directories, with a fixed number of telephone numbers per sampling unit, to provide a probability sample.

In both countries, from within each household contacted, respondents were screened selected using the "most recent birthday" method.

6.3 Training of Interviewers and Trainers

A training workshop for the Household Survey countries was held at WHO in Geneva from 3-6 July 2000 and two participants per country were invited to participate as trainers for the country survey teams. The ten participating countries were China, Colombia, Egypt, Georgia, India, Indonesia, Lebanon, Nigeria, Slovakia and Turkey. The purpose of the workshop was to train the trainers in conducting the household survey, and share ideas and concerns in a multicultural setting. The training covered sampling, interviewing techniques, questionnaire review and practice, and general issues related to the survey.

Additionally, training manuals and videos were distributed to each site for ongoing training. The training manuals clearly described the question by question specifications for the survey instrument and provided instructions to interviewers in terms of the "dos and don'ts" for the interview. It also provided guidance as to what clarifications the interviewers could provide if asked by the respondents. The videos illustrated basic interviewing techniques and some difficult interview situations and clarified issues related to specific questions. They also demonstrated how the calibration tests and valuation exercises were to be carried out. All this material was made available on different media as well.

In countries where brief face-to-face surveys were undertaken, interviewers were trained to follow standard instructions in a rigorous manner under the responsibility of the Project Manager and the Fieldwork Supervisors. Site visits for training and supervision of interviews were carried out in all countries, for quality assurance purposes.

6.4 Implementation of Survey in the Field

Surveys were conducted in various countries in three different modes. Sampling plans approved by WHO were implemented with specifications of the sampling units and stratification procedures at each sampling stage (primary, secondary and tertiary sampling levels). Several contact calls (at least four in the BFTF and ten in the household mode) were attempted and interviewers tried to contact each selected household at different times of the day and days of the week. Each contact call was recorded together with reasons for non-response.

Interviewers were supervised on a regular basis during fieldwork to ensure that expectations and production requirements were met, interviewers were performing well, information was kept confidential and professional ethics were followed, questionnaires and other materials were completed accurately and submitted on time, and lastly, that any problems were reported as soon as they arose. WHO asked supervisors to sit in on at least 10 interviews during the pilot phase to check

that interviews were conducted in a standardized way. The data was entered in the following days of paper-pencil instrument finalization after editing and approval by the supervisors. Each country made a report on the following aspects of the survey implementation:

- Details on each stage of sampling
- Quality control procedures implemented in the fieldwork
- Response rates and efforts undertaken to increase this, and the effects of these incentives
- Qualitative reports on the implementation process from the fieldworkers.

6.5 Quality Assurance

In order to monitor the quality of the data and ensure that countries complied with WHO guidelines in all household surveys the conditions under which the interviews were conducted and the problems that survey teams encountered were observed by supervisors first hand. Supervisors reviewed 10% of the questionnaires to check if options had been recorded appropriately and if questions were skipped correctly. About 10% of respondents were called or visited by the supervisor to ensure that the interview had been done, and 10% of all interviews were repeated by another interviewer within a period of one week to check for the reliability of the interview.

In addition, a site visit was scheduled to all full-length household survey sites during data collection. During these site visits several activities were undertaken:

- Overall survey management: sampling procedures; training/supervision; selection of respondent; and timing of survey were discussed.
- Interview assessment: the WHO staff sat in at least 4 interviews to see how the interview was conducted, the interaction between interviewer and respondent, and the timing of the interview.
- A meeting with the survey team was held to discuss contacting procedures, interviews, data and logistics.
- The data in questionnaires was checked by examining the survey records and data entry program.

Site visits made in the early phases of the data collection detected any problems, ensured that the questionnaire was administered and completed correctly, and confirmed that calibration tests were performed according to the instructions provided by WHO.

6.6 Feedback During Data Collection

The data was sent to WHO in a weekly or fortnightly basis such that a quick assessment could be made of the survey for each country in terms of missing information, reliability, use of appropriate skips, etc. Following data submission certain computerized algorithms were run to identify possible errors whilst the survey teams were in the field. Feedback regarding the data quality was routinely given to the site coordinator who took relevant action to ensure good quality data.

6.7 Ethical Aspects

To ensure that the set of survey studies was carried out with the highest standards, a working group was established with varied expertise in carrying out large-scale multinational surveys. This group met on a weekly basis to review the state-of-the-art in survey design and implementation so that benefits of the surveys to the participants and public were maximized and potential hazards minimized. In addition consultations were held with a committee of experts who agreed to the overall design of the questionnaire. Periodically, on an as-and-when-needed basis, discussions were also held with international experts to get further guidance with regard to this survey. The survey study core protocol and processes were cleared by the Sub-Committee for Research Involving Human Subjects (SCRIHS) at WHO.

Informed consent was obtained from every respondent in the survey. The consent form was carefully drafted in keeping with internationally accepted standards and in discussion with WHO's SCRIHS. Further, confidentiality agreements with each collaborating site were executed whereby the principal investigator at each site assured confidentiality of data. A draft confidentiality declaration from collaborators was discussed with SCRIHS and agreed upon. Ethical bodies at each site also reviewed the confidentiality issues of particular local relevance as part of the process of obtaining local ethical clearance.

Some personal identifying information was required for purposes of re-contact for testing the validity of the questions. However, once data was collected from the respondent it was stored in an anonymous manner with all identifying information stripped off the data set. Data transmitted to WHO contained no personal identifying information. Locally stored data was under the personal control of the principal investigator at each site and kept secure in agreement with local ethical committees.

7. DATA COLLECTION AND MANAGEMENT

7.1 Data Coding

At each site the data was coded by investigators to indicate the respondent status and the selection of the modules for each respondent within the survey design. After the interview was edited by the supervisor and considered adequate it was entered locally.

7.2 Data Entry Program

A data entry program was developed in WHO specifically for the survey study and provided to the sites. It was developed using a database program called the *I-Shell* (short for Interview Shell), a tool designed for easy development of computerized questionnaires and data entry (34). This program allows for easy data cleaning and processing.

The data entry program checked for inconsistencies and validated the entries in each field by checking for valid response categories and range checks. For example, the program didn't accept an age greater than 120. For almost all of the variables there existed a range or a list of possible values that the program checked for.

In addition, the data was entered twice to capture other data entry errors. The data entry program was able to warn the user whenever a value that did not match the first entry was entered at the second data entry. In this case the program asked the user to resolve the conflict by choosing either the 1st or the 2nd data entry value to be able to continue. After the second data entry was completed successfully, the data entry program placed a mark in the database in order to enable the checking of whether this process had been completed for each and every case.

7.3 Data Transfer

The data entry program was capable of exporting the data that was entered into one compressed database file which could be easily sent to WHO using email attachments or a file transfer program onto a secure server no matter how many cases were in the file.

The sites were allowed the use of as many computers and as many data entry personnel as they wanted. Each computer used for this purpose produced one file and they were merged once they were delivered to WHO with the help of other programs that were built for automating the process.

The sites sent the data periodically as they collected it enabling the checking procedures and preliminary analyses in the early stages of the data collection.

7.4 Data quality checks

Once the data was received it was analyzed for missing information, invalid responses and representativeness. Inconsistencies were also noted and reported back to sites.

7.5 Data Cleaning and Feedback

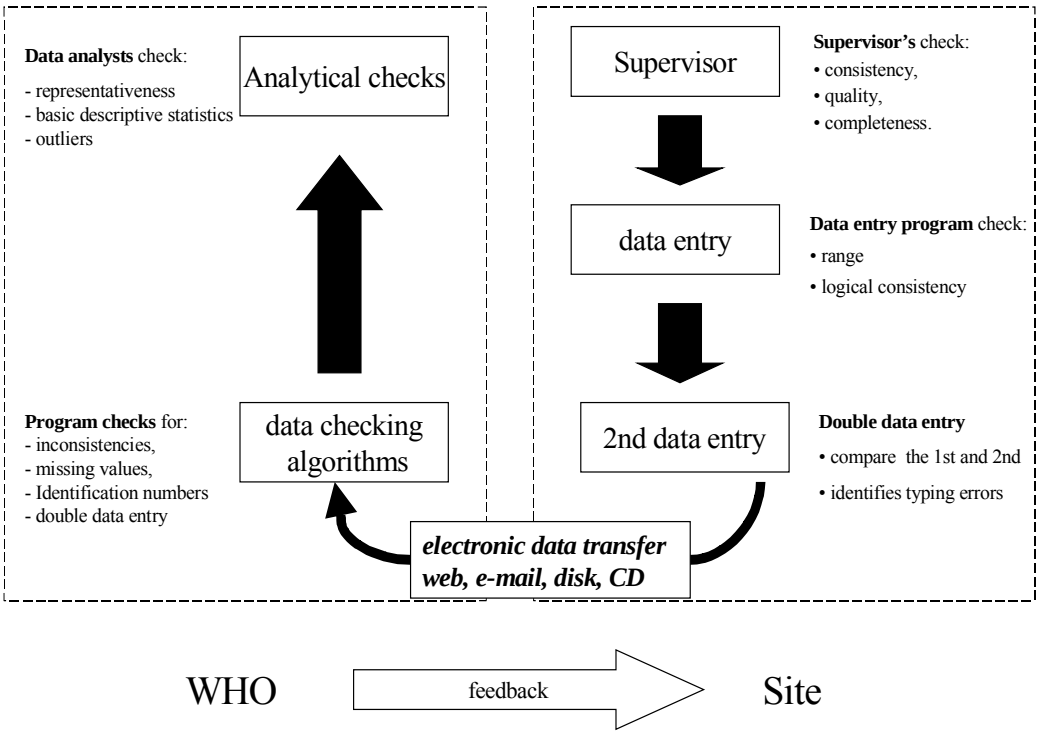
After receipt of cleaned data from sites, another program was run to check for missing information, incorrect information (e.g. wrong use of center codes), duplicated data, etc. The output of this program was fed back to sites regularly. Mainly, this consisted of cases with duplicate IDs, duplicate cases (where the data for two respondents with different IDs were identical), wrong country codes, missing age, sex, education and some other important variables.

7.6 Quality Assurance Steps for Data Collection

As noted above each record was entered twice using the data entry program that checked for inconsistencies between the two entries in order to minimize errors. Besides, the steps outlined above also ensured periodic corrections and checking.

The figure below summarizes the steps in the Quality Assurance process.

Figure 2: Quality Assurance Steps for Data



8. DATA ANALYSIS STRATEGIES

8.1 Basic Descriptive Analysis

8.1.1 Univariate, Bivariate and Multivariate Statistics

Once the data had been cleaned using standard procedures, univariate (i.e. frequencies, descriptives, etc.), bivariate (i.e. Pearson's and Spearman's correlation coefficients, crosstabulations, etc.) and multivariate (i.e. analysis of variance, regression analysis) analyses were carried out overall, for each country, and between countries using a pooled sample.

8.1.2 Item Analysis: Classical Approach

All the items of the survey were subject to traditional item analysis, using standard statistical procedures.

Endorsement rates were determined for each survey item by calculating the proportion (*p)* of people choosing a particular answer (i.e. '*Yes*').

Cronbach's alpha coefficient was calculated on the scales (factors) resulting from the exploratory factor analysis (EFA) and confirmatory factor analysis (CFA), to estimate the internal-consistency of each new composite score. The classical index of discrimination was obtained by calculating the corrected item-total correlation coefficients (r) for each item with its scale.

Test-retest reliability indices were computed within and across populations using *kappa statistics* for categorical and *intraclass correlation coefficient* for continuous variables to give estimates of chance corrected agreement rates for concordance between test and retest applications to indicate the stability of the application. These were estimated only for the full-length household surveys.

8.1.3 Item Analysis: Item Response Theory Approach and HOPIT

In addition to the methods described above, the items from the health status description section were also calibrated using the responses from the subjects on the vignettes and the performance tests. Since self-reports are dependent on the semantic value ascribed to words in a rating scale, in order to make valid comparisons between subjects within and across populations, it is imperative to correct for systematic biases in self-report such that comparisons indeed reflect true similarities and differences in the underlying trait of interest, in this case 'health'. Strategies such as the IRT approach and a hierarchical ordered probit model (HOPIT) that WHO developed for this purpose were used. These strategies are designed to correct for biases in cut-points in rating scales and are fully described elsewhere(21).

8.2 Specific Analyses for Health State Valuations

The results of the valuation component of the survey were descriptions and valuations of 11 health states by each respondent. The 11 health states include the individual's own health state plus 10 other hypothetical states drawn from four different sets. Health state descriptions consist of levels on each of the seven domains of health included in the survey. The first level of analysis was an examination of the different descriptions provided by respondents for the health state labels. Frequency distributions and summary statistics were compared across sites, and analysis of determinants of variation both across and within sites were undertaken using ordered probit and other models. This analysis was combined with the analytical efforts to estimate cut points on underlying domain scales using calibration vignettes and tests as mentioned above.

Health state valuations for each state, obtained from the visual analogue thermometer scale, were examined for differences across and within sites. The primary analytical objective in the valuation section was to estimate the valuation function through which health state profiles are mapped to valuations, i.e. the relative weights and interactions of different domains. Hierarchical models were developed in order to estimate these valuation functions in a way that accounts for variation in these functions according to different locations and individual characteristics such as age, sex, etc.[5]

8.3 Specific Analyses for Health System Responsiveness

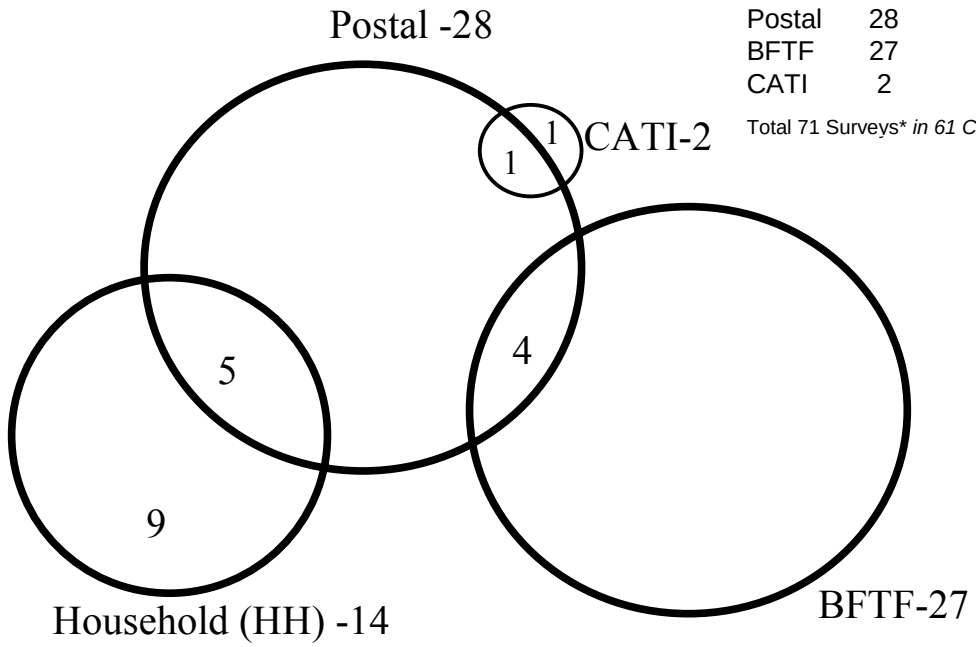
As in the case of health state measurements, an important part of the measurement and analysis strategy for responsiveness was the inclusion of vignettes, which described different responsiveness experiences. Analyses of responsiveness scores (inpatient and outpatient) for all elements with the variables of sex, age, health state, health service utilization, income, rural/urban and education were conducted to see how responsiveness results differ across these subgroups. Inpatient and outpatient element scores were compared across all elements. Classical test theory approaches were used to examine the factor structure of the construct of responsiveness. If a unidimensional structure is demonstrated, and a set of common items are identified that have similar properties across populations, the scale cut-off points for the responsiveness vignettes across countries is calculated using HOPIT and compound HOPIT (or CHOPIT) techniques. This exercise enabled calibration of the results if necessary, improving the interpretation of cross-country comparisons.

[5] In WHO Multi-country Survey Study Household Sites, an additional multi-method valuation study among educated respondents was undertaken in order to estimate the relationship between health state valuations elicited using the visual analouge scale and the underlying strength of preference function that is required in the construction of summary measures of population health

9. PRELIMINARY RESULTS

WHO Multi-country Survey Study in 2000-2001 was carried out in 61 different countries using 71 surveys. There were a total of 188,307 respondents in different modes including 14 full-length Household Surveys, 27 Brief Face to Face Surveys, 28 Postal Surveys and 2 Telephone (CATI) Surveys. Figure 3 shows the different modes and Figure 4 shows the regional distributions with Table 2 summarizing the geographical distribution by WHO Regional Offices.

Figure 3: Different Survey Modes in WHO Multi-country Survey Study, 2000-2001



Repeat designs in two modes

- HH and postal - China, Egypt, Indonesia, Lebanon, Turkey
- Postal and BFTF - Czech Republic, Finland, France, Netherlands
- Postal and CATI - Canada

*(*surveys in Iran, Lebanon (household), Singapore and Syria in progress)*

Figure 4: WHO Multi-Country Survey Study, 2000-2001

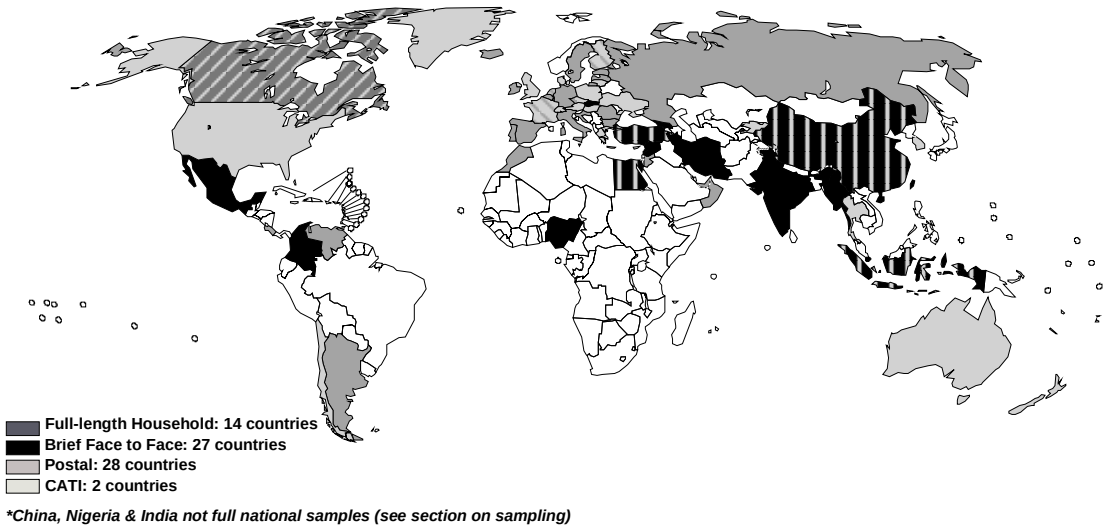


Table 2: Distribution of survey countries by WHO Regions

Region	Total Countries	Number of survey countries	% Region Coverage (by country)
AFRO	46	1	2.2
AMRO /PAHO	35	11	31.4
EMRO	22	8	36.4
EURO	51	33	64.7
SEARO	10	3	30.0
WPRO	27	5	18.5
TOTAL:	**191**	**61**	**31.9**

Full household surveys are currently in progress in Iran, Lebanon, Singapore and Syria.

9.1 Participating Countries

Table 3 lists the countries that have participated in the survey study.

Table 3: Countries Participating in the Survey Study

Household	Brief Face-to-Face	CATI	Postal
China†	Argentina	Canada	Australia
Colombia	Bahrain	Luxembourg	Austria
Egypt	Belgium		Canada
Georgia	Bulgaria		Chile
India†	Costa Rica		China
Indonesia	Croatia		Cyprus
Iran*	Czech Republic		Czech Republic
Lebanon*	Estonia		Denmark
Mexico	Finland		Egypt
Nigeria†	France		Finland
Singapore*	Germany		France
Slovakia	Iceland		Greece
Syria*	Ireland		Hungary
Turkey	Italy		Indonesia
	Jordan		Kyrgyzstan
	Latvia		Lebanon
	Malta		Lithuania
	Morocco		Netherlands
	Netherlands		New Zealand
	Oman		Poland
	Portugal		Rep. of Korea
	Romania		Switzerland
	Russian Federation		Thailand
	Spain		Trinidad and Tobago
	Sweden		Turkey
	United Arab Emirates		Ukraine
	Venezuela		United Kingdom
			USA
14 (10 completed)	**27**	**2**	**28**

** Surveys in progress - four household surveys in marked countries are due December 2001.*

† *Not full national sample*

9.2 Results by Survey Modes

9.2.1 Household Mode

In the household mode, trained interviewers carried out a face-to-face detailed interview with one randomly selected respondent from each household in the sample. Calibration tests were also carried out at the respondents' houses. Brief descriptions of survey samples each country are given in the Appendix 2 (*refer to pages 71-139*). The sampling was done as a general multistage cluster

sampling. The quality of the sample to represent the general population was monitored through the Sample Population Deviation Index, which shows the proportion of age and sex strata in the sample in comparison to the general population as taken from UN population statistics (refer to *Figure 5, page 37* for further explanation). In general the duration of the household surveys ranged between 65 - 119 minutes with an average of 87 minutes (see also *Section 9.4* Survey Length for detailed information on duration).

The average household response rate across all countries was 84%. The missing data rate overall was less than 5%, where missing data is defined as proportion of respondents missing greater than 10% of items in Household, 2% of items in BFTF and 5% of items in Postal modes. Test-retest reliability calculated by chance corrected agreement rates was in general higher than 0.6 which indicated overall good reliability. Design effects (i.e. the ratio of sampling variance in comparison to simple random sample) ranged between 3-8 in selected key variables. While most surveys make sample size estimates based on design effect estimates of around 2, the results of this study suggest that design effect for many items may indeed be much higher; the who Multi-country Survey Study had calculated required sample sizes based on a conservative design effect estimate of 2.5. All these summary diagnostic variables for different modes of surveys are shown in Table 4. Detailed findings for each country are given in the Tables 10A through C.

Table 4: Summary Survey Diagnostics for Different Modes

SURVEY MODE	Duration in min.	Sample Deviation Index	Response rate %	Respondent Missing Data*	Reliability	Design Effects
Household	**87** [65-119]	**2.06** [0.67-3.84]	**84%** [82 – 99]	**12.1 %** [0.8-31.9]	**0.67** [0.43-0.87]	**5.85** [2.0-10.7]
BFTF	**35** [15-40]	**2.28** [0.89-6.57]	**64%** [35-80]	**1.5%** [0-17.6]	*Not done*	*Not applicable*
Postal	**45** (estimated)	**2.91** [0.86-7.95]	**46%** [24- 92]	**6.8%** [0 - 26]	*Not done*	*Not applicable*
CATI	**30** [18-42]	**2.18** [1.50-2.87]	**40%** [25 - 55]	**2.1 %** [1.6-2.6]	*Not done*	*Not done*

** Respondents with missing data 10% in household, 2% in BFTF and CATI, 5 % in postal surveys.*

9.2.2 Brief Face-to-Face Mode

In the brief face-to-face mode, trained interviewers carried out a face to face brief interview (on average 35 minutes) with one randomly selected respondent from each household in the sample, using either the birthday method (an adult member of the household whose birthday was closest to the date of interview date was requested to complete the questionnaire) or the Kish Method (30). The response rates for the Brief Face to Face interviews ranged between 35 – 80% with a mean value of 64%. In comparison to longer household surveys, this response rate is lower mainly because of the fact that fewer calls were made to households in the case of non-response (three calls versus ten) and possibly because of the selection of countries for this mode (western countries tend to have higher refusal rates). The BFTF mode had a missing data rate of 1.5%, the lowest of all

modes, and a SDI of 2.28, lower than the Postal mode and not significantly different from the Household mode (p=0.61).

9.2.3 Postal Mode

Using a well described sampling frame of addresses, questionnaires (refer to Appendix 2, pages 71-139 for details) were mailed or couriered to respondents. Either the respondent to whom the questionnaire was addressed or the person with the closest birthday answered the questionnaire. The questionnaires were mailed back in most countries or collected in person in China, Egypt and Turkey. Personal collection yielded a very high response rate (over 90%). According to the reports from sites it took, on average, 45 minutes for the respondents to fill in the questionnaires. Postal mode application had higher level of missing data since there was no interviewer assistance to explain questions and individuals on their own took more liberty not to respond. The postal mode had a missing data rate of 6.8% and a SDI of 2.91, both higher than Household and Postal modes.

9.3 Representativeness of Samples

The sampling procedure used aimed to create a representative subgroup of the whole population from which population-based estimates can be generated. A simple indicator of representativeness is the sex ratio in the sample and the general population. Table 5 provides a comparison of the sex ratio between the population in the respective countries as reflected in the UN population data base and that in the samples from the household surveys.

Table 5: Sex Ratio (Male/Female) in the UN Population Database and in the Household Survey Country Samples

Country	Population	Survey (unweighted)	Survey (weighted)
China	1.04	1.14	1.29
Colombia	0.94	0.53	0.48
Egypt	1.01	0.78	0.80
Georgia	0.87	0.73	0.96
Indonesia	0.97	0.83	1.13
India	1.06	0.86	0.83
Mexico	0.94	1.47	1.39
Nigeria	0.96	0.64	0.82
Slovakia	0.92	0.82	0.80
Turkey	1.00	1.72	1.92
Overall	**0.97**	**0.95**	**1.04**

Table 5 reveals that at some sites post-stratification strategies may have to be employed to correct for less than perfect representativeness. For example, in Turkey, Mexico and China the male population was over-represented while in Colombia the female population was over-represented. In most countries women in the age groups between 35-54 were over-represented.

In addition to sex, the age distribution is an important indicator of the representativeness of the sample. We computed a "Sample-Population Deviation Index" by age categories which shows how closely the distribution of the study sample in the different age groups by sex matches that of the general population distribution as available in the UN population database. A perfect match between the study population and the UN database is indicated by 1, whereas values greater than 1 indicate that the population in that age group has been over-sampled in the survey. Similarly values less than 1 indicate that the population in that group is underrepresented in the study sample.

An example of a country-specific sample population deviation index is shown in Figure 4. The sample-population deviation index for all countries participating in the household, BFTF, and postal surveys is provided in Appendix 2 (pages 71-139).

Figure 5: Example of a Sample Population Deviation Index

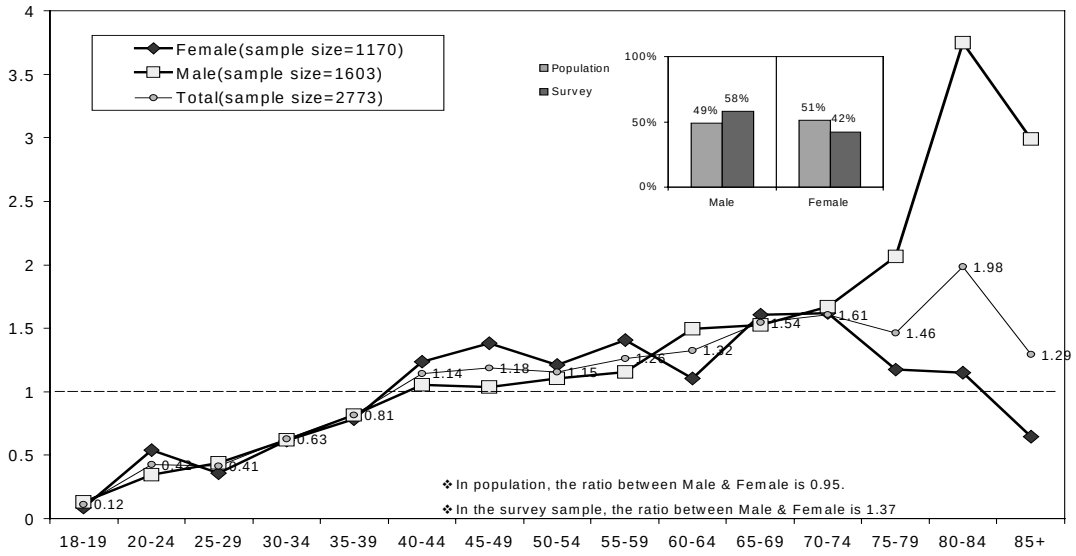


Figure 5 provides a comparison of the sex-age ratio between the WHO sample for the postal survey in a country and the UN population database for 2000. It is evident that between the ages of 20 to 74 the sample surveyed for the WHO postal survey for this particular country closely approximates UN population data. Further, the sex distribution (i.e. ratio of males to females) for the WHO postal survey is comparable to the UN population database.

9.4 Survey Length

For the full-length household surveys, the interview length varied by country, taking between an hour to nearly two hours to complete with an average of 87 minutes. Table 6 provides a more detailed view of the length of different sections of the interview duration for each section during the pilot phase of the study. Table 7 shows the average interview duration for the main phase of the study by country for the full-length household countries.

Table 6: Timing of Individual Sections of full-length household interviews from Pilot Phase (data from 10 countries)

Section	Timing (minutes)	
	Average	Range
Demographics	8	5-12
Health State Description	15	12-23
Health Conditions	4	3-7
Screening	4	3-6
Health State Valuations	24	15-40
Health System Responsiveness	20	15-30
Adult Mortality	4	3-7
Calibration Tests	15	10-19
Total (Minutes)	94	66-144

Given that the pilot phase of the study was intended to develop the survey instrument ,it is understandable that the interviews lasted this long. Since individual modules will be used singly or in combination the duration of the interview can be tailored according to the needs of countries and existing resources.

Table 7: Mean Interview Duration (Minutes) of full-length household interviews – Main phase (data from 10 countries)

Country	Mean Duration (minutes)
China	97
Colombia	68
Egypt	71
Georgia	97
Indonesia	119
India	87
Mexico	90
Nigeria	79
Slovakia	102
Turkey	65
Average	**87**

9.5 Reliability of the Survey Instrument

Reliability of the items in the survey instruments is measured using chance-corrected agreement rates. Kappa (κ) statistics were used for categorical variables and Intraclass correlation coefficients (ICC) for continuous variables. These measures assess reliability as a relative measure of agreement between observers compared to the agreement that can occur by mere chance. A weighted Cohen's kappa analysis (35) was employed to determine the test-retest reliability of the items. The differences between the multiple categorical responses at time 1 and time 2 were weighted such that,

a greater discrepancy on the categories across the two applications (e.g. between 2 and 4 compared to between 3 and 2) would receive a greater weight.

Reliability coefficients can range from +1 to - 1. A kappa coefficient of +1 means that there is complete agreement, while a coefficient close to 0 is consistent with the hypothesis that agreement occurs by chance. Any kappa coefficient less than 0 suggests that any agreement that occurs is less than that due to chance alone. Kappa values less than 0.0 indicate poor reliability; 0.0 to 0.2 slight reliability; 0.21 to 0.4 fair reliability; 0.41 to 0.6 moderate reliability; 0.61 to 0.80 substantial agreement; and 0.81 to 1.00 almost perfect agreement (36). Table 8a summarizes the reliability statistics for major sections of the health component of the survey study. Table 8b summarizes the reliability statistics for major sections of the responsiveness component of the survey study. The retest sample size was 5684 (range 98-1040) individuals in 10 countries for the health domains and 4625 (range 96-940) for the responsiveness domains. All retest subjects were readministered the original interview in its entirety.

Table 8a: Reliability Statistics for Major Sections of the Health Component of the Survey (data from 10 countries in full-length Household Interview mode)

	Overall Average Kappa Value Across All countries	Range
Health Items		
▪ Mobility	.71	.37-.92
▪ Self care	.60	.32-.95
▪ Work and household activities	.69	.24-.93
▪ Pain or discomfort	.67	.36-.94
▪ Distress, sadness, worry	.65	.31-.93
▪ Concentrating and remembering	.62	.16-.95
▪ Personal relationship; community participation	.58	.20-.93
Vignettes		
▪ Self-care	.53	.38-.54
▪ Pain	.56	.47-.56
▪ Mobility	.49	.45-.56
▪ Affect	.51	.41-.52
▪ Usual Activities	.55	.52-.63
▪ Cognition	.55	.49-.63
▪ Vision	.58	.46-.62
Performance Tests		
▪ Verbal fluency	.83	.35-.95
▪ Immediate recall	.57	.30-.86
▪ Delayed recall	.63	.26-.86
▪ Cancellation	.66	.30-1.00
▪ PLM	.39	.30-.98
▪ Vision	.80	.34-.92

Table 8b: Reliability Statistics for Major Sections of the Responsiveness Component of the Survey

	Average Kappa	Range	
		min	max
Utilization	0.80	0.78	0.84
Discrimination	0.50	0.00	0.80
Responsiveness Items-Outpatient			
Dignity	0.65	0.63	0.69
Confidentiality	0.75	0.74	0.76
Quality of Basic Amenities	0.73	0.73	0.74
Choice	0.75	0.71	0.77
Autonomy	0.74	0.70	0.78
Prompt Attention	0.77	0.63	0.88
Responsiveness Items(Total)-Inpatient	0.73	0.63	0.81
Vignettes			
Communication	0.57	0.48	0.58
Dignity	0.53	0.43	0.60
Confidentiality	0.55	0.54	0.66
Quality of Basic Amenities	0.48	0.45	0.61
Choice	0.54	0.50	0.57
Social Support	0.56	0.48	0.57
Autonomy	0.60	0.48	0.62
Prompt Attention	0.59	0.53	0.65
Importance	0.54	0.49	0.60

Most values reported in Tables 8a and b are within acceptable ranges. Some countries such as Colombia and Nigeria have lower reliabilities across a large number of variables. Tables 10A to C present the reliability coefficients for each country.

9.6 Implementation of Measured Calibration Tests

The calibration tests were tested in pilot studies. In order to ensure their uniform application, all principal investigators from household study sites were trained at Geneva in a workshop in July 2000 in their implementation. A training video was also prepared demonstrating the implementation of the tests and distributed to all sites. Further, during each of the site visits made by HQ staff the actual implementation of the calibration tests was observed and suggestions provided for a standardized application. In spite of these methods, it was observed that the calibration tests were not being consistently applied at some sites because the lay trained interviewers were not accustomed to carrying out such tests in the field.

The measurement properties of the calibration tests are reported in table 9. They are within acceptable ranges and are comparable to kappas for self-report described above. Completion rates were over 95% for all calibration tests.

Table 9: Summary Results of Calibration Tests

Calibration Test	Description	Measurement results	Reliability (Kappa/ICC)
Vision	*Near Vision*	20% of sample better than 6/6 10% of sample worse than 6/60	0.80 [0.34-0.92]
	Distant Vision	70% of sample better than 6/6	
PLM	*Mean Time to stand-up*	2.4 seconds	0.39 [0.30-0.98]
	Mean Time to Complete test	13.4 seconds	
Verbal Fluency	*Mean # of animals correctly named*	16.7	0.83 [0.35-0.95]
	Errors	0.7	
Verbal Recall	*Immediate Recall – mean # of words*	6.4	0.57 [0.30-0.86]
	Delayed Recall – mean # of words	5.5	0.63 [0.26-0.86]
Cancellation Test	*Mean Time to Complete test*	33 seconds	0.66 [0.30-1.00]
	Mean # of errors in cancellation	2.3	

Some practical problems will have to be dealt with during the refinement of their implementation in the future. Almost all the calibration tests (except for vision) are timed tasks. The timing of these tests need to be recorded accurately using stop watches, and a failure to use them in all settings may have led to some inaccuracies. The implementation of the Posturo-locomotion (PLM) test requires a distance of 6 metres as well as a standard weight of 2 kgs. In most households interviewers were able to find an object of that weight but variation may have occurred in the non-standard shape and size of the objects. This is also true for the seating from which the test is begun. For vision tests in some households it was difficult to find a space that is 6 metres long and this instruction may therefore not have been followed accurately. More rigorous training will probably be required to standardize the methods in future studies.

9.7 Vignette Implementation

Though care was taken to ensure that the vignettes were written in a culturally sensitive manner and were reviewed by all study sites, the very nature of the task is unfamiliar in some study settings. We believe that some of the respondents in some settings might not have wanted to think of such

situations lest they became ill themselves. Others responded by saying that they could not say how much difficulty the person described in the vignette would have because this depends on many other factors such as the person's life circumstances. However, overall the vignettes show good repeatability (shown in tables 8a and 8b) suggesting that respondents understand the task and are able to provide meaningful responses.

9.8 Overall Survey Metrics

To assess the quality of the survey process, the survey metrics of the WHO Multi-country Survey Study were examined in a systematic fashion using four components: (1) an aggregate or summary sample population deviation index[6] and the deviation from the UN population database as expressed in terms of chi-square values (2) response rate (i.e. those who completed the interviews among eligibles), (3) item missing values (i.e. percentage of items with > 5% missing data) as well as the percentage of respondents with missing data (>10% for household surveys, >2% for the BFTF surveys and >5% for the postal surveys) and (4) reliability coefficients (i.e. percentage of items which have higher than 0.4 kappa values or equivalent chance-corrected concordance coefficients). Tables 10 A-C provide a detailed description of these survey metrics.

The tables reveal that the survey samples were fairly representative of the national population structure as estimated by the summary sample population deviation indices. The household samples were closer to the UN population database figures than the postal survey samples. In Slovakia the population was the closest to that in the UN database. Whether this was a function of the smaller sample size in Slovakia is unclear.

The response rates for the full-length household survey were much higher than in the other two modes (with the exception of the Egypt and Turkey couriered postal surveys as mentioned earlier). This was perhaps a function of the more rigorous interviewer training and call back attempts also described earlier.

The missing rates at the respondent and item levels are higher in the full-length survey than for the brief survey. Interestingly, the item level missing data rate is the highest in the postal survey although the respondent level missing data is highest in the full-length survey. A possible interpretation is that the length of the full version may have been more than optimal. However, the presence of an interviewer (unlike the way a postal survey is conducted) leads to more complete data collection on all items.

---

[6] Aggregate Deviation Index or summary index is calculated using the following formula:

$$\text{Summary Index} = \Sigma \mid 1 - \text{Age Group Index} \mid$$

The summary index was calculated for all age groups between 20 and 74, as they represent the most stable age groups across survey countries and modes.

Table 10A: Survey Metrics for the Household Survey –Health & Responsiveness Components

						Missing Data				Reliability (kappa)					
						Health Module		Responsiveness Module		Health Module		Responsiveness Module			
												Outpatient		Inpatient	
Country	Summary SPDI (unweighted)	Summary SPDI (weighted)	Deviation from UN Population Chi-square critical value = 49.58 (29 df, p=0.01)	Chi-square p-value	Response Rate %	Individuals with > 10% of Missing Data	Items with > 5 % Missing data	Individuals with > 10% of Missing Data	Items with > 5 % Missing data	Overall Items	% of Items > 0.40	Overall Items	% of Items >0.40	Overall Items	% of Items >0.40
China	1.95	3.62	425.2	0.0 E -6	99.0	10	4.6	21.1	7.3	0.75	99.0	0.8	100.0	0.8	100.0
Colombia	1.70	1.38	201.8	0.0 E -6	82.0	31.9	1.7	52.0	5.2	0.43	49.0	0.4	12.0	0.5	75.0
Egypt	1.77	1.39	283.9	0.0 E -6	99.0	12.1	6.1	59.2	1.0	0.75	97.0	0.8	100.0	0.9	100.0
Georgia	1.00	0.91	214.3	0.0 E -6	87.0	5.4	6.0	24.7	4.2	0.53	69.0	0.4	52.0	0.6	92.0
India	2.29	2.84	309.9	0.0 E -6	98.0	0.8	4.7	31.3	4.2	0.66	91.0	1.0	100.0	1.0	100.0
Indonesia	1.69	1.69	512.7	0.0 E -6	99.0	6.4	7.3	23.7	3.1	0.81	98.0	0.7	100.0	0.7	100.0
Mexico	3.30	3.47	408.6	0.0 E -6	96.0	3.9	10.5	-	4.2	0.83	98.0	N/A	N/A	N/A	N/A
Nigeria	2.36	2.80	411.6	0.0 E -6	98.0	11.5	5.6	21.2	14.6	0.37	45.0	0.3	36.0	0.1	8.0
Slovakia	0.67	1.08	28.8	4.8 E -1	84.0	16.5	6.0	43.9	7.3	0.87	97.0	0.9	100.0	0.9	83.0
Turkey	3.84	3.96	171.5	0.0 E -6	90.0	22.6	8.0	65.5	8.3	0.66	93.0	0.8	100.0	0.5	8.0
Overall	**2.06**	**2.31**	**296.83**	**0.0 E -6**	**93.2**	**12.12**	**6.1**	**34.2**	**5.9**	**0.67**	**83.6**	**0.6**	**70.0**	**0.6**	**66.6**

Table 10B: Survey Metrics for the Brief Face-to-Face Survey –Health & Responsiveness Components

Country	Summary SPDI	Deviation from UN Population Chi-square critical alue=49.56 (26 df, p=0.01)	Chi-square p-values	Response Rate %	Health Module		Responsiveness Module	
					Individuals with > 5 % of Missing data	Items with > 5 % Missing data	Individuals with >6 % of missing data	Items with >5 % of Missing data
Argentina	1.31	62.0	3.5 E -4	36	3.9	2.4	0.0	9.1
Bahrain	4.23	408.8	0.0 E -6	35	0.2	3.5	0.2	3.0
Belgium	1.36	83.6	0.0 E -6	48	5.5	1.7	8.3	15.9
Bulgaria	0.91	46.7	7.6 E -3	88	1.7	3.5	14.2	27.0
Costa Rica	1.33	36.4	8.5 E -2	37	2.1	0.8	0.0	1.5
Croatia	0.83	94.3	0.0 E -6	68	0.5	4.1	5.6	13.6
Czech Republic	1.39	54.2	9.6 E -4	60	0.4	1.0	16.5	36.0
Estonia	2.42	120.2	0.0 E -6	71	0.0	2.0	17.2	33.3
Finland	1.86	70.1	0.0 E -6	52	0.2	2.1	9.7	14.4
France	1.59	104.2	0.0 E -6	77	1.0	1.4	5.1	10.6
Germany	1.28	63.5	5.6 E -5	80	3.7	2.0	6.1	16.3
Iceland	1.44	77.1	0.0 E -6	53	0.8	2.0	20.0	33.0
Ireland	1.99	69.8	0.0 E -6	39	0.0	1.8	6.5	21.2
Italy	1.22	118.0	0.0 E -6	61	0.2	1.6	7.4	19.3
Jordan	3.19	149.4	0.0 E -6	74	0.0	2.7	0.2	4.5
Latvia	2.18	190.9	0.0 E -6	53	0.0	2.2	87.6	4.5
Malta	2.19	64.1	4.6 E –5	63	0.2	1.3	5.0	12.1
Morocco	1.57	72.2	3.0 E -6	69	0.0	1.4	0.1	3.0
Netherlands	0.86	115.0	0.0 E -6	59	0.4	2.1	12.1	17.4
Oman	6.00	437.9	0.0 E -6	71	0.0	3.2	0.2	3.0
Portugal	1.53	62.5	4.6 E -5	61	0.0	1.6	9.2	20.1
Romania	2.58	99.5	0.0 E -6	52	1.0	2.9	10.8	40.5
Russian Federation	0.89	166.8	0.0 E -6	25	0.1	1.6	14.5	22.4
Spain	6.57	59.4	2.0 E -4	75	1.6	1.4	6.8	15.5
Sweden	1.45	59.6	1.9 E -4	53	0.1	1.3	13.5	21.2
United Arab Emirates	4.90	469.5	0.0 E -6	72	17.6	3.4	0.9	3.0
Venezuela	4.49	232.3	0.0 E -6	66	0.4	6.8	0.1	1.5
Overall	**2.28**	**132.9**	**0.0 E -6**	**59**	**1.5**	**2.3**	**10.3**	**15.66**

Table 10C: Survey Metrics for the Postal Survey – Health and Responsiveness Component

Country	Summary SPDI	Deviation from UN Population Chi-square critical value=49.56 (26 df, p=0.01)	Chi-square p-values	Response Rate %	Health Module: Individuals with > 5 % Missing data	Health Module: Items with > 5 % Missing data	Responsiveness Module: Idividuals with > 10 % Missing data	Responsiveness Module: Items with > 5 % Missing data
Australia	3.00	746.0	0.0 E -6	35	0	13	N/A	N/A
Austria	3.93	310.0	0.0 E -6	56	7.3	15.1	0.0	31.3
Canada	3.68	158.4	0.0 E -6	55	2.4	4.8	0.0	20.2
Chile	4.80	410.7	0.0 E -6	42	4.2	8.4	0.0	29.8
China	7.86	1969.3	0.0 E -6	50	0.4	4.1	0.3	12.6
Cyprus	2.95	331.8	0.0 E -6	27	3.2	2.2	0.0	17.7
Czech Republic	3.24	359.8	0.0 E -6	40	4	1.7	0.1	42.2
Denmark	0.98	91.9	0.0 E -6	54	5.3	9.7	0.1	42.7
Egypt	1.00	85.1	0.0 E -6	92	1.9	1.2	0.1	19.7
Finland	1.49	300.8	0.0 E -6	54	10.1	8.1	0.3	23.2
France	2.13	239.3	0.0 E -6	31	6.3	6.5	0.2	25.3
Greece	3.19	359.6	0.0 E -6	35	3.3	2.6	1.2	38.4
Hungary	0.86	171.6	0.0 E -6	72	4.1	15.6	0.0	29.3
Indonesia	3.40	986.5	0.0 E -6	60	6.1	5.8	0.2	13.6
Kyrgyzstan	2.94	273.1	0.0 E -6	44	13.2	9	0.6	32.8
Lebanon	3.19	66.4	2.2 E -5	44	1.4	5.1	N/A	N/A
Lithuania	1.13	120.9	0.0 E -6	70	4.6	6.5	1.4	10.6
Netherlands	0.96	275.2	0.0 E -6	40	5	6.1	0.5	17.7
New Zealand	2.81	474.2	0.0 E -6	68	5.5	1.7	0.2	22.5
Poland	1.52	370.9	0.0 E -6	34	4.7	6.2	0.1	26.8
Rep. of Korea	7.95	506.4	0.0 E -6	24	0	1.2	0.4	18.7
Switzerland	1.85	63.4	5.8 E -5	38	9.7	16.9	0.3	41.0
Thailand	1.89	179.4	0.0 E -6	46	0	1	0.5	26.5
Trinidad & Tobago	2.48	231.7	0.0 E -6	52	22.3	7.9	0.3	53.3
Turkey	3.57	511.4	0.0 E -6	90	23.9	10.1	0.1	70.5
Ukraine	1.02	129.9	0.0 E -6	31	4.5	15	0.4	54.3
United Kingdom	2.26	182.7	0.0 E -6	40	26.7	16.5	0.0	30.3
USA	5.28	857.5	0.0 E -6	35	10.9	2.5	0.0	48.5
Overall	**2.91**	**384.43**	**0.0 E -6**	**48.54**	**6.82**	**7.31**	**0.3**	**30.8**

Figure 6: Sample-Population Representativeness by Survey Mode

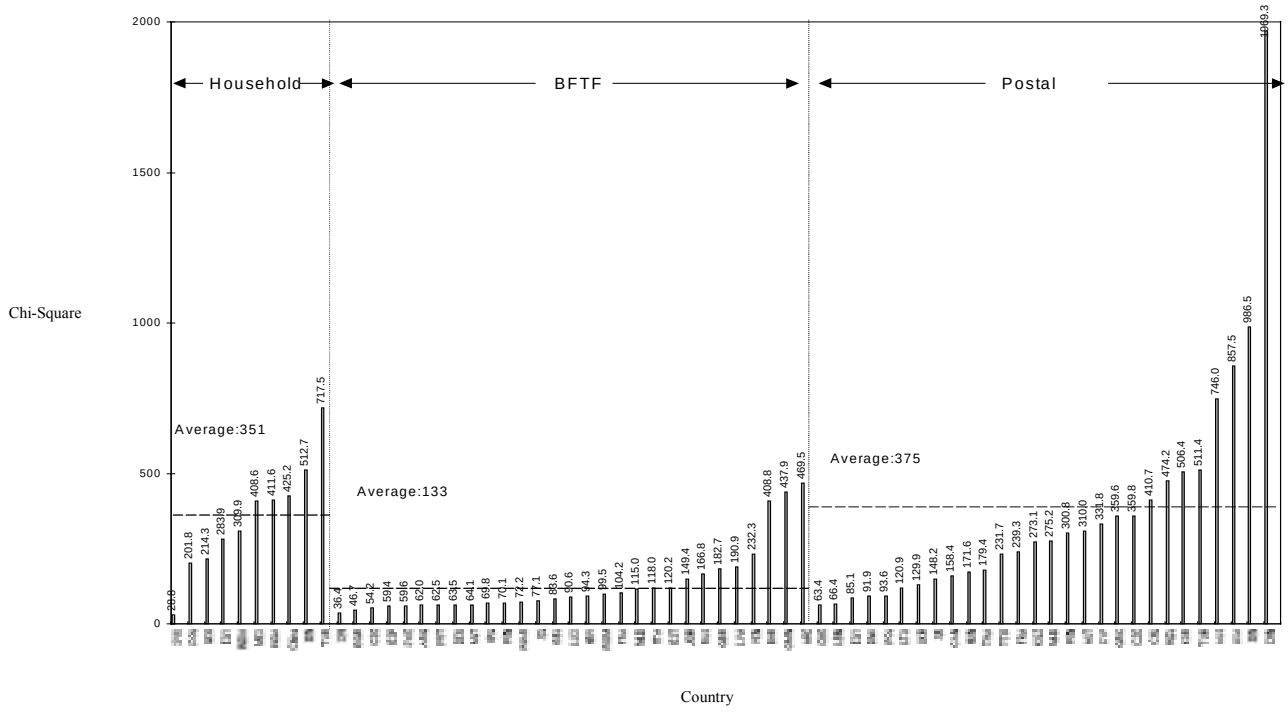


Figure 6 illustrates the country and mode-specific chi-square values for the deviation of the WHO sample (observed) from the UN population database for the year 2000 (expected), collapsed across all age groups between 20 and 74 years. Within each survey mode, the countries are rank ordered according to their aggregate chi-square. For each survey mode the average chi-square across all countries is also indicated. It is evident that there is variation in the aggregate chi-square between countries within each survey mode, the greatest variation being between countries in the postal survey mode. In the countries with the greatest deviation from the expected sample age and sex distribution, attempts are being made to understand what happened and how this can be overcome.

9.9 Comparison of Survey Mode Effects

One of the objectives of the WHO Multi-country Survey Study was to compare the feasibility and efficiency of different survey modes, so it is of scientific interest to consider the survey metrics in countries where multiple modes were used.

Figure 7a provides a comparison of the chi-squares values of the deviation of the samples from the UN population database for 2000 for countries that participated in both the postal and household surveys. This figure illustrates that for Egypt and Turkey the WHO sample deviation from the UN population database for the postal surveys was considerably less than that for the household surveys, perhaps because the postal surveys were hand delivered in these countries using a sample frame of postal addresses. In contrast, for Indonesia and China the postal surveys had a much higher deviation from the UN population database than the household surveys. This is perhaps because the postal surveys were actually mailed and the sampling frame may not have been adequate. Also the response rates for the mailed surveys in China and Indonesia were much less than the household surveys (50-60% compared to 99%).

Figure 7a: Comparison of Household vs. Postal Surveys for Representativeness

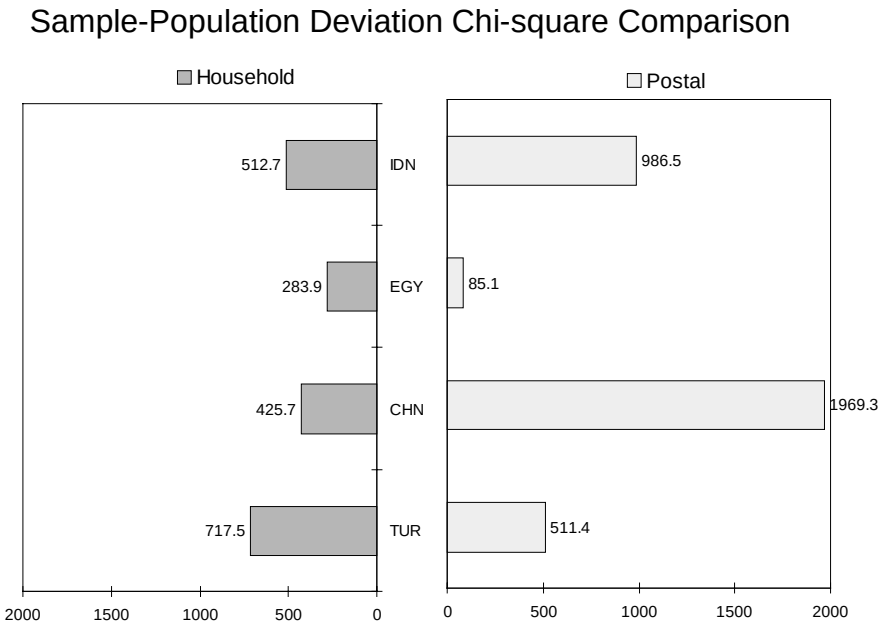


Figure 7b compares the chi-square for countries that participated in both the postal and brief face-to-face (BFTF) surveys. In all countries, but particularly in the Czech Republic and Finland, the sample deviation from the UN population database was considerably higher for the postal surveys than for the brief face to face (BFTF) surveys. This is perhaps due to the lack of an adequate sampling frame for the postal surveys.

Figure 7b: Comparison of BFTF vs. Postal Surveys for Representativeness

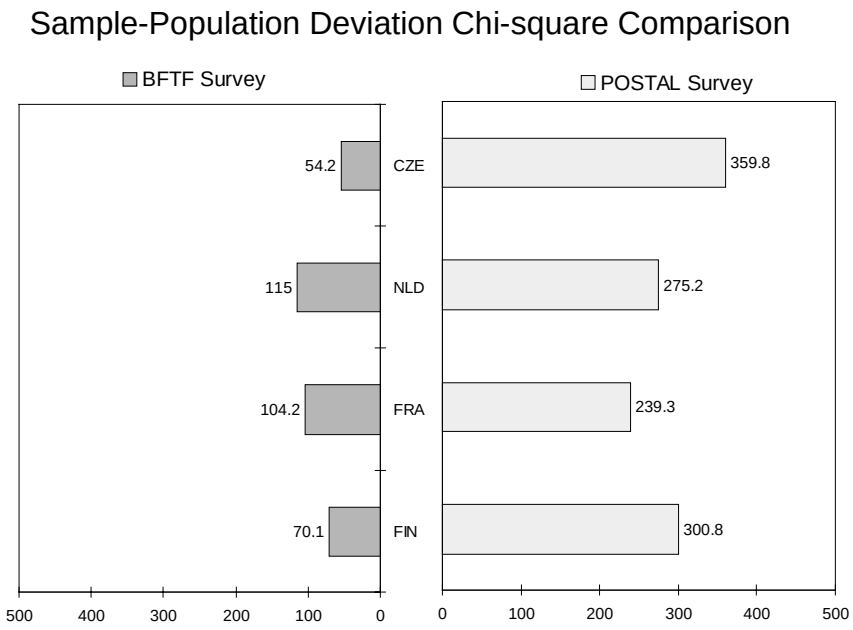


Figure 8a: Summary Quality Assessment of Household Surveys by Reliability and Missing Value

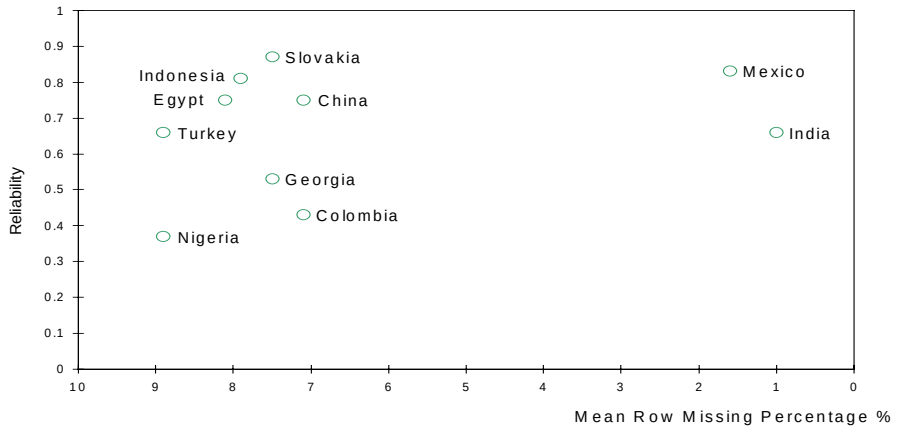


Figure 8b: Summary Quality Assessment of Household Surveys by Reliability and Representativeness

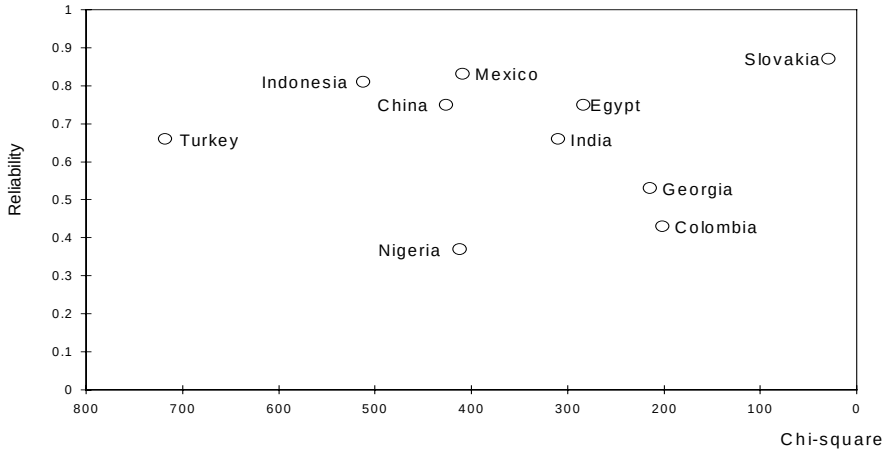


Figure 8c: Summary Quality Assessment of Household Surveys by Representativeness and Missing Value

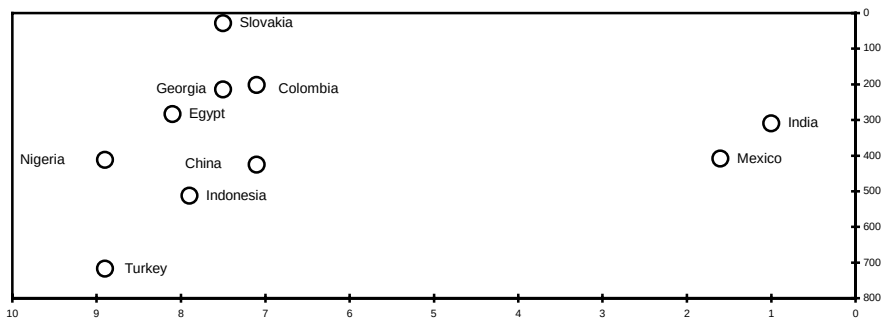


Figure 8d: Summary Quality Assessment of WHO Surveys by Representativeness and Missing Value

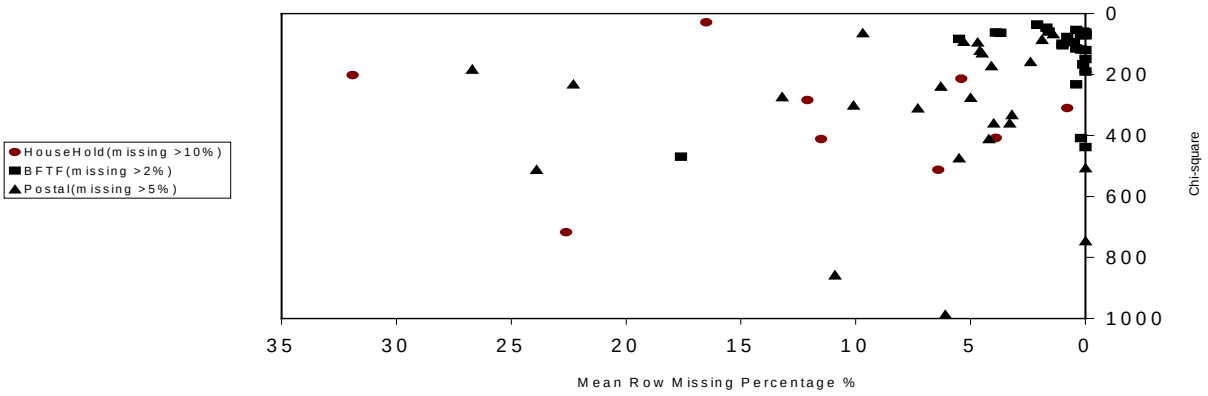


Figure 8e: Summary Quality Assessment of WHO Surveys by Response Rate and Missing Value

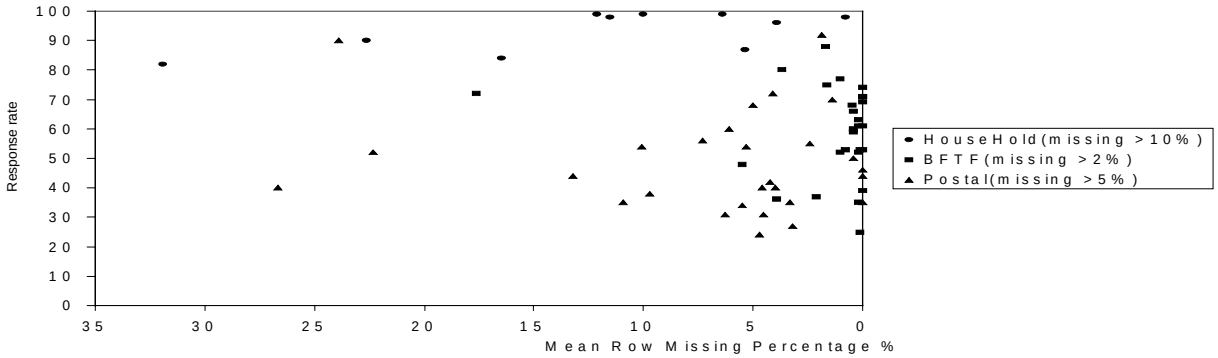
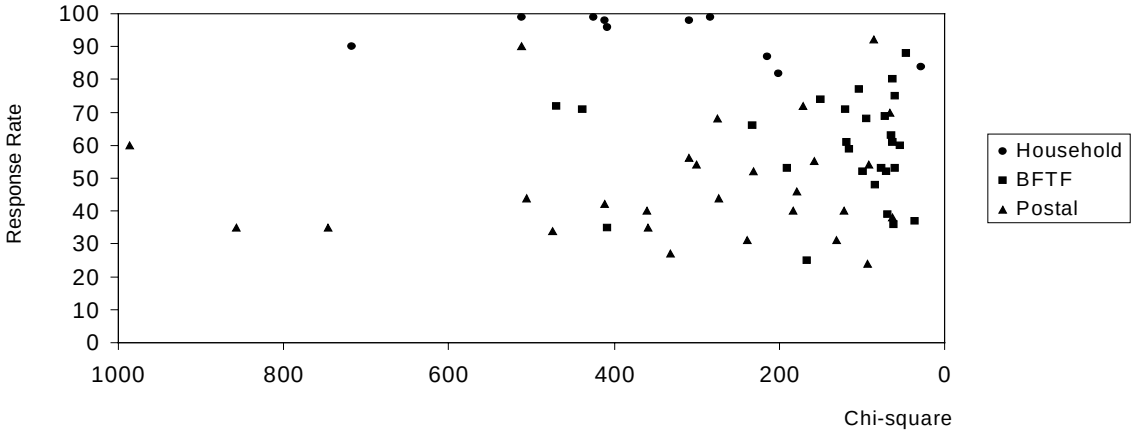


Figure 8f: Summary Quality Assessment of WHO Surveys by Representativeness and Response Rate



Figures 8a-f provide an overview of the performance of survey metrics in countries participating in the household surveys, in terms of their population deviation (chi-square), their test-retest reliability as measured by the kappa coefficient, the missing values and the response rates. The five figures show the relationship between these four quality metrics of the surveys.

From these figures it appears that the different surveys modes performed differently depending on the metric being examined. For example, though the surveys in India and Mexico had very little missing data and high reliabilities the survey samples in these countries were not as representative of the overall country population as for example in Slovakia. Though the survey in Turkey had acceptable reliability the missing data and deviation from the country population were relatively high. The survey in Nigeria had higher missing values and lower reliabilities while the chi-square values were close to the overall average.

Overall, the results suggest that there is no consistent relationship between the different survey metrics for any mode, e.g. reliability is not correlated with missingness or with representativeness. Brief face-to-face interviews have higher response rates and less missing data relative to the other modes, but the trade-off is that they collect less data. We are in the process of developing a single index of quality that can combine all these different parameters using multiple-indicator models, and examining how costs, sample size, country of implementation etc. relate to that index.

9.10 Comparison to other multi-country surveys

The results of the survey study in terms of the quality metrics can be better understood when they are compared with surveys such as the Living Standards Measurement Survey (LSMS) (37) and the Demographic and Health Survey (DHS) (38). For example in the DHS surveys the proportion of missing items that have more than 5% missing data ranges from 0.82% for Colombia in DHS III to 25.24% for Thailand for DHS I. There is a substantial decrease in the percentage of items with more than 5% missing data over the three waves of the DHS. The chi-square values comparing the survey sample (observed) with the UN population database for 2000 (expected) for the DHS surveys (when considering all women in the age group 15-49 years, critical value=26.22, 6df, p=.01) ranges between 6.21 in El Salvador for DHS I to 351.03 in Mali for DHS III. This is shown below in Figures 9a and 9b.

Figure 9a: Representativeness of DHS surveys by country

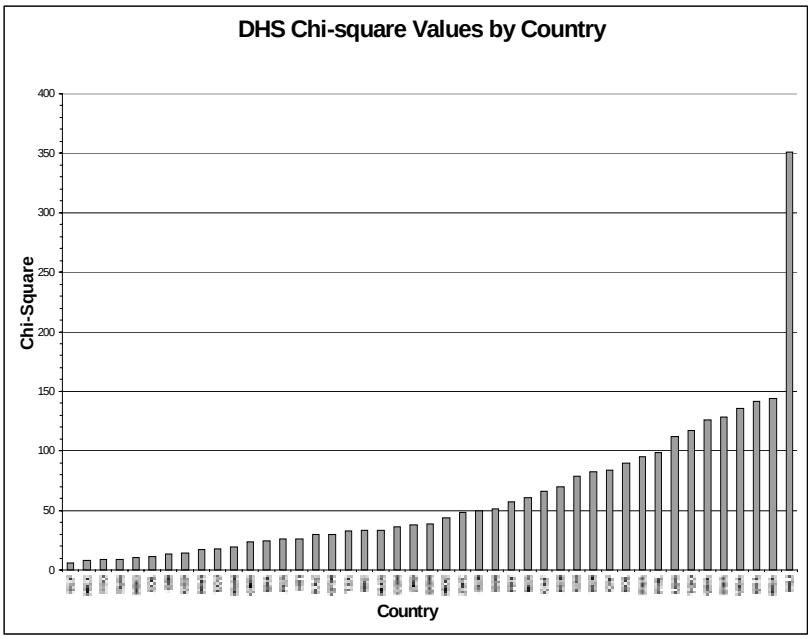
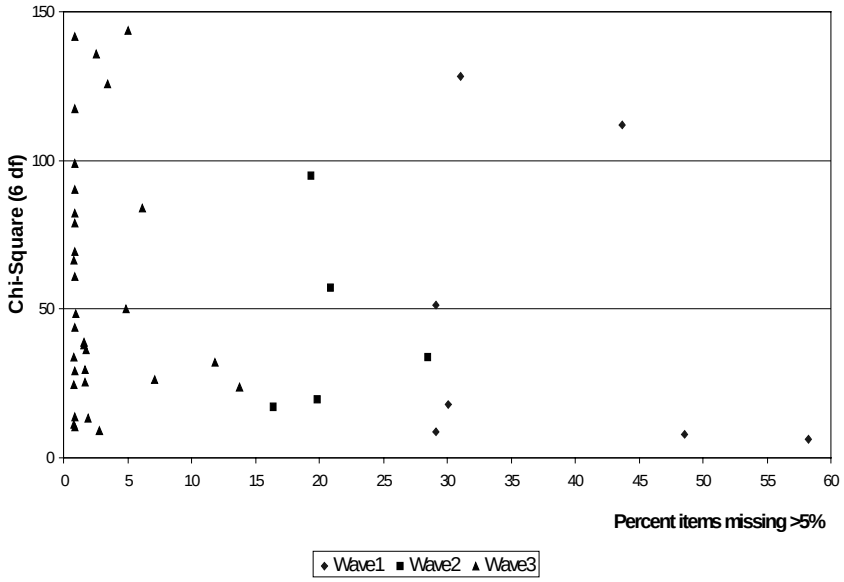


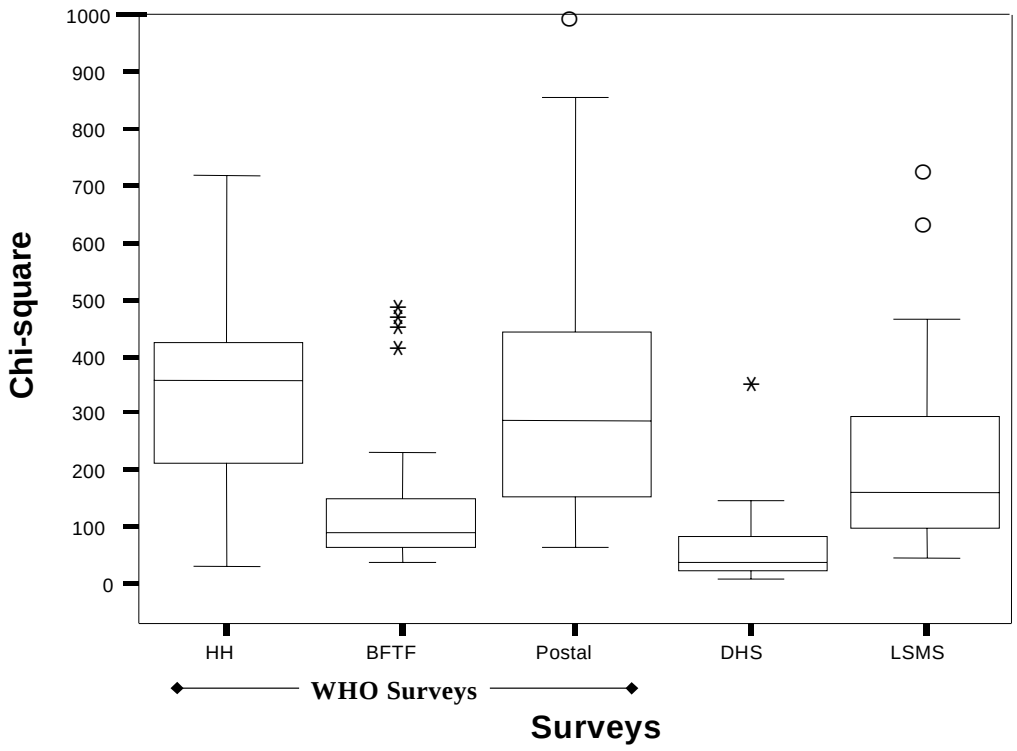
Figure 9b: Relationship between missing data and representativeness for DHS surveys



Similarly for the LSMS and related surveys the chi-squares for the representativeness of the sample ranged between 46.28 for Guyana to 1609.28 for the Lebanon Family Expenditure Survey (critical value=44.31, 25df, p=.01).

Figure 10 shows the comparisons in the chi-square values as measures of representativeness of the samples across the WHO and other surveys. The DHS surveys have a very narrow range of chi-square values as compared to the LSMS or WHO surveys and are the least deviant from the UN population database. Though the BFTF surveys also have a narrow range and smaller chi-square values, all the WHO and LSMS surveys are within overlapping ranges, suggesting that all the surveys were comparable in terms of their deviation from the UN population data.

Figure 10: Comparison of representativeness across different surveys



9.11 Costs

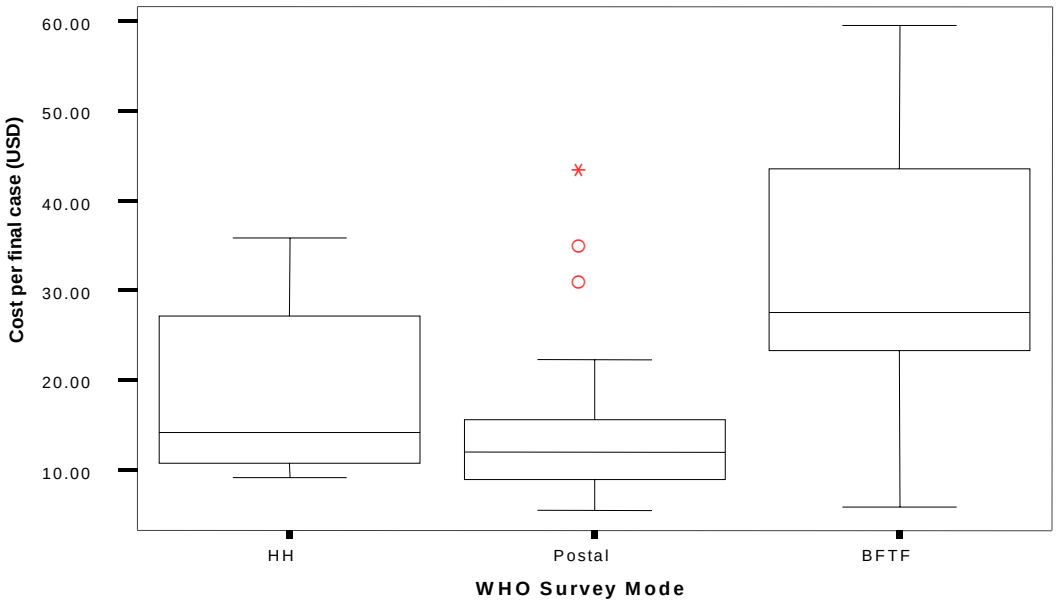
It is difficult to compare the costs of surveys in different contexts, since many factors vary across sites such as, the equivalence of modes, variation in purchasing power, and individual arrangements with the sites. Nonetheless, it is still important to give an idea for the WHO Multi-

country Survey Study since there is a general erroneous belief that surveys are costly. The average in-country interview is shown in the table below.

Table 11: Average Cost Per Completed Interview

Survey Mode	Average Cost Per Final Case [range]
Household	$ 16 [8 - 35]
Brief face-to-face	$ 34 [6 - 59]
CATI (avg. 30 minutes)	$ 30 [27 - 33]
Postal	$ 15 [5 - 43]

Figure 11: Comparison of Cost by Mode



The findings confirm that face to face surveys are generally more expensive than postal surveys. The costs of the BFTF surveys, on average, were higher, perhaps because they were carried out mainly in European countries. Nonetheless the overlapping ranges suggest that costs are comparable across survey modes.

It is useful to compare these costs with other various survey programmes, although detailed costs are not always available. The average costs in the European Social Survey was about $222 for a completed interview. The Multiple Indicator Cluster Survey (MICS) of UNICEF cost about $112,000 on an average per country excluding costs for UNICEF staff, etc., for between 4,000 and 8,000 interviews per country (39). Comparatively surveys in WHO study ranged between $16,000 to $180,000 for sample sizes between 1,200 to 10,000 respondents. This information is

not easily available for the other major surveys. If a conclusion could be drawn from these comparisons, WHO surveys are probably less costly than other survey programmes partly because they were conducted directly through governmental agencies or universities rather than intermediary agencies, and partly because they do not rely on expensive international consultants to maintain them. More importantly, these survey costs are quite reasonable and affordable in view of the information gains and they, therefore, offer a good means of supporting routine health information systems (HIS).

9.12 Selected Results on Health Module

One of the main aims of the results of the WHO Multi-country Survey Study has been developing appropriate analytical methods to ensure cross-population comparability of the survey data. The analytical methods reported in section 2.2 and 8.1.3 served this goal to ensure that the same level of 'true' health can be estimated in different populations from self-reported data in order to make valid comparisons. Figure 12 illustrates the results for difficulties in moving around in a given country. The circles show the mean response by age, and triangles show the adjusted responses. There is only a slight decrease in average mobility with age without any adjustment. Once the way in which respondents use the categorical scale is used by way of HOPIT, the drop-off in mobility with age is more pronounced. This indicates that the methods developed to adjust self-reports for ensuring cross-population comparability seem to adjust self-reports in the expected directions allowing for more appropriate comparisons.

Figure 12: Adjusted and Unadjusted Mobility by Age

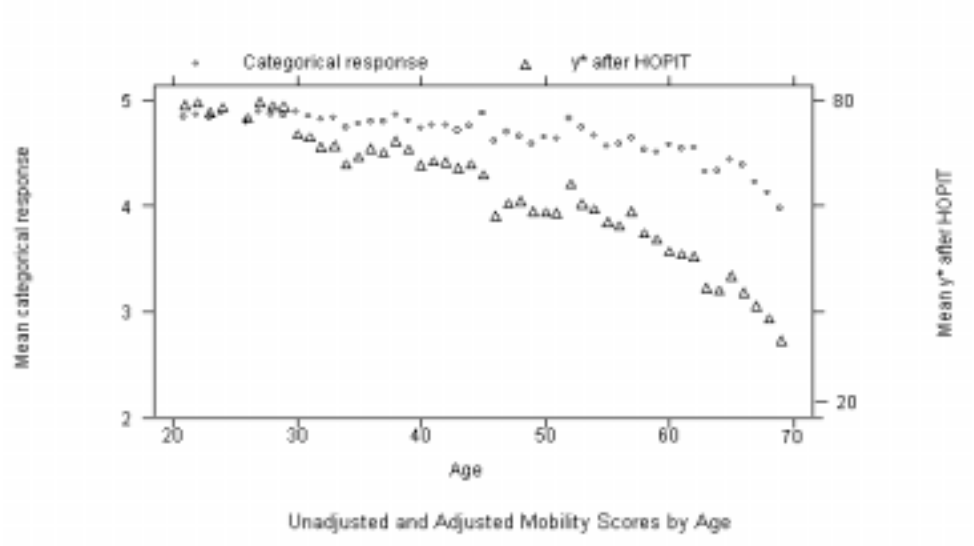
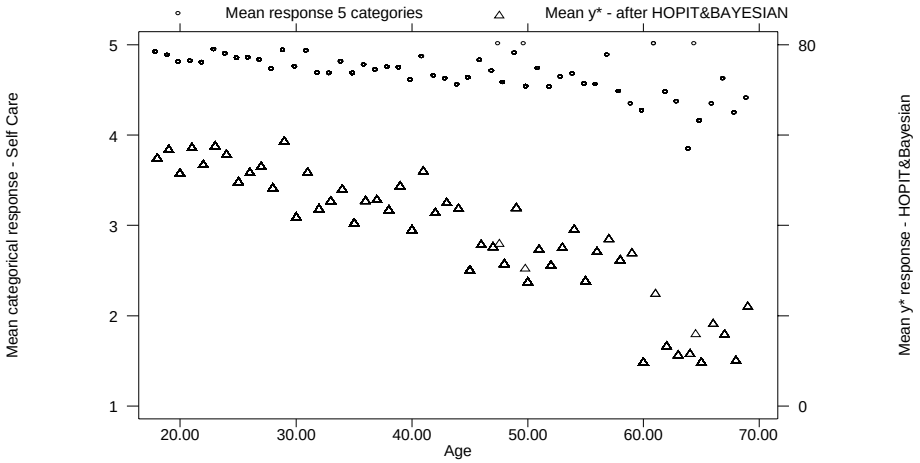


Figure 13 illustrates similar findings with self-care. The adjustment shows that mean responses decline much more rapidly with age than unadjusted self-reports.

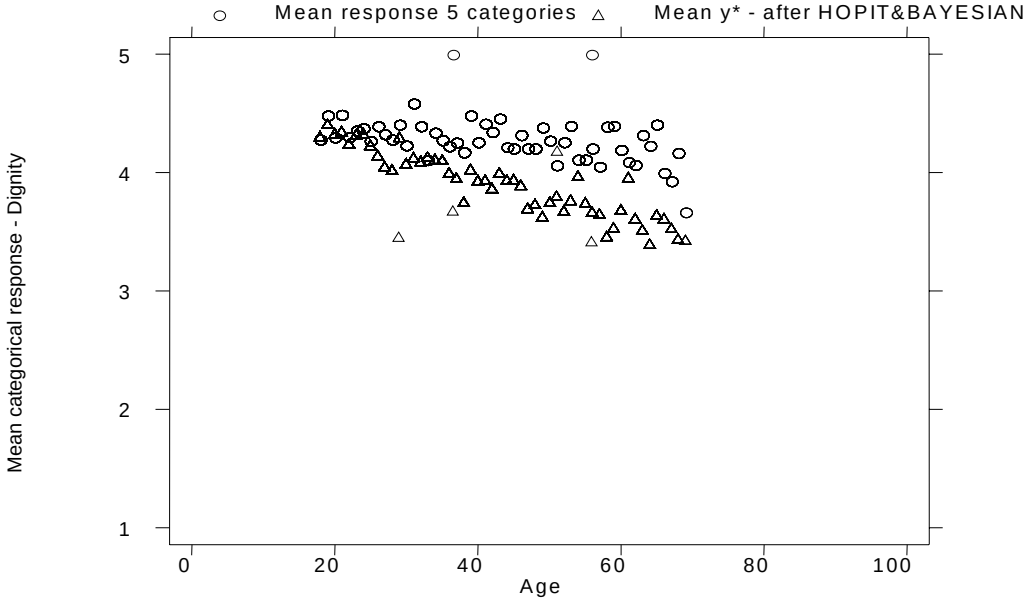
Figure 13: Adjusted and Unadjusted Self-care by Age



9.13 Selected Results on Responsiveness Module

The Figure 14 illustrates an example after adjusting the self-report using the HOPIT developed for the aim of enhancing cross-population comparability using vignettes in the analysis of this survey results. In the data shown in the figure, before adjustment (circles) there are no significant changes with age in the way that respondents describe their experience of being treated with dignity when they come into contact with health services. However, once the results are adjusted (triangles) it is clear that older people in this sample actually experience that they are treated with less dignity by the health care services than the younger persons. Once again the methods identify differences that were not evident before the results were adjusted for the cut-point shifts.

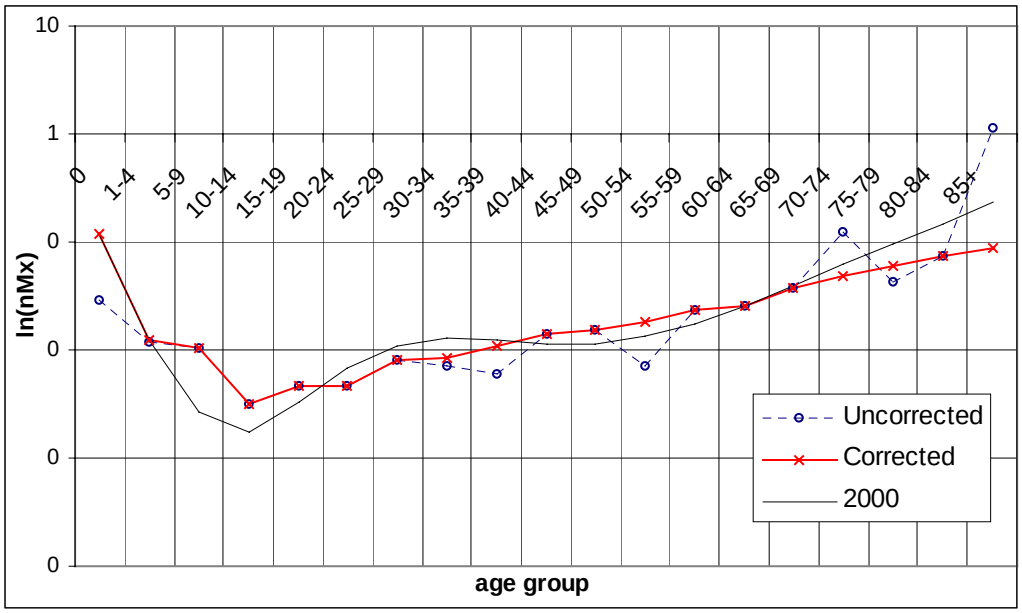
Figure 14: Adjusted and Unadjusted Dignity by age



9.14 Selected Results on Adult Mortality Module

Results from the full version of the household surveys were analysed for the information gathered with regard to adult mortality. Information gathered from the household surveys was compared to the estimates made by WHO from all other data sources such as vital registration, census, UN population data bases, etc. Figure 15 illustrates that data gathered from the surveys on mortality for females in Nigeria, compares favourably with other WHO estimates for the adult female age groups in Nigeria. This is one piece of evidence that suggests that surveys may be an accurate and efficient way to supplement information on mortality in countries where vital registration data is limited or unavailable in order to construct life tables for the population.

Figure 15: Plot of weighted and smoothed log of death rates from survey, compared with corresponding WHO 2000 estimates for females in Nigeria



10. DISCUSSION AND CONCLUSIONS

10.1 Feasibility and Utility

The current phase of the WHO multi-country survey study has demonstrated the feasibility of carrying out large surveys on selected outcomes to which the health system contributes, in a way which supplements the information provided by routine national health information systems. This was necessary because some data, such as that on responsiveness, are not routinely collected. It was also necessary because the data on health that are collected are not standardized, do not include multiple domains, nor has the issue of cross-population comparability been addressed.

The collaborating sites in countries have shown a great deal of interest in being a part of this endeavor. They participated actively in all phases of the survey and completed the data collection in a timely manner. The survey study has used a rather complex design with many different modules and randomization of respondents across different components of the survey. Despite these complexities, completion rates have been high and the rate of missing data has been within acceptable limits.

The implementation of a complex survey study involving 61 countries and 71 surveys within 18 months requires substantial organizational skills on the part of the collaborating partners and continuous monitoring and support from WHO. This has been ensured through periodic phone conversations, email contacts and actual visits to sites for on-site observation. The logistics of the WHO Multi-country Survey Study have been examined and several lessons learnt for future applications.

10.2 Use of Data

The WHO Multi-country Survey Study has proven useful in providing reliable and relevant data on health of the populations, responsiveness of health systems and other additional components such as adult mortality, health care expenditure modules for international comparisons. The annex tables included in the World Health Report 2000 were based largely on secondary analyses of existing various data sources. WHO now has developed an instrument which can be used to collect primary data on key aspects of health system performance.

10.3 Approaches for Cross-population Comparability

In the analysis survey data use of novel techniques, such as the vignettes and the calibration techniques, have proven useful to achieve cross population comparability. Using these data and novel statistical methods (i.e. HOPIT) we were able to overcome the known biases in the self reports data such as those between people from different cultural and educational backgrounds.

The study has demonstrated the feasibility and utility of using vignettes and performance tests to subsequently adjust self-reports. However, further work needs to be done in this area. The current set of vignettes needs to be expanded to capture the breadth of each domain better. In addition, some of the domains of health are inherently multi-dimensional (e.g. vision which may be a combination of near and distant vision, color vision and adaptation to light). This leads to difficulties in rating vignettes according to severity since the different aspects of a domain may be viewed as being independent of each other. Further, the questions being asked of each respondent also need to span all levels of difficulty such that segments of the population that are in good health can be discriminated from those that are in average health and in turn from those that are in poor health.

Performance tests have also been implemented effectively in this study. The feasibility of lay interviewers using them in the general population in survey settings has been demonstrated. However, tests have also shown variation across sites in the manner in which they were conducted and steps need to be taken to ensure they are more uniformly executed. More rigorous training of interviewers with periodic supervision to maintain quality might be the appropriate response.

10.4 Survey Costs

The overall costs of carrying out the study has been considerably less than comparable surveys such as the Living Standards Measurement Survey (LSMS) which costs between 500,000 – 1,000,000 USD per country (38). However survey costs cannot be compared directly given the differences in the nature of the different surveys. Nonetheless the surveys in this study have maintained similar standards of quality. This suggests that surveys of this kind may be an efficient way of gathering data. The implementation of such surveys may often be the best method of collecting specific data in a reliable manner that may not be available from routine health information systems.

10.5 Plans for the World Health Survey (2002-2003)

The demand for internationally comparable information on health and health systems requires a search for low cost ways of supplementing the information routinely gathered by national health information systems (HIS). Based on the experience gained in the WHO Multi-country Survey Study 2000-2001, the World Health Survey (WHS) has been designed specifically to fill these critical information-gaps in routine HIS. The Director–General of WHO has decided to launch the WHS. This will systematically gather data from representative populations to facilitate the task of monitoring health and components of the health system. In addition, it will be an ideal platform to obtain information on health system performance.

The WHO Multi-country Survey Study has provided valuable scientific opportunities for testing different survey modes, developing appropriate questions and techniques for ensuring cross-population comparability. It has also proved to be a low cost way of obtaining data. Some important lessons have been learnt that have helped us to develop the WHS.

10.5.1 *Questionnaires*

The WHO Multi-country Survey Study provided extensive empirical data on the applicability and psychometric properties of individual survey questions. It identified certain questions that have low reliabilities and response rates across several sites (e.g. question on overall health, or overall interpersonal relationships and participation in the community, questions related to alcohol and substance use, vignettes describing risky sexual behavior, etc). Such questions have been reworked to ensure better measurement qualities and uniform applicability across cultures. Questions that are long or complicated were shortened and made more precise.

10.5.2 *Measuring health in the relatively well populations*

It is also important to capture the health level of the relatively well members of the population – in technical terms, avoid ceiling effects in the questionnaire. Given that the respondents were selected randomly from the general population, most were in a relatively well functioning level in any given health domain. Traditional health survey questions are designed to pick up problems and they worked well with people who have moderate and severe problems and do not discriminate between relatively mild problems. Thus, more specific questions have been developed which attempt to capture this range of health spectrum. Given the existing database it will now be possible to develop adaptive questioning strategies allowing us to tailor the questionnaires to the respondents' experience and allow the discrimination of the health states of persons who are at the relatively healthy end of the continuum.

10.5.3 *Modes*

Given the survey quality metrics in the postal surveys, we plan to focus mainly on in-person household or telephone interviews. Although this will increase the costs of surveys slightly, this would be offset by the corresponding gains in quality.

10.5.4 *Sampling*

(a) We used vigorous sampling techniques to ensure that the sample is representative of the general population. However, further work is necessary to ensure that the elderly and the more sick members of the population are not excluded from the surveys. Because these populations may not be in households we need to sample from institutions or oversample the elderly so as to provide information on the health of this segment of the population.

(b) The current phase of the study has focused exclusively on the adult population. Inclusion of youth and children in the World Health Survey would be valuable since data on this age group is particularly scarce. More importantly, in developing countries a large proportion of the population (30-50%) is in this age group.

(c) We also plan to geo-code the data and to use a Geographical Information System (GIS) in order to improve the sampling designs. This would also reduce possible clustering effects.

10.5.5 *Vignettes*

The WHO Multi-country Survey Study has demonstrated the feasibility and utility of using vignettes for cross-population comparability. To be useful, vignettes need to span the whole continuum of a given domain. Some of the sets of vignettes that were piloted did not do this adequately. In addition, some domains of health are inherently multi-dimensional (e.g. the domain of "vision" which may be a combination of near and distant vision, colour vision and adaptation to light). Vignettes need to yield a unidimensional" stimulus to respondents for comparability purposes.

10.5.6 *Performance tests*

To calibrate the range of responses, known tests were used in the survey study (e.g. Snellen eye chart for vision, and some mobility and cognition tests). We have demonstrated that they can be applied to the general population by lay interviewers in survey settings, suggesting that they can be introduced in a larger survey programme. However, these tests have also shown variation across sites in the manner in which they are implemented – (e.g. measured mobility tests showed variability because respondents were asked to walk on different types of surfaces). These tests need to be more uniformly executed in different sites. Also more rigorous training of interviewers specifically on performance tests, with periodic supervision to maintain quality, is required.

10.5.7 *New Survey Modules*

Given the modular design, there has been a demand from countries to develop new modules.

Coverage - Access - Utilization: Ways of ensuring the effective coverage of health interventions has long been of concern to health policy-makers. Despite separate studies on particular aspects of coverage such as immunization and antenatal care, no overall framework which allows regular measurement of what proportion of the population is effectively covered by critical interventions that they need is available. Conceptual development work is under way and following pilot studies such a module will be incorporated in the survey instrument.

Risk factors: Given the importance of risk factors in explaining the current and future health status of individuals and populations it is useful to describe the risk factors influencing the health in a survey module. Such a module would include risk factors in detail such as water and sanitation, air pollution, malnutrition, lack of breast feeding, smoking, alcohol and drugs, physical activity, obesity, unsafe sex, behavioral factors, cholesterol and blood pressure, etc. The current survey questionnaire has several questions on these topics, but a more systematic approach to the measurement ofrisk factors is under development.

Health financing: The Multi-country Study questionnaire included several questions on household income and health expenditures. With the demonstration of the feasibility of this approach, it will be useful to gather more information on health expenditures from all sources in countries that do not collect this information routinely.

10.5.6 *Capacity building and sustainability*

In the future waves it is important to establish the common goals of the survey programme with outside partners taking into account their information needs. Given the need for a comprehensive health system performance assessment at national and sub-national level, it is important to build capacity to carry out periodic surveys in countries and sustain this platform with appropriate resources and skills. Also clear strategies for moving from diagnosis to intervention with regard to health systems need to be identified such that the survey results are used as the evidence to inform policy and for monitoring and evaluation of performance.

The WHS platform will, thus, provide a modular approach that will allow member states to prioritise areas of immediate concern for purpose of data collection in order to inform and monitor decisions. It will be an ongoing programme such that member states can decide the frequency with which the survey needs to carried out in their respective countries. Attempts will be made to incorporate these surveys into national HIS and to harmonise data elements already being gathered by as part of national HIS with those included in the WHS.

The World Health Survey offers an ideal platform to seek information on the prevalence of health states, health state valuations, responsiveness levels and distributions, household health expenditures, risk factors, coverage, basic demography and permanent income which can be useful for health systems performance assessment.

Special efforts to build consensus on collaborative approaches with other agencies sponsoring or conducting surveys will be made e.g. with Demographic and Health Surveys, Living Standards Measurement Surveys, EURO Barometer, and national surveys.

Efforts will be made to build capacity to carry out periodic surveys in countries and sustain this platform with appropriate resources and skills. Clear strategies for moving from diagnosis to intervention with regard to health systems will need to be identified such that the survey results are used as the evidence to inform policy and for the monitoring and evaluation of performance.

11. REFERENCES

1. Eurostat. Self-reported health in the European community: Statistics in focus, population and social conditions. 1997. ISSN 1024-4352.

2. Mathers CD, Sadana R, Salomon JA, Murray CJL and Lopez AD. Estimates of DALE for 191 countries: methods and results: GPE discussion Paper No 16. Geneva: WHO,2000

3. Mathers CD, Murray CJL, Lopez AD, Salomon JA, Sadana R, Ustün TB, Chatterji s. *Estimates of healthy life expectancy for 191 countries in the year 2000: methods and results.* Global Programme on Evidence for Health Policy, Discussion Paper # 38. World Health Organization, 2001.

4. Sadana R, Mathers C, Lopez AD, Murray CJL, and Iberg K. Comparative analysis of more than 50 household surveys on health status, GPE discussion Paper No. 15. Geneva: WHO, 2000.

5. Mathers CD, Sadana R, Salomon JA, Murray CJL and Lopez AD. Healthy life expectancy in 191 countries, 1999. Lancet 2001; 357: 1685-91

6. Murray CJL, Tandon A, Salomon JA and Mathers CD. Enhancing cross-population comparability of survey results, GPE discussion Paper No 35. Geneva : WHO, 2000.

7. Üstün TB, Chatterji S, Rehm J, Kennedy C, Prieto L, Epping-Jordan J, Saxena S, and Pull C in collaboration with WHO/NIH Joint Project Collaborators World Health Organization Disability Assessment Schedule II (WHO DAS II) : Development and Psychometric Testing. In press.

8. Üstün TB, Chatterji S, Bickenbach JE, Trotter RT, Room R, Rehm J and Saxena S, Gottingen, Hogrefe and Huber. Disability and Culture : Universalism and Diversity. 2001.

9. Hambleton, R. K., Swaminathan, H., Rogers, H. J. Fundamentals of Item Response Theory. Newbury, Sage 1991.

10. Streiner DL and Norman GR. Health Measurement Scales: a practical guide to their development and use. 2nd ed. Oxford : Oxford University Press, 1995.

11. Graziano, AM and Raulin, ML. Research methods: A process of inquiry. 1997. New York. Longman.

12. World Health Organization. World Health Report 2000. Geneva: WHO, 2000.

13. Chatterji S, B. Ustun, R. Sadana, J. Salomon, C. Mathers, C.J.L. Murray : The conceptual basis for measuring and reporting on health.

14. World Health Organization . International Classification of Diseases and related health problems – Tenth Revision (ICD - 10). Geneva: WHO, 1992

15. World Health Organization . International Classification of Functioning, Disability and Health (ICF). Geneva: WHO, 20

16. Coren S. Reporting the visual acuity of groups: the relation among alternate measures. Am J Optom Physiol Opt 1987 Dec;64(12):897-900

17. Kokko SM, Paltamaa J, Ahola E and Malkia E. The assessment of functional ability in patients with Parkinson's disease: the PLM-test and three clinical tests. Physiother Res Int 1997;2(2):29-45

18. Spreen O and Strauss E. A compendium of neuropsychological tests. Oxford University Press. Oxford, 1998.

19. Trennery MR, Crosson B, DeBoe J and Leber WR. Visual search and attention test. 1990. Odessa, FL: Psychological Assessment Resources.

20. Wilson B, Cockburn J and Halligan PW. Development of a behavioural test of visuospatial neglect. Arch Phys Med Rehabil 1987; 68: 98-102

21. Tandon A, Murray CJL, Salomon J and King G. *Statistical models for enhancing cross population comparability.* Global Programme on Evidence for Health Policy, Discussion Paper. World Health Organization. Geneva, 2001.

22. Salomon J, Chatterji S, Murray CJL, Üstün TB. WHO Multimethod Valuation Study. (In Press).

23. J. Salomon, C.J.L. Murray, B. Ustun, S. Chatterji. Health state valuations in summary measures of population health.

24. Valentine NB, de Silva A and Murray CJL. Estimating Responsiveness Level and Distribution for 191 Countries: Methods and Results: GPE Discussion Paper Series: No. 22. Geneva: WHO, 2000

25. Ortiz, J. P., Valentine, N, and Liu, B. Comparison of household surveys and key informant surveys as modes to measure responsiveness. (WHO unpublished document, 2001).

26. World Health Organization . Composite International Diagnostic Interview (CIDI). Geneva: WHO, 1992 (Updated 2000 http://www.who.int/msa/cidi/)

27. Murray CJL, Feachem RS, Phillips MA, Willis C. Adult morbidity: limited data and methodological uncertainty. Chapter3 in Feachem RG et. Al. (eds). The Health of adults in the Developing World, p. 113-160. New York : Oxford University Press, 1992.

28. Lopez, A., Ahmad, O., Guillot, M., Inoue, M., and Ferguson, B. Life tables for 191 countries for 2000: data, methods, results.

29. World Health Organization. Environmental Health Indicators: Framework and Methodologies. Protection of the Human Environment Occupational and Environmental Health Series. Geneva : WHO, 1999.

30. Xu, K, Kawabata, K., Hanvoravongchai, P., Murray, C. J. L., Ortiz, J. P., Zeramdini, R., and Klavus, J. Household health system contributions and capacity to pay: definitional, empirical and technical challenges.

31. Kish L. Survey sampling. New York: John Wiley, 1965

32. Dillman D.A. Mail and Internet Surveys: The Tailored Design Method. 2000. New York: John Wiley and Sons, Inc. 2nd Edition.

33. Dillman D.A. Mail and Telephone Surveys: The Total Design Method. 1978.New York: Wiley-Interscience.

34. World Health Organization. I-Shell. http://www.who.int/evidence/hhsr-survey/

35. Cohen J. A coefficient of agreement for nominal scales. Educ. Psych. Meas 1960; 20: 37-46.

36. Fleiss, J.L. Statistical Methods for Rates and Proportions. 1981. New York. John Wiley and Sons, Inc. 2nd Edition.

37. Grosh M and Glewwe P. Designing household survey questionnaires for developing countries: Lessons from 15 years of the Living Standards Measurement Study. 2000. Washington, D.C., The World Bank.

38. Macro International. Demographic and Health Surveys. http://www.measuredhs.com/

39. United Nations Children's Fund. Evaluation of Multiple Indicator Cluster Surveys. New York: UNICEF, 1997

APPENDICES

Appendix 1: List of Collaborating Institutions

A. Household Surveys

1. **China**	The Institute of Social Medicine and Health Policy, Shandong
2. **Colombia**	Pontificia Universidad Javeriana, Santafé de Bogotá DC
3. **Egypt**	Health Care International, Cairo
4. **Georgia**	Institute of Polling and Marketing, Tbilisi
5. **India**	Institute of Health Systems, Hyderabad
6. **Indonesia**	National Institute of Health Research and Development, JL.Jakarta
7. **Iran**	Ministry of Health, Tehran
8. **Lebanon**	University of Balamand, Beirut
9. **Mexico**	Instituto Nacional de Salud Publica, Mexico City
10. **Myanmar**	Ministry of Health, Yangon
11. **Nigeria**	The University of Ibadan, Ibadan
12. **Singapore**	Ministry of Health, Singapore
13. **Slovakia**	Public Health Institute of Slovak Republic, Bratislava
14. **Syria**	Ministry of Health, Damascus
15. **Turkey**	WHO Collaborating Center, AMATEM, Istanbul

B. Postal Surveys

1. **Australia**	TQA Research
2. **Austria**	Statistik Österreich
3. **Canada**	Environics Research Group
4. **Chile**	University of Concepción
5. **China-Shandong**	Shandong Medical University
6. **Cyprus**	MEMRB International-Research and Consultancy Group
7. **Czech Republic**	Institute of Health Information and Statistics
8. **Denmark**	Statistics Denmark
9. **Egypt**	Health Care International
10. **Finland**	National Research and Development Centre for Welfare and Health Outcome and Equity research Group-STAKES
11. **France**	Erik Consulting
12. **Greece**	National School of Public Health Department of Health
13. **Hungary**	SZONDA IPSOS-Market Research Inst.
14. **Indonesia**	National Institute of Health Research and Development
15. **Kyrgyzstan**	Siar-Bishkek company
16. **Lebanon**	University of St Joseph
17. **Lithuania**	Institute for Biomedical Research, Kaunas University

18.	**Netherlands**	Netherlands Organization for Applied Scientific Research (TNO)
19.	**New Zealand**	University of Otago-Linz Research Unit
20.	**Poland**	CBOS (Public Opinion Research Center)
21.	**Rep. of Korea**	School of Public Health, Seoul
22.	**Switzerland**	LINK Inst. Market & Social Research
23.	**Thailand**	Mahidol University-Salaya Campus
24.	**Trinidad & Tobago**	University of West Indies
25.	**Turkey**	Plaza Ltd. Research
26.	**Ukraine**	Kiev International Institute of Sociology
27.	**United Kingdom**	National Centre for Social Research
28.	**USA**	Washington State University

C. Brief Face-to-Face Surveys

1.	**Belgium**	International Research Associates (INRA EUROPE)
2.	**Bulgaria**	International Research Associates (INRA EUROPE)
3.	**Czech Republic**	International Research Associates (INRA EUROPE)
4.	**Estonia**	International Research Associates (INRA EUROPE)
5.	**Finland**	International Research Associates (INRA EUROPE)
6.	**France**	International Research Associates (INRA EUROPE)
7.	**Germany**	International Research Associates (INRA EUROPE)
8.	**Iceland**	International Research Associates (INRA EUROPE)
9.	**Ireland**	International Research Associates (INRA EUROPE)
10.	**Italy**	International Research Associates (INRA EUROPE)
11.	**Luxembourg**	International Research Associates (INRA EUROPE)
12.	**Malta**	International Research Associates (INRA EUROPE)
13.	**Netherlands**	International Research Associates (INRA EUROPE)
14.	**Portugal**	International Research Associates (INRA EUROPE)
15.	**Romania**	International Research Associates (INRA EUROPE)
16.	**Russian Federation**	International Research Associates (INRA EUROPE)
17.	**Spain**	International Research Associates (INRA EUROPE)
18.	**Sweden**	International Research Associates (INRA EUROPE)
19.	**Argentina**	GALLUP EUROPE
20.	**Bahrain**	GALLUP EUROPE
21.	**Costa Rica**	GALLUP EUROPE
22.	**Jordan**	GALLUP EUROPE
23.	**Latvia**	GALLUP EUROPE
24.	**Morocco**	GALLUP EUROPE
25.	**Oman**	GALLUP EUROPE
26.	**United Arab Emirates**	GALLUP EUROPE
27.	**Venezuela**	GALLUP EUROPE
28.	**Croatia**	Market, Media, and Public Opinion Research

Appendix 2: Detailed Sampling Plans for the Participating Countries

There were four different modes used in the WHO Multi-country Survey Study: 1) face-to-face long household survey, 2) brief face-to-face survey, 3) CATI survey and 4) postal (mailed or couriered) survey.

A. HOUSEHOLD MODE

In most countries where the long version of the survey was carried out the following sampling design was used:

a) Sample Size and frame

Between 5,000 and 10,000 individuals, at least 18 years old, who were not institutionalized, were selected from a sampling frame that was reasonably representative of the country's population (such as census data or electoral rolls).

b) Sampling strategy

The samples were a nationally representative multistage stratified sample in urban and rural areas in all countries. The sample population consisted of male and female adults living in private households, i.e. non-institutionalized population. One respondent per household was randomly selected either from registries or using the Kish method for selecting individuals from those who are eligible.

B. BRIEF FACE-TO-FACE MODE

The sampling design for the brief face-to-face surveys was as follows:

a) Sample Size

A sample of 1,000 to 1,500 individuals aged 18+ and non-institutionalized was selected from a sampling frame that was reasonably representative of the country's population.

b) Sampling strategy

For all brief face to face countries a multi-stage random probability sampling strategy was used.

The sampling points represented the whole territory of the country surveyed and were selected proportionally to the distribution of the population in terms of metropolitan, urban and rural areas. (except for a few countries where only urban areas were covered as they represented most of the country and rural areas are very remote and difficult to access).

C. CATI MODE

Two countries, Canada and Luxembourg carried out a CATI survey.

a) Sample Size

A sample of 1,000 individuals aged 18+, non-institutionalized and living in private households, was taken into the sampling frame from the most recently published telephone directories.

b) Sampling strategy

Telephone numbers were selected from the most recently published telephone directories. The random digit dialling method was applied enabling better sampling coverage.

D. POSTAL MODE

Most of the postal countries had the following sampling design:

a) Sample Size

A sample of 5,000 individuals aged 18+ and non- institutionalized was selected from a sampling frame that was reasonably representative of the country's population.

b) Sampling strategy

- Where possible eligible individuals were selected from acceptable frames and mailed the questionnaire. If an acceptable sampling frame of individuals was not available, households were selected. In this case, the recipient of the mailed survey in a household was asked to select from the persons in the household who are 18 or older the person with the closest birthday to complete the questionnaire. That person completed the questionnaire about themselves and their experience and not for the entire household. The questionnaire included a question about how many persons 18 or older reside in the household to assist in developing weights for the data.

- An up-to-date registry of all persons residing in the country was preferred. Alternatives such as telephone directories, registries providing partial coverage, post office listings, etc., were considered.

The country specific sampling plans, a summary of the survey characteristics as well as a graphical representation of the population deviation index for those participating in the survey is given in the following pages.

APPENDICES

Appendix 1: List of Collaborating Institutions

D. Household Surveys

16. **China**	The Institute of Social Medicine and Health Policy, Shandong
17. **Colombia**	Pontificia Universidad Javeriana, Santafé de Bogotá DC
18. **Egypt**	Health Care International, Cairo
19. **Georgia**	Institute of Polling and Marketing, Tbilisi
20. **India**	Institute of Health Systems, Hyderabad
21. **Indonesia**	National Institute of Health Research and Development, JL.Jakarta
22. **Iran**	Ministry of Health, Tehran
23. **Lebanon**	University of Balamand, Beirut
24. **Mexico**	Instituto Nacional de Salud Publica, Mexico City
25. **Nigeria**	The University of Ibadan, Ibadan
26. **Singapore**	Ministry of Health, Singapore
27. **Slovakia**	Public Health Institute of Slovak Republic, Bratislava
28. **Syria**	Ministry of Health, Damascus
29. **Turkey**	WHO Collaborating Center, AMATEM, Istanbul

E. Postal Surveys

29. **Australia**	TQA Research
30. **Austria**	Statistik Österreich
31. **Canada**	Environics Research Group
32. **Chile**	University of Concepción
33. **China-Shandong**	Shandong Medical University
34. **Cyprus**	MEMRB International-Research and Consultancy Group
35. **Czech Republic**	Institute of Health Information and Statistics
36. **Denmark**	Statistics Denmark
37. **Egypt**	Health Care International
38. **Finland**	National Research and Development Centre for Welfare and Health Outcome and Equity research Group-STAKES
39. **France**	Erik Consulting
40. **Greece**	National School of Public Health Department of Health
41. **Hungary**	SZONDA IPSOS-Market Research Inst.
42. **Indonesia**	National Institute of Health Research and Development
43. **Kyrgyzstan**	Siar-Bishkek company
44. **Lebanon**	University of St Joseph
45. **Lithuania**	Institute for Biomedical Research, Kaunas University
46. **Netherlands**	Netherlands Organization for Applied Scientific Research

	(TNO)
47. **New Zealand**	University of Otago-Linz Research Unit
48. **Poland**	CBOS (Public Opinion Research Center)
49. **Rep. of Korea**	School of Public Health, Seoul
50. **Switzerland**	LINK Inst. Market & Social Research
51. **Thailand**	Mahidol University-Salaya Campus
52. **Trinidad & Tobago**	University of West Indies
53. **Turkey**	Plaza Ltd. Research
54. **Ukraine**	Kiev International Institute of Sociology
55. **United Kingdom**	National Centre for Social Research
56. **USA**	Washington State University

F. Brief Face-to-Face Surveys

29. Belgium	International Research Associates (INRA EUROPE)
30. Bulgaria	International Research Associates (INRA EUROPE)
31. Czech Republic	International Research Associates (INRA EUROPE)
32. **Estonia**	International Research Associates (INRA EUROPE)
33. **Finland**	International Research Associates (INRA EUROPE)
34. **France**	International Research Associates (INRA EUROPE)
35. **Germany**	International Research Associates (INRA EUROPE)
36. **Iceland**	International Research Associates (INRA EUROPE)
37. **Ireland**	International Research Associates (INRA EUROPE)
38. **Italy**	International Research Associates (INRA EUROPE)
39. **Luxembourg**	International Research Associates (INRA EUROPE)
40. **Malta**	International Research Associates (INRA EUROPE)
41. **Netherlands**	International Research Associates (INRA EUROPE)
42. **Portugal**	International Research Associates (INRA EUROPE)
43. **Romania**	International Research Associates (INRA EUROPE)
44. **Russian Federation**	International Research Associates (INRA EUROPE)
45. **Spain**	International Research Associates (INRA EUROPE)
46. **Sweden**	International Research Associates (INRA EUROPE)
47. **Argentina**	GALLUP EUROPE
48. **Bahrain**	GALLUP EUROPE
49. **Costa Rica**	GALLUP EUROPE
50. **Jordan**	GALLUP EUROPE
51. **Latvia**	GALLUP EUROPE
52. **Morocco**	GALLUP EUROPE
53. **Oman**	GALLUP EUROPE
54. **United Arab Emirates**	GALLUP EUROPE
55. **Venezuela**	GALLUP EUROPE
56. **Croatia**	Market, Media, and Public Opinion Research

Appendix 2: Detailed Sampling Plans for the Participating Countries

There were four different modes used in the WHO Multi-country Survey Study: 1) face-to-face long household survey, 2) brief face-to-face survey, 3) CATI survey and 4) postal (mailed or couriered) survey.

A. HOUSEHOLD MODE

In most countries where the long version of the survey was carried out the following sampling design was used:

a) Sample Size and frame

Between 5,000 and 10,000 individuals, at least 18 years old, who were not institutionalized, were selected from a sampling frame that was reasonably representative of the country's population (such as census data or electoral rolls).

b) Sampling strategy

The samples were a nationally representative multistage stratified sample in urban and rural areas in all countries. The sample population consisted of male and female adults living in private households, i.e. non-institutionalized population. One respondent per household was randomly selected either from registries or using the Kish method for selecting individuals from those who are eligible.

B. BRIEF FACE-TO-FACE MODE

The sampling design for the brief face-to-face surveys was as follows:

a) Sample Size

A sample of 1,000 to 1,500 individuals aged 18+ and non-institutionalized was selected from a sampling frame that was reasonably representative of the country's population.

b) Sampling strategy

For all brief face to face countries a multi-stage random probability sampling strategy was used.

The sampling points represented the whole territory of the country surveyed and were selected proportionally to the distribution of the population in terms of metropolitan, urban and rural areas. (except for a few countries where only urban areas were covered as they represented most of the country and rural areas are very remote and difficult to access).

C. CATI MODE

Two countries, Canada and Luxembourg carried out a CATI survey.

a) Sample Size

A sample of 1,000 individuals aged 18+, non-institutionalized and living in private households, was taken into the sampling frame from the most recently published telephone directories.

b) Sampling strategy

Telephone numbers were selected from the most recently published telephone directories. The random digit dialling method was applied enabling better sampling coverage.

D. POSTAL MODE

Most of the postal countries had the following sampling design:

a) Sample Size

A sample of 5,000 individuals aged 18+ and non- institutionalized was selected from a sampling frame that was reasonably representative of the country's population.

b) Sampling strategy

- Where possible eligible individuals were selected from acceptable frames and mailed the questionnaire. If an acceptable sampling frame of individuals was not available, households were selected. In this case, the recipient of the mailed survey in a household was asked to select from the persons in the household who are 18 or older the person with the closest birthday to complete the questionnaire. That person completed the questionnaire about themselves and their experience and not for the entire household. The questionnaire included a question about how many persons 18 or older reside in the household to assist in developing weights for the data.

- An up-to-date registry of all persons residing in the country was preferred. Alternatives such as telephone directories, registries providing partial coverage, post office listings, etc., were considered.

The country specific sampling plans, a summary of the survey characteristics as well as a graphical representation of the population deviation index for those participating in the survey is given in the following pages.

The sample was a stratified random multi-stage sample representative of all inhabitants of Argentina aged 18 +. At the first stage of sampling, using the stratification as per geographical criteria, the country was divided into six regions.

The sampling selection criteria adopted for Argentina was as follows:

a) Buenos Aires

The official cartography provided by the National Census was used for the sample frame. The procedure consisted in stratifying the census ratio according to two criteria:

- Geographical location;
- And social class, defined by the educational level of the head of the household.

Using this stratification, the census ratio was selected and within each one, a block was randomly selected.

b) Rest of the Country

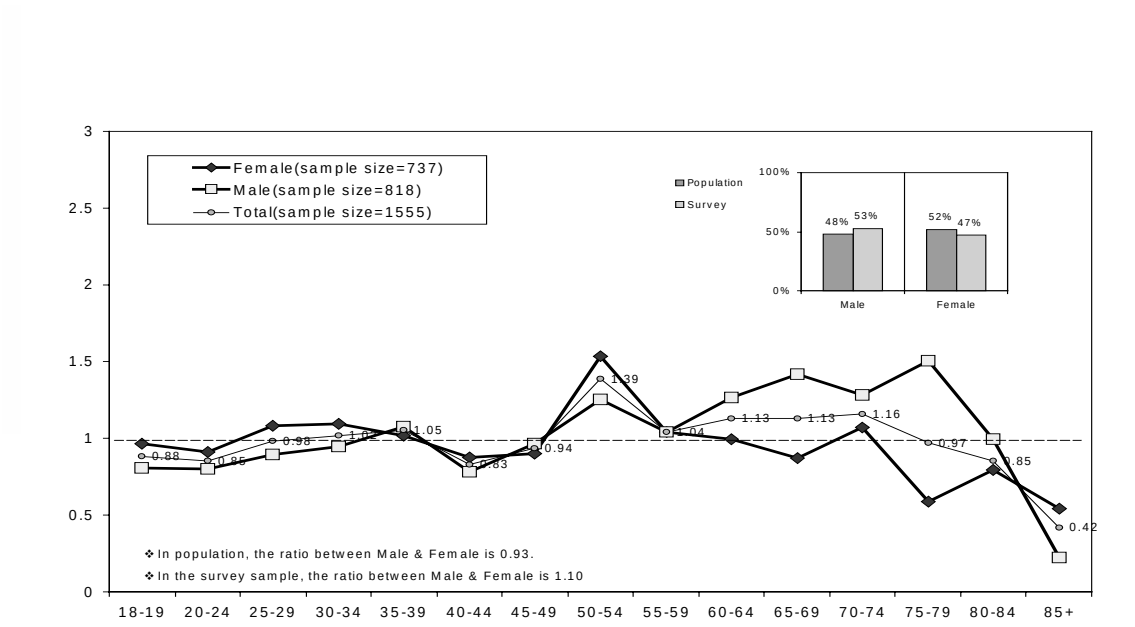
Within each locality selected (representing the first stage sampling unit), the census ratio (second stage unit) was ordered by social class and a sample within them was chosen using a random start. In each census ratio that was selected, the same criteria used for Buenos Aires area was applied to get to the final unit sample. (the respondent).

c) General Sampling Aspects:

Over 250 different sampling points were selected on a mathematically random basis from within localities. In each sampling point, four interviews were conducted. Only one person per household was interviewed. If the person who opened the door matched the quota requirements (sex and age), this person was interviewed. If not, the correct target was looked for in the household.

Survey summary characteristics

Final Sample size	% Response rate	% Missing data	Deviation from UN population Chi-square Critical Value = 49.58 (29 df, p=0.01)
1,555	36	3.9	469.5



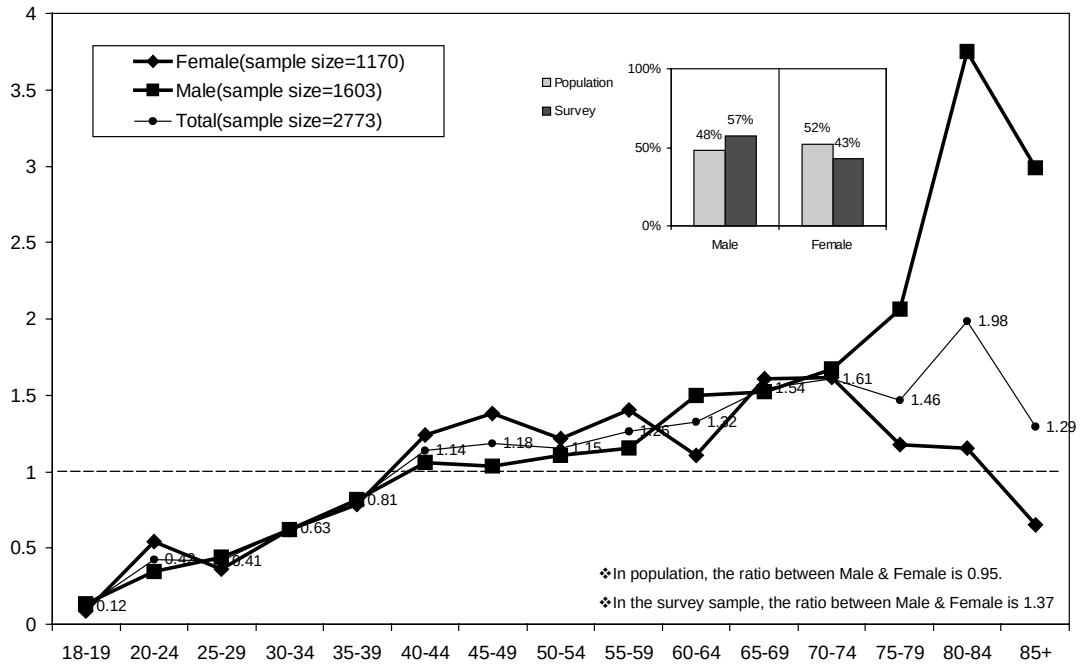
Two possible sampling frames were considered by the survey company: the telephone directory and the electoral roll data. The selected sampling frame was the electoral roll data, which enables a broader coverage of the Australian adult population in all areas.

A sample of 8000 individuals over the age of 18 years was randomly selected from the Australian Electoral Role for use in this study. The database, which was up to date as of February 2000 (and continuously updated since), was provided by ASIS List Services. The ASIS database currently includes over 12 million individual records and is guaranteed to be above the 97% level of accuracy.

Accordingly, the sample chosen should be very representative of the Australian population.

Survey summary characteristics

Final Sample size	% Response rate	% Missing data	Deviation from UN population Chi-square Critical Value = 49.58 (29 df, p=0.01)
2,773	35	0	746



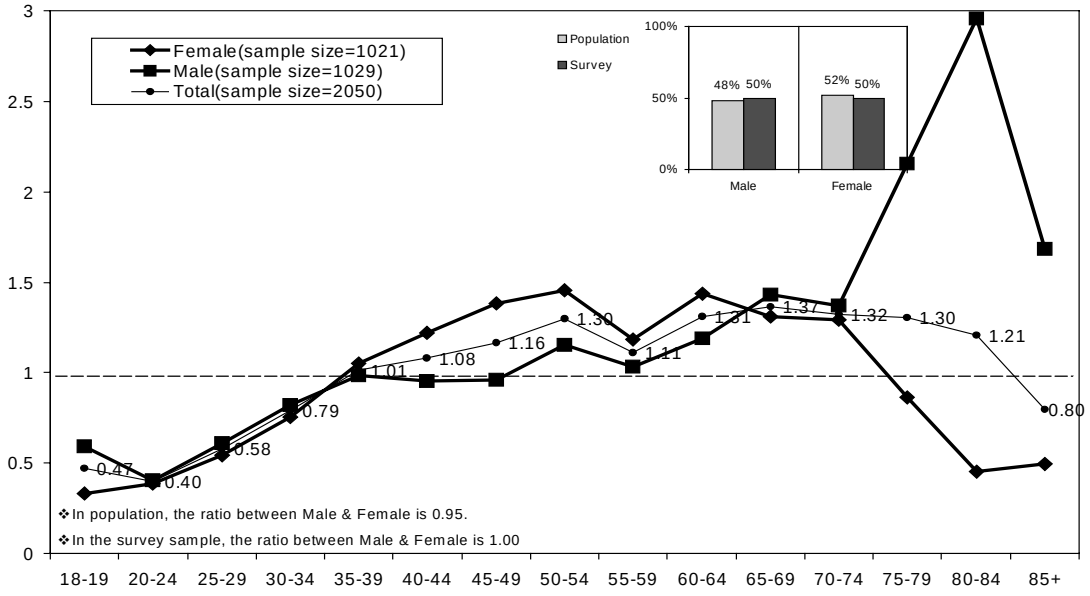
The Austrian Microcensus is the main household sample survey of Statistics Austria.
The gross sample size is 31,500 dwellings and the net sample size of about 23,000 households. It includes nine samples for the Austrian Länder, ranging between 2,700 and 4,600 dwellings (gross sample size).

In all Länder, except for the city of Vienna and Vorarlberg where the sample is a one-stage stratified-random sample, there is a two-stage-stratified-random sample.

Addresses were drawn from the housing census from 1991 and from the yearly register of newly built dwellings.

Survey summary characteristics

Final Sample size	% Response rate	% Missing data	Deviation from UN population Chi-square Critical Value = 49.58 (29 df, p=0.01)
2,089	56	7.3	310



The sample was a multi-stage random probability sample representative of the population residing in urban and rural areas of Bahrain. A sample design method in 2 stages was used:

1. The survey was conducted in the three main municipalities of Bahrain.
2. The second stage consisted in the selection of the households within the primary sampling areas.

The sampling design employed for the main urban centres was as follows:

- The three main urban centers were divided into Administrative Units which themselves were subdivided into clusters (consisting of an agglomeration within a determined polygon of roads and streets).
- Each cluster has a certain number of blocks. The latter are defined as the smallest tract of land outlined by streets or roads that contain houses and buildings.
- In each block, buildings and houses are identified and counted. The selection of the respondent was done using the Kish Method.

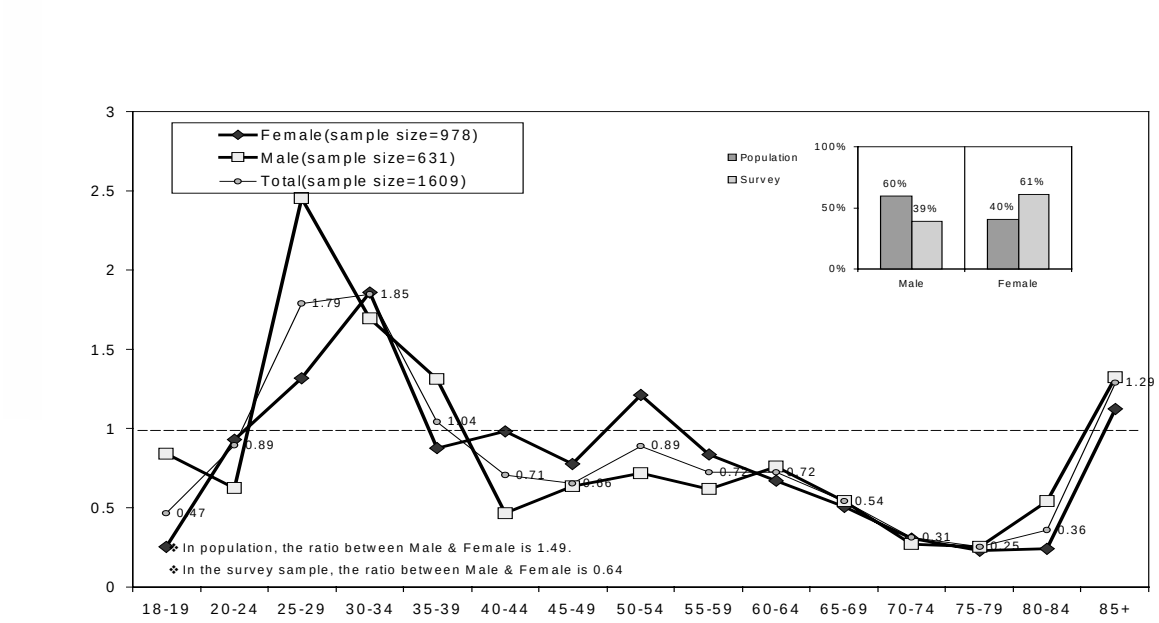
Statistical data acquired from the Block census was used in the sample design of this study. The density of the population varies from one Administrative Unit to another. They were classified into 3 categories:

- high density
- medium density
- low density areas.

The number of sampling units assigned for interviewing per Administrative Unit varied so that the population density was adequately represented.

Survey summary characteristics

Final Sample size	% Response rate	% Missing data	Deviation from UN population Chi-square Critical Value = 49.58 (29 df, p=0.01)
1,609	35	0.2	408.8



The metropolitan, urban and rural population and all "administrative regional units" as defined in Official Europe Union Statistics (NUTS 2) covered proportionately the respective population aged 18 and above. The country was divided into an appropriate number of areas, grouping NUTS regions at whatever level appropriately. The NUTS covered in Belgium were the following; Antwerpen, Vlaams Brabant, Brabant Wallon, Brussels, West-Vlaanderen, Oost-Vlaanderen, Hainaut, Liege, Limburg, Luxembourg and Namur.

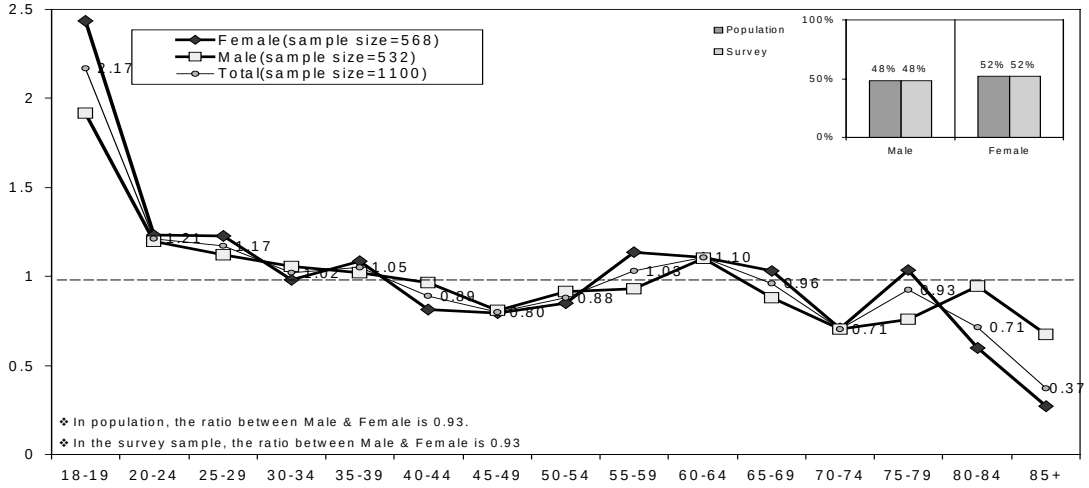
The basic sample design was a multi-stage, random probability sample. 100 sampling points were drawn with probability proportional to population size, for a total coverage of the country. The sampling points were drawn after stratification by NUTS 2 region and by degree of urbanisation. They represented the whole territory of the country surveyed and were selected proportionally to the distribution of the population in terms of metropolitan, urban and rural areas.

In each of the selected sampling points, one address was drawn at random. This starting address formed the first address of a cluster of a maximum of 20 addresses. The remainder of the cluster was selected as every Nth address by standard random route procedure from the initial address. In theory, there was no maximum number of addresses issued per country. Procedures for random household selection and random respondent selection were independent of the interviewer's decision and controlled by the institute responsible. They should be as identical as possible from country? to country, full functional equivalence being a must.

At every address up to 4 recalls were made to attempt to achieve an interview with the selected respondent. There was only one interview per household. The final sample size is 1,100 completed interviews.

Survey summary characteristics

Final Sample size	% Response rate	% Missing data	Deviation from UN population Chi-square Critical Value = 49.58 (29 df, p=0.01)
1,100	48	5.5	83.6



The metropolitan, urban and rural population and all "administrative regional units" as defined in Official Europe Union Statistics (NUTS 2) covered proportionately the respective population aged 18 and above. The country was divided into an appropriate number of areas, grouping NUTS regions at whatever level appropriately. The NUTS covered in Bulgaria were the following; Sofia Stolitsa, Severna Balgarija, Yuzhna Balgarija.

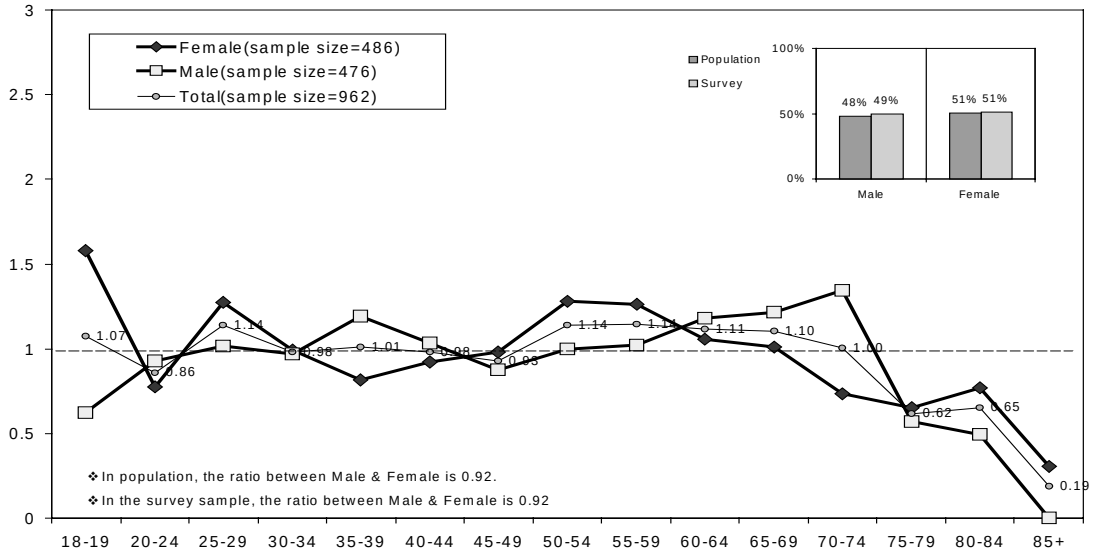
The basic sample design was a multi-stage, random probability sample. 100 sampling points were drawn with probability proportional to population size, for a total coverage of the country. The sampling points were drawn after stratification by NUTS 2 region and by degree of urbanisation. They represented the whole territory of the country surveyed and were selected proportionally to the distribution of the population in terms of metropolitan, urban and rural areas.

In each of the selected sampling points, one address was drawn at random. This starting address formed the first address of a cluster of a maximum of 20 addresses. The remainder of the cluster was selected as every Nth address by standard random route procedure from the initial address. In theory, there was no maximum number of addresses issued per country. Procedures for random household selection and random respondent selection were independent of the interviewer's decision and controlled by the institute responsible. They should be as identical as possible from to country, full functional equivalence being a must.

At every address up to 4 recalls were made to attempt to achieve an interview with the selected respondent. There was only one interview per household. The final sample size is 1,010 completed interviews.

Survey summary characteristics

Final Sample size	% Response rate	% Missing data	Deviation from UN population Chi-square Critical Value = 49.58 (29 df, p=0.01)
1,010	88	1.7	46.7

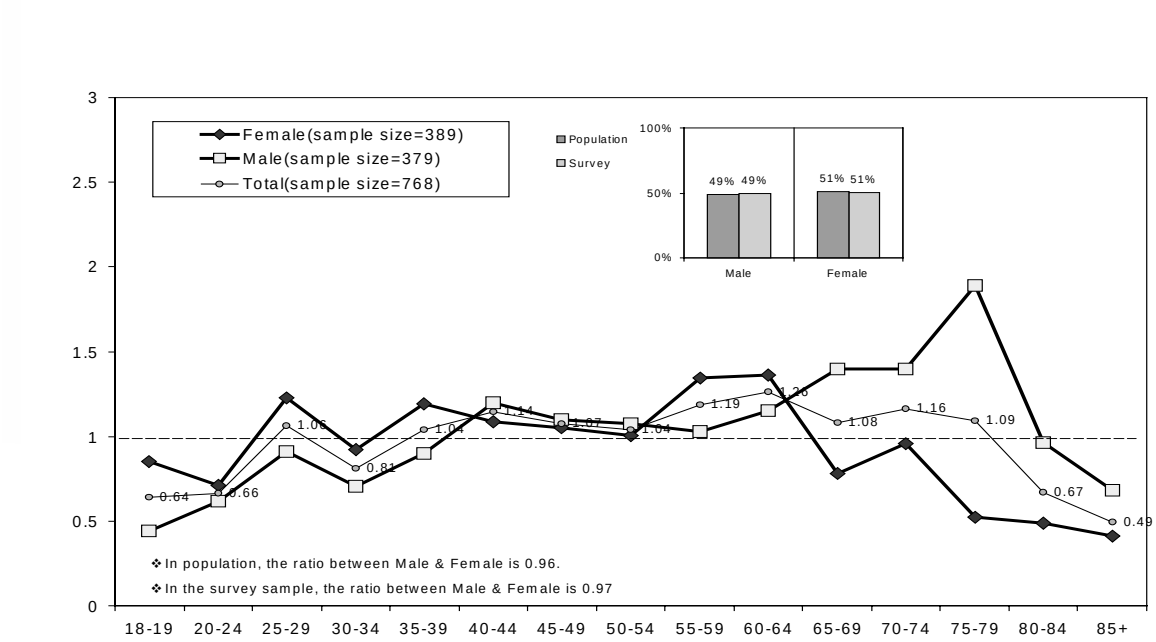


1,487 named individuals were selected from the Karom Group of Companies, Dialogue Canada household mail panel. This mail panel includes a cross-section of Canadians, with the exception of those living in the Yukon, Northwest Territories or Nunavut, from which a sample can be obtained to represent the Canadian population according to the most recent Statistics Canada data. The panel file was stratified by regions in Canada: city size, French Quebec and rest of Canada and ordered by postcode. The 1,487 named individuals were selected from the Dialogue Mail panel file, using a random method on the sample sorted by postcode.

Individual members of each household who were asked to complete the survey were identified by birth date and gender with this identifying information
From the initial 1,487 mailed out, 816 questionnaires came back hence reaching a response rate of 55%.

Survey summary characteristics

Final Sample size	% Response rate	% Missing data	Deviation from UN population Chi-square Critical Value = 49.58 (29 df, p=0.01)
816	55	2.4	158.4



The sample was drawn in such a way that it represented the Canadian population with the exception of the Canadians living in the Yukon, Northwest Territories or Nunavut.

The sampling model relied on the stratification of the population by ten provinces and by six community sizes.

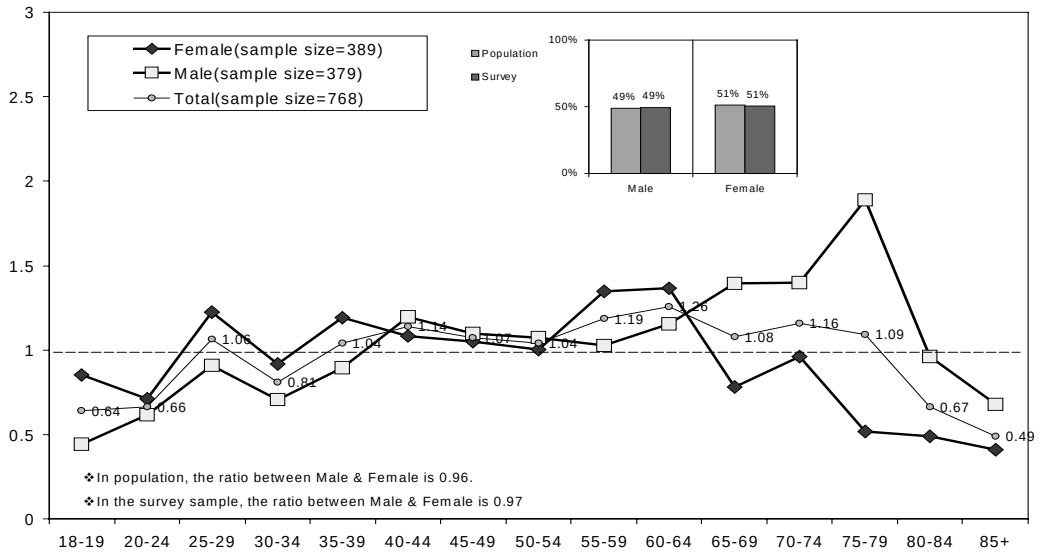
Telephone numbers were selected from the most recently published telephone directories. These numbers acted as "seeds" from which the sample was actually generated. The original "seed" telephone numbers were not used in the sample. Both unlisted numbers and numbers listed after the directory publication are included in the sample.

From within each household contacted, respondents 18 years of age and older were screened for random selection using the most recent birthday method.

From the 12,350 total calls made, 778 calls completed the interview. Among the 12,350 calls, 8,466 were ineligibles and from the latter, 5,305 calls for which the respondent was unavailable. The net response rate is therefore 24.6%.

Survey summary characteristics

Final Sample size	% Response rate	% Missing data	Deviation from UN population Chi-square Critical Value = 49.58 (29 df, p=0.01)
778	24.6	2.6	85.1



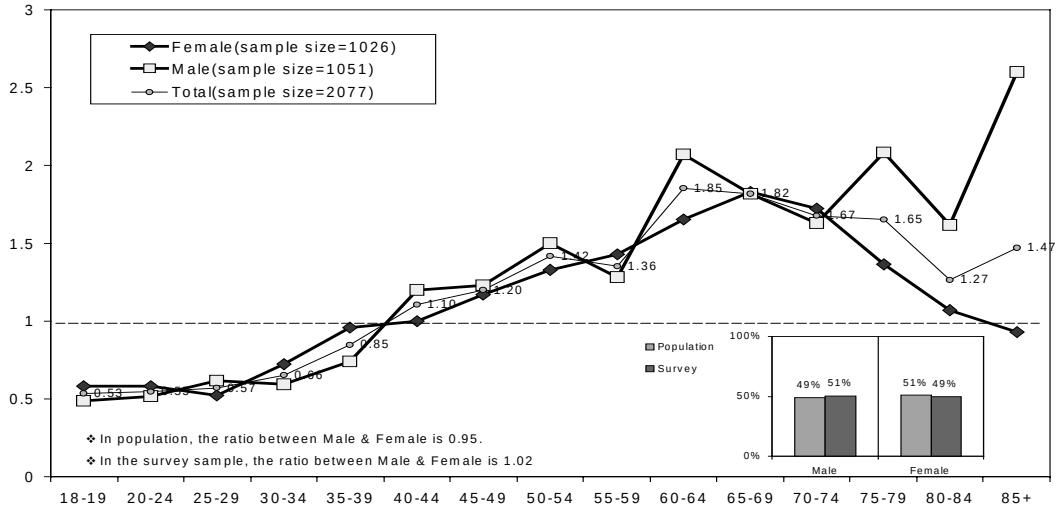
The telephone directory was used as the sampling frame since it is considered as the most reliable registry available. The following percentage of inhabitants from each of the following regions were selected to obtain a sample of 5,000 individuals.

REGIONS	**%**	**No. of individuals**
Tarapaca	2.6	130
Antofagasta	3.1	155
Atacama	1.8	90
Coquimbo	3.8	190
Valparaiso	10.3	515
Lib. O'higgins	5.2	260
Maule	6.0	300
Bio-bio	12.8	640
Araucanía	5.8	290
De los lagos	7.0	350
Aisen	0.6	30
Magallanes	1.0	50
r.m. santiago	40.0	2000

Each region was divided into provinces. The provinces are composed of "comunas" or municipalities from within which individuals were randomly selected. However, with this design, there may be a bias towards the population without a telephone.

Survey summary characteristics

Final Sample size	% Response rate	% Missing data	Deviation from UN population Chi-square Critical Value = 49.58 (29 df, p=0.01)
2,078	42	4.2	410.7



A nationally representative sample of male and female adults age 18+ was used. Three provinces from 3 economic levels were sampled as follows: Shandong (high), Henan (middle), Gansu (low).

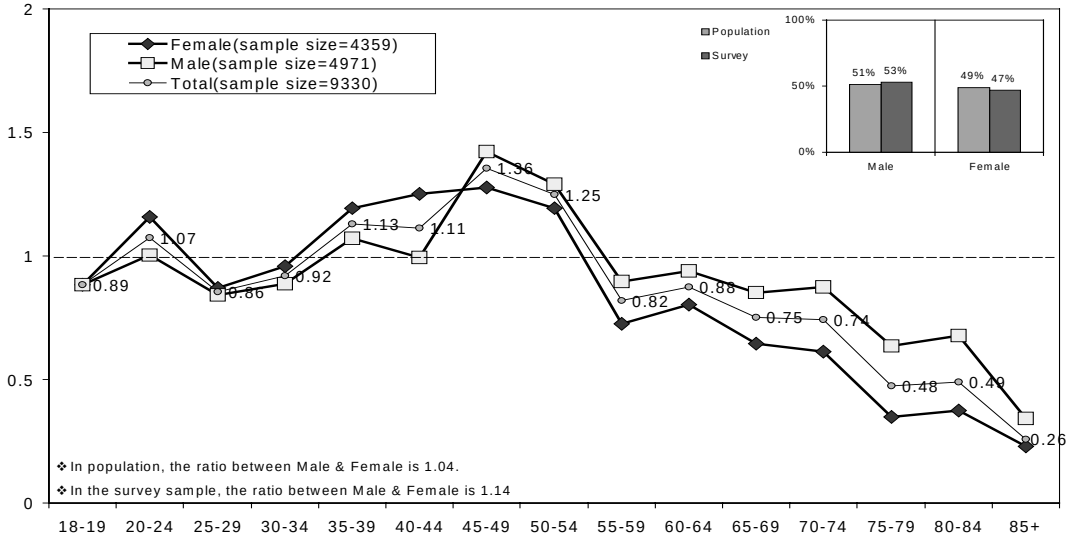
5,000 people in Shandong, 3,000 in Henan and 2,000 in Gansu were sampled. From the sample, 53.2% males vs. 46.8% females were interviewed.

In each province 33.7% of the respondents were interviewed in urban area, and 66.3% in rural area. According to the economic level of each province, 3-6 counties were chosen randomly. Respondents were selected randomly according to their household number.

Missing rates were quite low, as respondents generally tended to cooperate. Illiterate respondents found some questions were too difficult (health state valuations, HSR ranking, calibration tests). Interviews were also too long and the average time for one interview was at least 2 hours if the respondent had little education.

Survey summary characteristics

Final Sample size	% Response rate	% Missing data	Deviation from UN population Chi-square Critical Value = 49.58 (29 df, p=0.01)	% of items with Kappa > 0.4
9,442	99	10	425.7	99



Shandong province was selected for the survey. The respondents were selected from urban and rural areas respectively, using the household registry system which provides information on names, age, gender, education and address for each household member.

In rural areas, two counties, one high-income and the other low-income, were selected. 4,000 respondents were selected from these two counties.

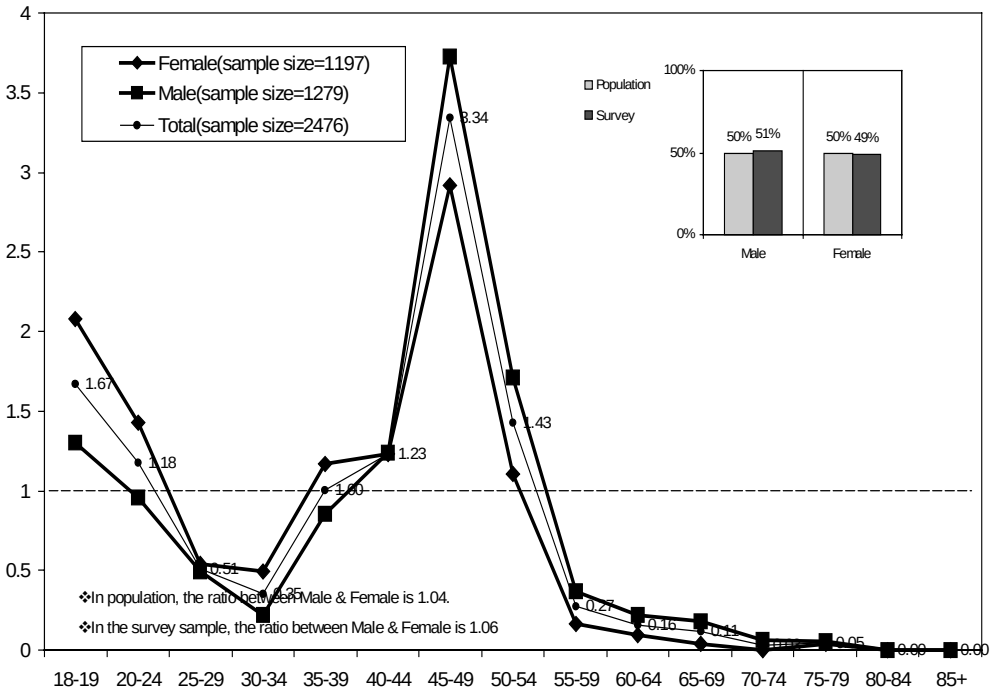
In the urban area, Jinan City was selected. 1,000 respondents were selected from six districts within this city.

Sample sizes in rural and urban areas were estimated according to the present rural/urban population ratio in Shandong.

The individual selected from the household was 18+ years in age and the closest birthday method was used to select the respondent.

Survey summary characteristics

Final Sample size	% Response rate	% Missing data	Deviation from UN population Chi-square Critical Value = 49.58 (29 df, p=0.01)
2,480	50	0.4	1969.3



The sample was provided by the National Institute of Statistics (DANE) and corresponded to the 1993 national census. The sample used was a two-stage, non-equal cluster, probabilistic and stratified sample.

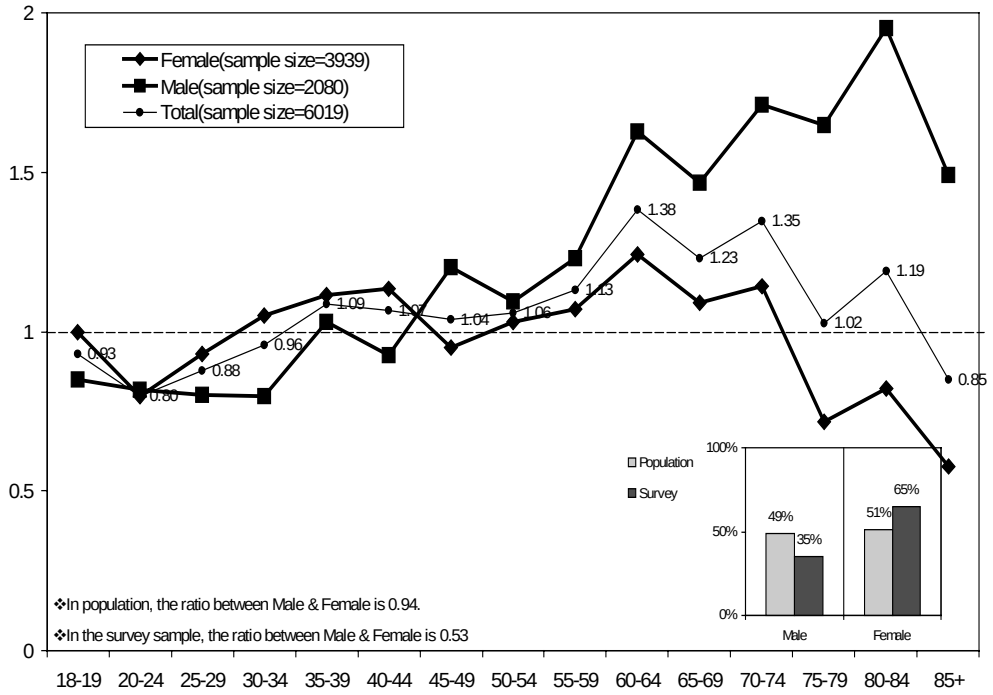
Thirty-three municipalities were sampled. Some PSUs (Orinoquia, the Amazonian Triangle, and the Pacific Coast - Coast from Nariño, Cauca and Valle, and the state of Chocó- with an urban population that represents less than 2% of the urban population of the country) were eliminated as access was difficult (vast areas, bad roads and high transport cost).

From the sample of 6,000, females accounted for 65.4% and males for 34.6%.

The major problems reported which affected the response rate were the violence, fear of kidnapping, and political instability.

Survey summary characteristics

Final Sample size	% Response rate	% Missing data	Deviation from UN population Chi-square Critical Value = 49.58 (29 df, p=0.01)	% of items with Kappa > 0.4
6,019	82	31.9	201.8	49



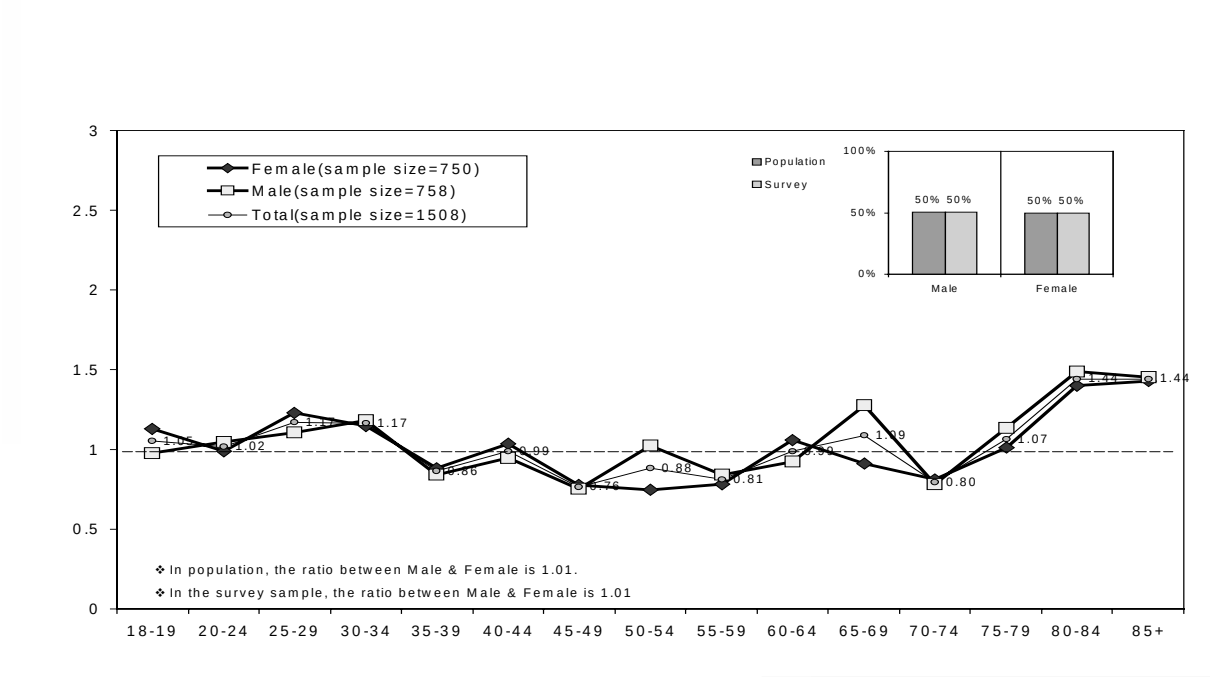
The country is divided into two regions, the Greater Metropolitan Area and the Rest of the country. The number of interviews in each area was proportionate to its population.

The sampling selection criteria adopted for Costa Rica was that the sampling segments where the interviews were to be held were randomly selected using the maps provided by the National Institute of Statistics and Census of Costa Rica. The procedure took into account the size of the sampling points according to the number of households in each segment (equal selection probability for each household). The resulting sample was self-weighted in terms of geographical population distribution.

General aspects about the sampling strategy in Costa Rica is that 170 different sampling points were selected on a mathematically random basis from the main geographical areas. In each sampling point, eight to ten interviews were conducted. Respondents were selected using the birthday method.

Survey summary characteristics

Final Sample size	% Response rate	% Missing data	Deviation from UN population Chi-square Critical Value = 49.58 (29 df, p=0.01)
1,508	37	2.1	36.4



The sample was a multi-stage stratified random sampling, representative of the Croatian population.

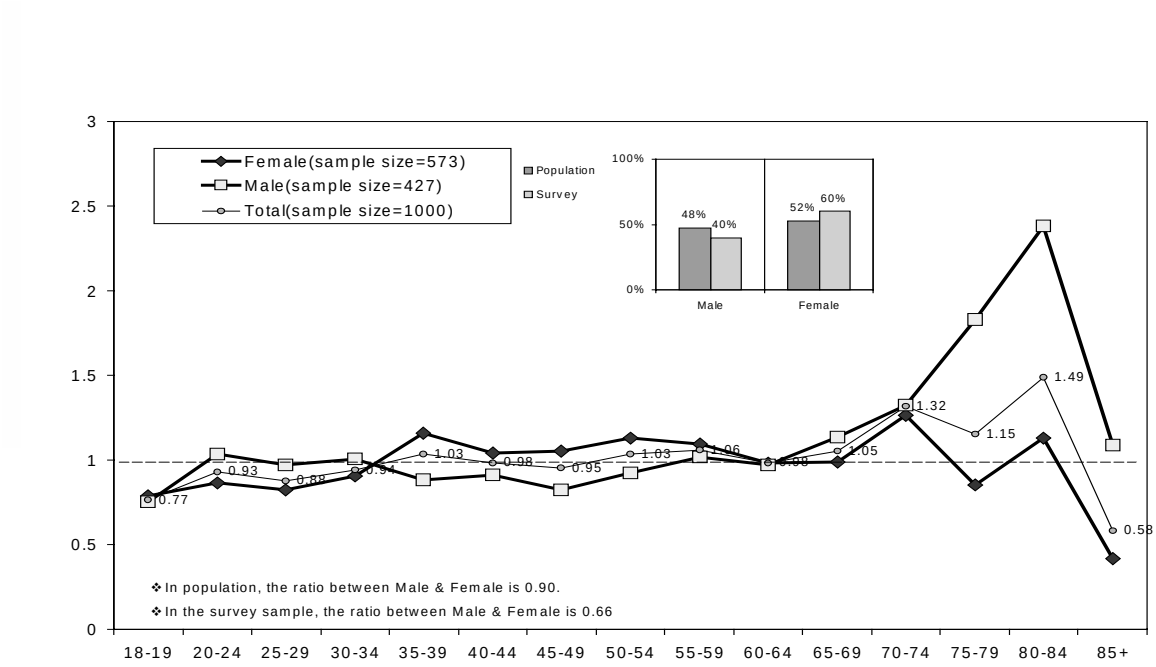
Croatia was divided into 7 distinct geographic regions; Zagreb & suroundings, Istria, Rijeka and Kvarner, Dalmatia, North Croatia, Slavonia, Lika & Banovina. Each of these regions covers various number of cities ("very small" with less than 2,000 inhabitants, "small" with 2,000-10,000, "medium" with 10,000-100,000 inhabitants and "large" with over of 100,000 inhabitants).

The planned fixed size of the entire stratified sample was 3,000 of adult persons, the proportionate allocation of units, according to the Census data, was constructed. The cities and villages were sampled at random using random digit generator with probabilities proportionate to the size (the measure of size being "the number of inhabitants") to ensure that in the further stage of the sampling procedure every person has the same probability of selection.
It was decided that every "large city" (with more than 100.000 inhabitants) has to be included in the sample.

The addresses of the households were chosen first by sampling of the several "Census areas" in the particular place and defining the starting point. The size of the sampling point was 16 respondents. The interviewers were provided with addresses of the starting point and systematic sampling from defined starting point (random walk - right hand rule) chose households. The choice of the respondent in the particular household was done according to the rules of Trohdal-Carter-Bryant (TCB) method.

Survey summary characteristics

Final Sample size	% Response rate	% Missing data	Deviation from UN population Chi-square Critical Value = 49.58 (29 df, p=0.01)
3,000	68	0.5	94.3



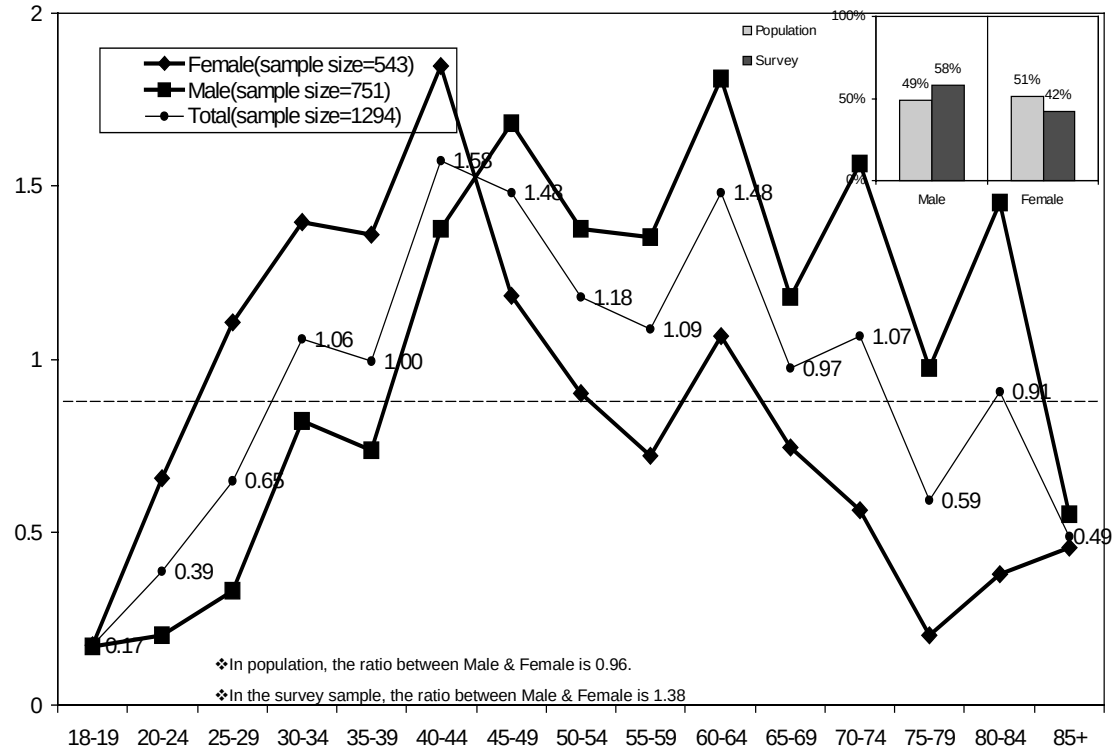
10,000 individuals were selected from the telephone directory, 5,000 of whom were subsequently removed, leaving a desired sample of 5000 participants. The survey was carried out in both urban and rural areas, but only in the Government controlled areas of the island, which accounts for approximately two-thirds of the island.

The addresses that were removed initially were chosen on the basis of their postal unsuitability (i.e. they were not private addresses, or were somehow inadequate) and with an intention to represent each district as well as urban and rural areas fairly.

The remaining extra addresses were removed randomly and the usable sample was divided into 8 sub-samples of 625 participants.

Survey summary characteristics

Final Sample size	% Response rate	% Missing data	Deviation from UN population Chi-square Critical Value = 49.58 (29 df, p=0.01)
1,374	27	3.2	331.8

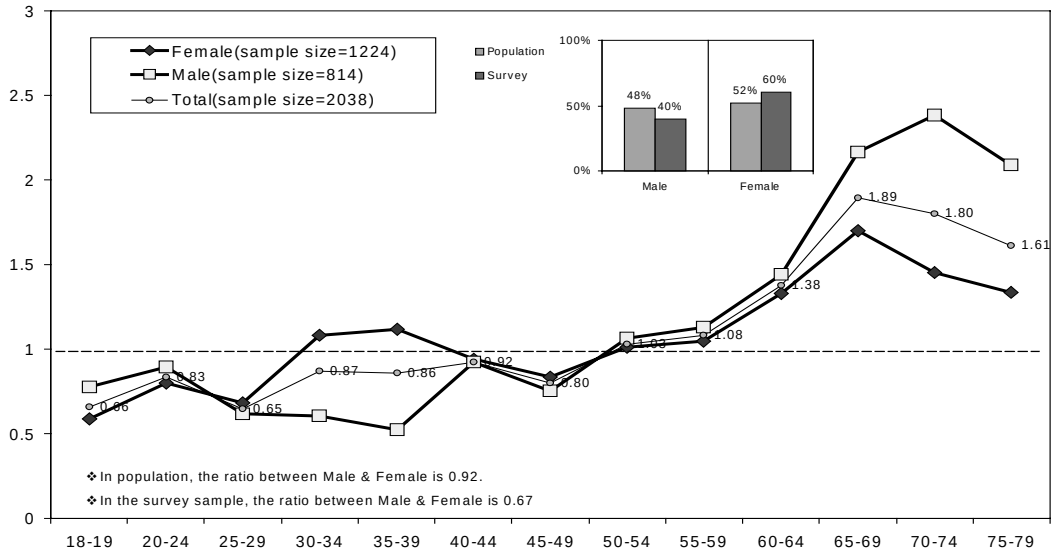


The sample was drawn from the Central Population Registry of the Czech Republic. It covers both urban and rural areas and is an up-to-date registry of the population living in the country.

A representative sample of 5,700 individuals, born between 1922 and 1982, was randomly selected.

Survey summary characteristics

Final Sample size	% Response rate	% Missing data	Deviation from UN population Chi-square Critical Value = 49.58 (29 df, p=0.01)
2,020	40	4.0	359.8



The metropolitan, urban and rural population and all "administrative regional units" as defined in Official Europe Union Statistics (NUTS 2) covered proportionately the respective population aged 18 and above. The country was divided into an appropriate number of areas, grouping NUTS regions at whatever level appropriately. The NUTS covered in the Czech Republic were the following; Praha, Stredni Cechy, Jihozapad, Severozapad, Severovychod, Jihovychod, Stredni Morova, Ostravsko.

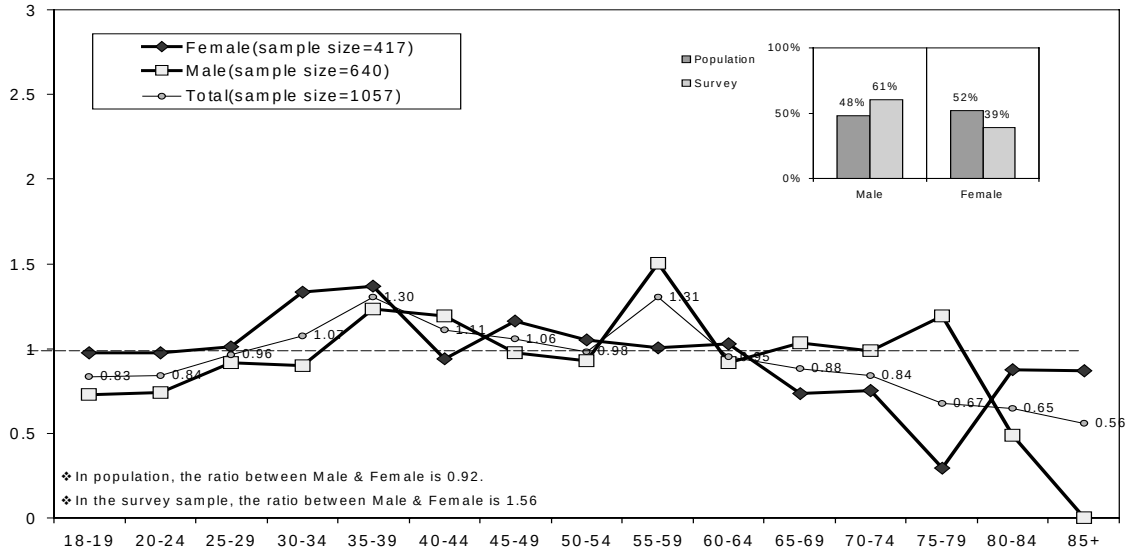
The basic sample design was a multi-stage, random probability sample. 100 sampling points were drawn with probability proportional to population size, for a total coverage of the country. The sampling points were drawn after stratification by NUTS 2 region and by degree of urbanisation. They represented the whole territory of the country surveyed and are selected proportionally to the distribution of the population in terms of metropolitan, urban and rural areas.

In each of the selected sampling points, one address was drawn at random. This starting address formed the first address of a cluster of a maximum of 20 addresses. The remainder of the cluster was selected as every Nth address by standard random route procedure from the initial address. In theory, there was no maximum number of addresses issued per country. Procedures for random household selection and random respondent selection were independent of the interviewer's decision and controlled by the institute responsible. They should be as identical as possible from to country, full functional equivalence being a must.

At every address up to 4 recalls were made to attempt to achieve an interview with the selected respondent. There was only one interview per household. The final sample size was 1,090 completed interviews.

Survey summary characteristics

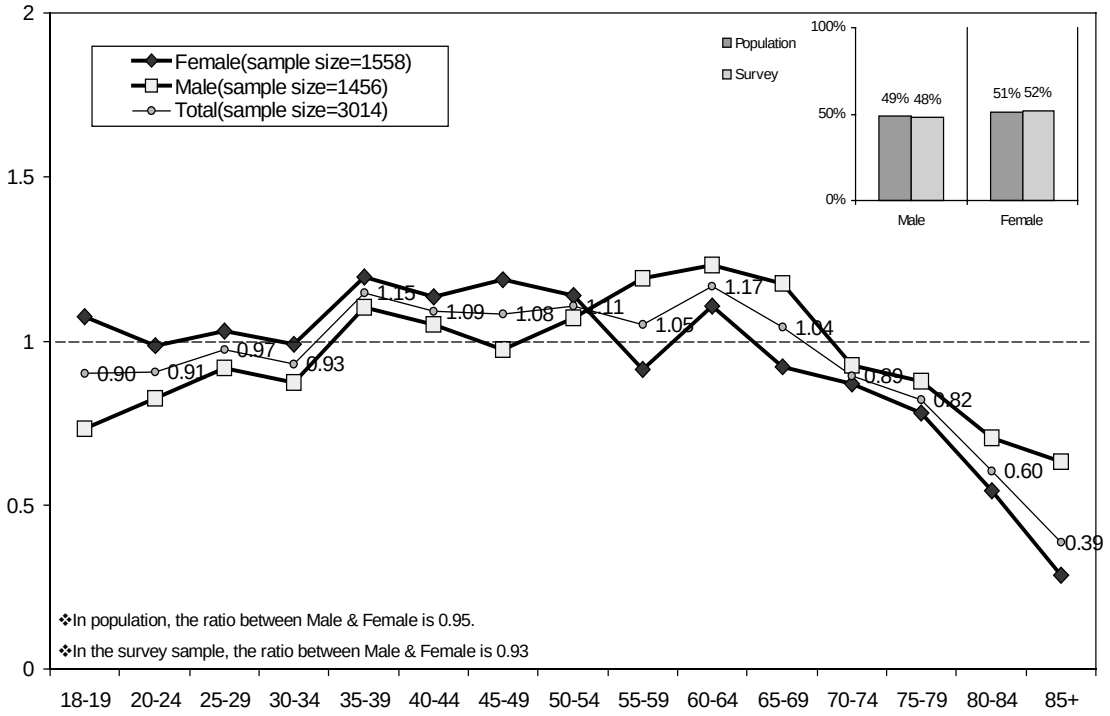
Final Sample size	% Response rate	% Missing data	Deviation from UN population Chi-square Critical Value = 49.58 (29 df, p=0.01)
1,090	60	0.4	54.2



A representative sample of 5,200 individuals aged 18+ years old residing in private households was selected randomly from the Central Population Register. This is an up-to-date registry of the population residing in the country.

Survey summary characteristics

Final Sample size	% Response rate	% Missing data	Deviation from UN population Chi-square Critical Value = 49.58 (29 df, p=0.01)
2,684	54	5.3	91.9



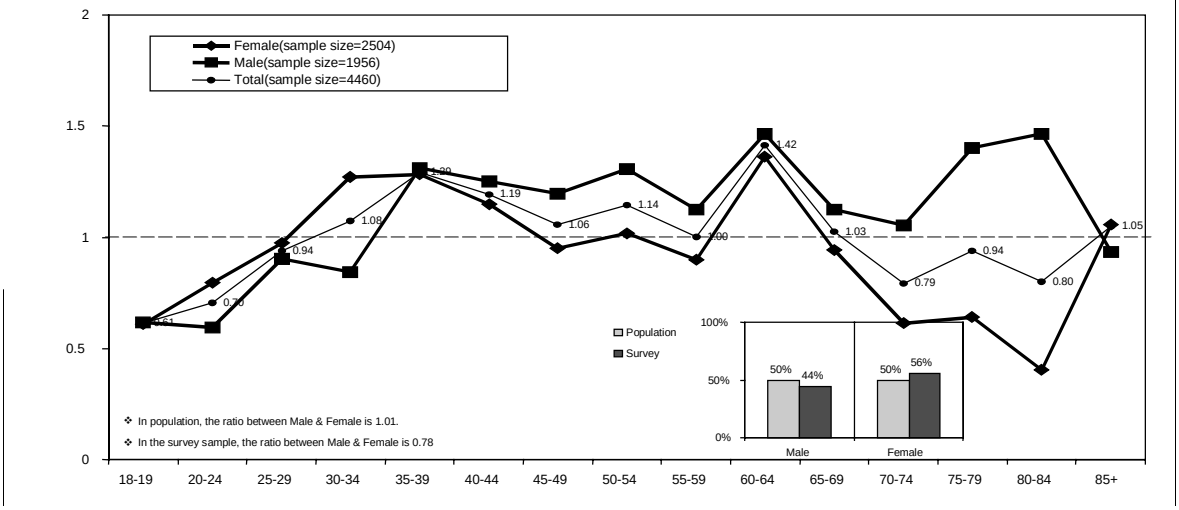
The sample was selected and approved by CAPMAS (Central Agency for Public Mobilization and Statistics) from six Governorates: Alexandria, Gharbeia, Sharkeya, Fayoum, Luxor, Matrouh.

These Governorates represented metropolitan, lower, upper, and frontier Governorates. The sample criteria for selection depended on gender, age, residence and education. A sample of 5,000 respondents was selected and was representative of the population density, socioeconomic pattern, age groups, gender, and regions (urban & rural). More females (56.2%) than males (43.8%) were interviewed, which is likely due to more women being at home.

Respondents found some questions in the questionnaire quite sensitive such as those dealing with intimate family relationships, personal habits, and illegal practices, which caused embarrassment, anxiety, shame, psychological stress and, sometimes a strong emotional reaction. Other problems reported were the difficulty of some questions (HSV) which created confusion and resulted in a low accurate response rate. Some questions were found to be unnecessary, other unusual (calibration tests). The reaction of some respondents was negative towards some questions as they saw them as childish or insulting. Further, some tests created (visual test, PLM) constraints due to unavailability of space in houses and impracticality of doing the test in open areas outside houses. In general, Ramadhan month affected the pace of work, data collection, retest and entry.

Survey summary characteristics

Final Sample size	% Response rate	% Missing data	Deviation from UN population Chi-square Critical Value = 49.58 (29 df, p=0.01)	% of items with Kappa > 0.4
4,486	99	12.1	283.9	97



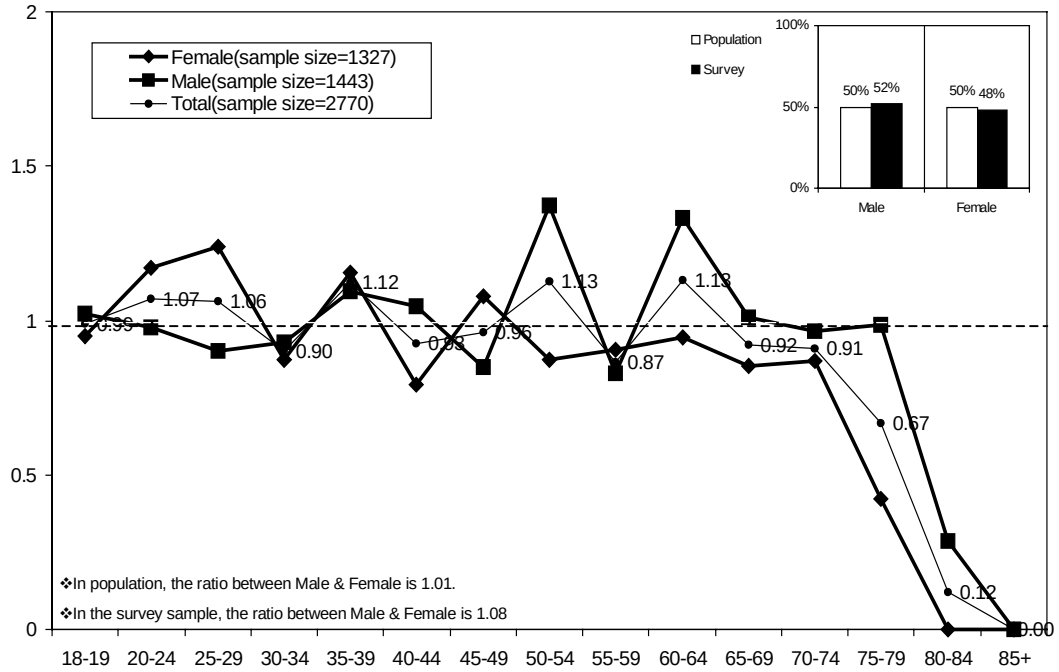
A sample of 3,000 respondents was selected and approved by CAPMAS from seven Governorates representing metropolitan, lower-Egypt, and upper-Egypt Governorates.

Based on the published 1996 census, a set of selection criteria was adopted to ensure that respondents representing age, gender, education, socioeconomic, and other categories of variables were included. In Egypt there is no updated registry system. CAPMAS, therefore, depended on the block system developed during the 1996 census.

A stratified sample was drawn to ensure representativeness of the sample based on predetermined important characteristics of the population.

Survey summary characteristics

Final Sample size	% Response rate	% Missing data	Deviation from UN population Chi-square Critical Value = 49.58 (29 df, p=0.01)
2,771	92	1.9	85.1



The metropolitan, urban and rural population and all "administrative regional units" as defined in Official Europe Union Statistics (NUTS 2) covered proportionately the respective population aged 18 and above. The country was divided into an appropriate number of areas, grouping NUTS regions at whatever level appropriately. The NUTS covered in Estonia were the following; Tallinn, North-Estonia, West-Estonia, Tartu Region, Soutth Estonia, Virumaa.

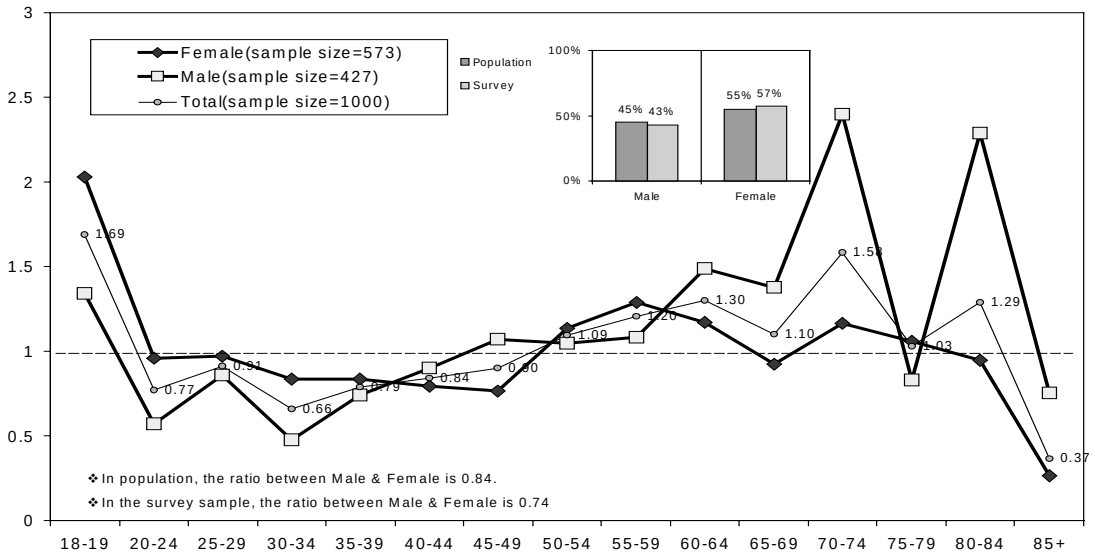
The basic sample design was a multi-stage, random probability sample. 100 sampling points were drawn with probability proportional to population size, for a total coverage of the country. The sampling points were drawn after stratification by NUTS 2 region and by degree of urbanisation. They represented the whole territory of the country surveyed and are selected proportionally to the distribution of the population in terms of metropolitan, urban and rural areas.

In each of the selected sampling points, one address was drawn at random. This starting address forms the first address of a cluster of a maximum of 20 addresses. The remainder of the cluster was selected as every Nth address by standard random route procedure from the initial address. In theory, there is no maximum number of addresses issued per country. Procedures for random household selection and random respondent selection are independent of the interviewer's decision and controlled by the institute responsible. They should be as identical as possible from to country, full functional equivalence being a must.

At every address up to 4 recalls were made to attempt to achieve an interview with the selected respondent. There was only one interview per household. The final sample size is 1,000 completed interviews.

Survey summary characteristics

Final Sample size	% Response rate	% Missing data	Deviation from UN population Chi-square Critical Value = 49.58 (29 df, p=0.01)
1,000	71	0	120.2



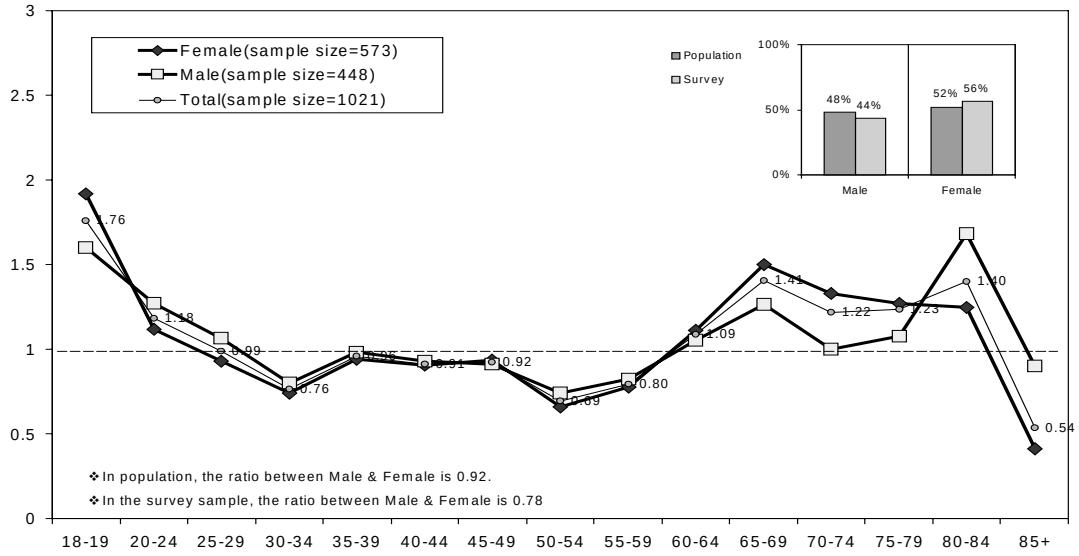
The metropolitan, urban and rural population and all "administrative regional units" as defined in Official Europe Union Statistics (NUTS 2) covered proportionately the respective population aged 18 and above. The country was divided into an appropriate number of areas, grouping NUTS regions at whatever level appropriately. The NUTS covered in Finland were the following; Etalä-Savo, Etelä-Karjala, Etelä-Pohjanmaa, Häme, Kainuu, Keski-Suomi, Kymenlaakso, Lappi, Pirkanmaa, Pohjois-Karjala, Pohjois-Pohjanmaa, Pohjois-Savo, Satakunta, Uusimaa, Vaasan rannikkoseutu, Varsinais-Suomi.

The basic sample design was a multi-stage, random probability sample. 100 sampling points were drawn with probability proportional to population size, for a total coverage of the country. The sampling points were drawn after stratification by NUTS 2 region and by degree of urbanisation. They represented the whole territory of the country surveyed and are selected proportionally to the distribution of the population in terms of metropolitan, urban and rural areas. In each of the selected sampling points, one address was drawn at random. This starting address forms the first address of a cluster of a maximum of 20 addresses. The remainder of the cluster was selected as every Nth address by standard random route procedure from the initial address. In theory, there is no maximum number of addresses issued per country. Procedures for random household selection and random respondent selection are independent of the interviewer's decision and controlled by the institute responsible. They should be as identical as possible from to country, full functional equivalence being a must.

At every address up to 4 recalls were made to attempt to achieve an interview with the selected respondent. There was only one interview per household. The final sample size is 1,021 completed interviews.

Survey summary characteristics

Final Sample size	% Response rate	% Missing data	Deviation from UN population Chi-square Critical Value = 49.58 (29 df, p=0.01)
1,021	52	0.2	70.1

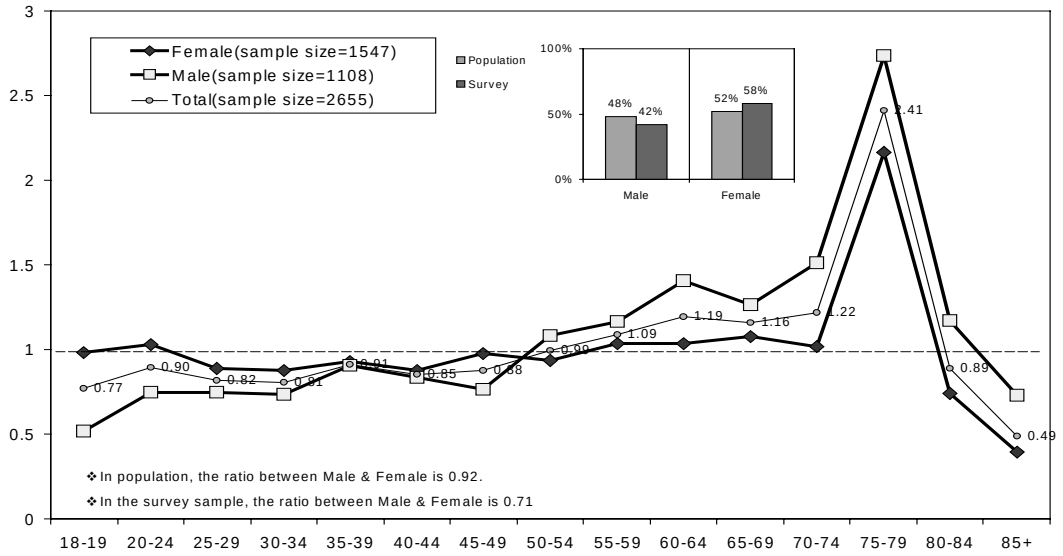


The sample (N=5000) was randomly drawn from the Finnish Population Register, which is an up-to-date registry of all persons residing in the entire country in the beginning of the year 2000.

The register includes the following information: name, address, date of birth, gender and native language. The sample was restricted to non-institutionalized persons aged 18 years or above with Finnish as their native language. The Swedish speaking population (only 5 % of inhabitants) was excluded from the sampling frame.

Survey summary characteristics

Final Sample size	% Response rate	% Missing data	Deviation from UN population Chi-square Critical Value = 49.58 (29 df, p=0.01)
2,692	54	10.1	300.8



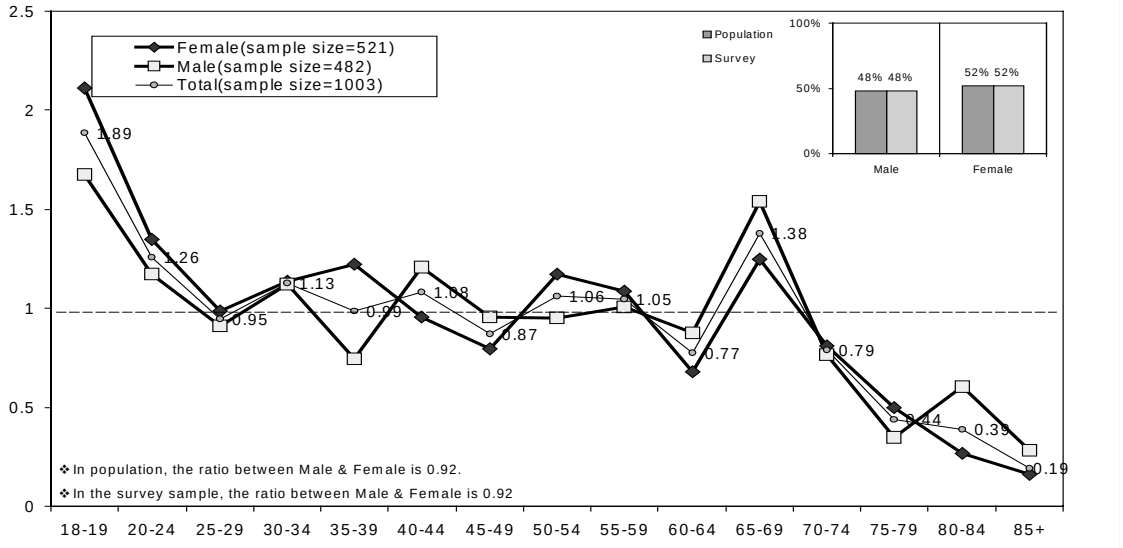
The metropolitan, urban and rural population and all "administrative regional units" as defined in Official Europe Union Statistics (NUTS 2) covered proportionately the respective population aged 18 and above. The country was divided into an appropriate number of areas, grouping NUTS regions at whatever level appropriately. The NUTS covered in France were the following; Alsace, Aquitaine, Auvergne, Basse Normandie, Bourgogne, Bretagne, Centre, Champagne-Ardennes, Corse, Franche-Comté, Haute Normandie, Ile de France, Languedoc-Roussillon, Limousin, Lorraine, Midi-Pyrénées, Nord/Pas-de-Calais, Pays de la Loire, Picardie, Poitou-Charentes, Provence-Alpes-Côte d'Azur, Rhône-Alpes.

The basic sample design was a multi-stage, random probability sample. 100 sampling points were drawn with probability proportional to population size, for a total coverage of the country. The sampling points were drawn after stratification by NUTS 2 region and by degree of urbanisation. They represented the whole territory of the country surveyed and are selected proportionally to the distribution of the population in terms of metropolitan, urban and rural areas. In each of the selected sampling points, one address was drawn at random. This starting address forms the first address of a cluster of a maximum of 20 addresses. The remainder of the cluster was selected as every Nth address by standard random route procedure from the initial address. In theory, there is no maximum number of addresses issued per country. Procedures for random household selection and random respondent selection are independent of the interviewer's decision and controlled by the institute responsible. They should be as identical as possible from to country, full functional equivalence being a must.

At every address up to 4 recalls were made to attempt to achieve an interview with the selected respondent. There was only one interview per household. The final sample size is 1,003 completed interviews.

Survey summary characteristics

Final Sample size	% Response rate	% Missing data	Deviation from UN population Chi-square Critical Value = 49.58 (29 df, p=0.01)
1,003	77	1.0	104.2



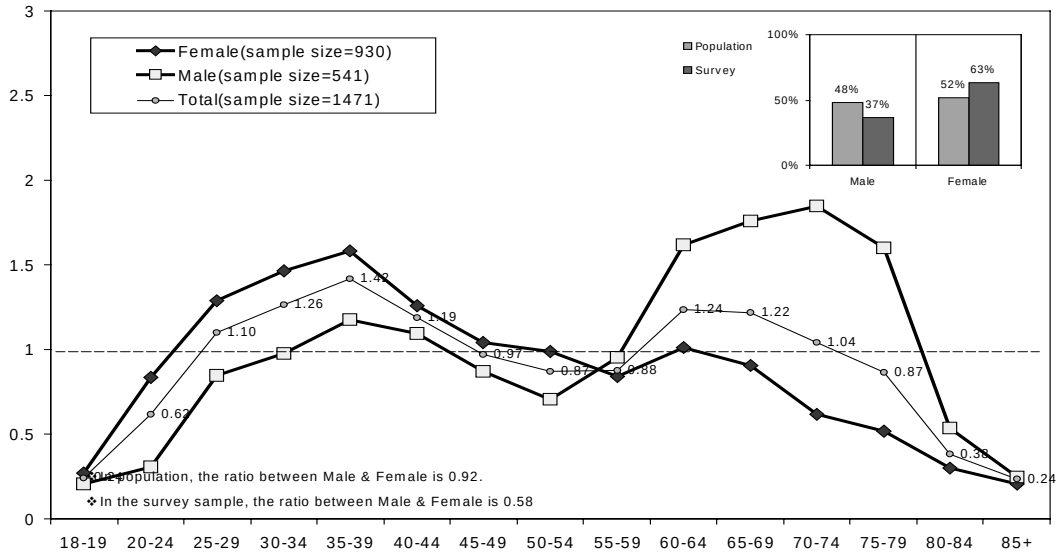
5,000 named individuals were selected randomly from a customer panel which consisting of 1,117,913 singles and 3,175,342 couples.

The sample covered urban and rural areas and included all socio-professional groups.

Each socio-professional group was represented proportionally.

Survey summary characteristics

Final Sample size	% Response rate	% Missing data	Deviation from UN population Chi-square Critical Value = 49.58 (29 df, p=0.01)
1525	31	6.3	239.3



The last census was carried out in Georgia in 1989. Because of various political and economical events in the country, such as conflict in Abkhazia and Tskhinvali region, civil war, etc., which caused migration, there are no population lists available that could be used for the sampling purposes. Lists prepared for elections are inaccurate. Based on the existing statistical data, a random sample design was used and a Random Walk Procedure was followed. This design was exceptionally accepted by WHO. A total of 10 regions were sampled and 10,000 were drawn from these regions:

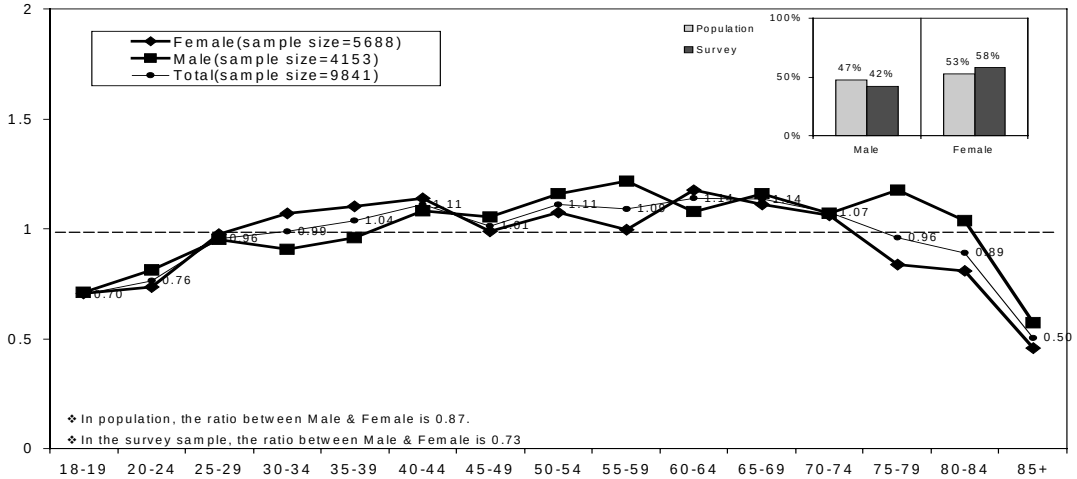
Region 1: Tbilisi
Region 2: Ajara
Region 3: Guria
Region 4: Imereti
Region 5: Kakheti
Region 6: Mstkheta-Mtianeti
Region 7: Samegrelo
Region 8: Samtskhe-Javakheti
Region 9: Kvemo Kartli
Region 10: Shida Kartli

The sampling frame covered urban and rural areas, however due to the political situation the Abkhazia and Tskhinvali regions were excluded. More females (57.8%) than males (42.2%) were interviewed.

Because of the questionnaire size and the difficult winter period of the fieldwork a higher non-response rate was anticipated. However, the total percentage of non-responses was much lower than expected. The main reasons of refusals to participate in interviews were mistrust, fear, and irritation due to their bad socioeconomic conditions. As well, interview duration was reported as being a problem. Further, in regions and sub regions of Georgia with a predominant non-Georgian population the language barrier became one additional negative factor, even if a bilingual questionnaire was used. In the Kvemo Kartli region, the Azeri population hardly understood either Georgian or Russian. Another problem was religion. Female Muslim respondents were not allowed to participate in the survey without the permission of their husbands who often were present during the interviews.

Survey summary characteristics

Final Sample size	% Response rate	% Missing data	Deviation from UN population Chi-square Critical Value = 49.58 (29 df, p=0.01)	% of items with Kappa > 0.4
9.847	87	5.4	214.3	69

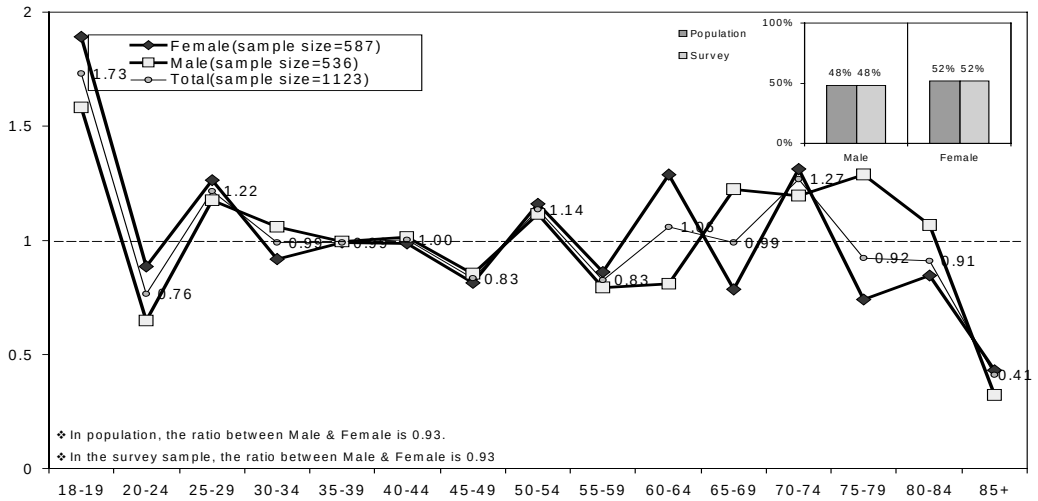


The metropolitan, urban and rural population and all "administrative regional units" as defined in Official Europe Union Statistics (NUTS 2) covered proportionately the respective population aged 18 and above. The country was divided into an appropriate number of areas, grouping NUTS regions at whatever level appropriately. The NUTS covered in Germany were the following; Arnsberg, Berlin-Ost, Berlin-West, Bremen, Chemnitz, Cottbus, Darmstadt, Detmold, Dreden, Leipzig, Düsseldorf, Frankfurt/Oder, Gera, Suhl, Giessen, Hall, Erfurt, Hamburg, Kassel, Koblenz, Köln, Magdeburg, Mittelfranken, Münster, Neubrandenburg, Niederbayern, Nordbaden-Karlsruhe, Nordwürttemberg-Stuttgart, Oberbayern, Oberfranken, Oberpfalz, Potsdam, RB Lüneburg, RB Braunsweig, RB Weser-EMS, RB Hannover, Rheinhessen-Pfalz, Rostock, Saarland, Schleswig Holstein, Schwaben, Schwerin, Südbaden-Freiburg, Südwürttemberg-Tübingen, Trier, Unterfranken.

The basic sample design was a multi-stage, random probability sample. 100 sampling points were drawn with probability proportional to population size, for a total coverage of the country. The sampling points were drawn separately in West Germany and East Germany. The sampling points were drawn after stratification by NUTS 2 region and by degree of urbanisation. They represented the whole territory of the country surveyed and are selected proportionally to the distribution of the population in terms of metropolitan, urban and rural areas. In each of the selected sampling points, one address was drawn at random. This starting address forms the first address of a cluster of a maximum of 20 addresses. The remainder of the cluster was selected as every Nth address by standard random route procedure from the initial address. In theory, there is no maximum number of addresses issued per country. Procedures for random household selection and random respondent selection are independent of the interviewer's decision and controlled by the institute responsible. They should be as identical as possible from to country, full functional equivalence being a must. At every address up to 4 recalls are made to attempt to achieve an interview with the selected respondent. There was only one interview per household. The final sample size is 1,123 completed interviews.

Survey summary characteristics

Final Sample size	% Response rate	% Missing data	Deviation from UN population Chi-square Critical Value = 49.58 (29 df, p=0.01)
1,123	80	3.7	63.5



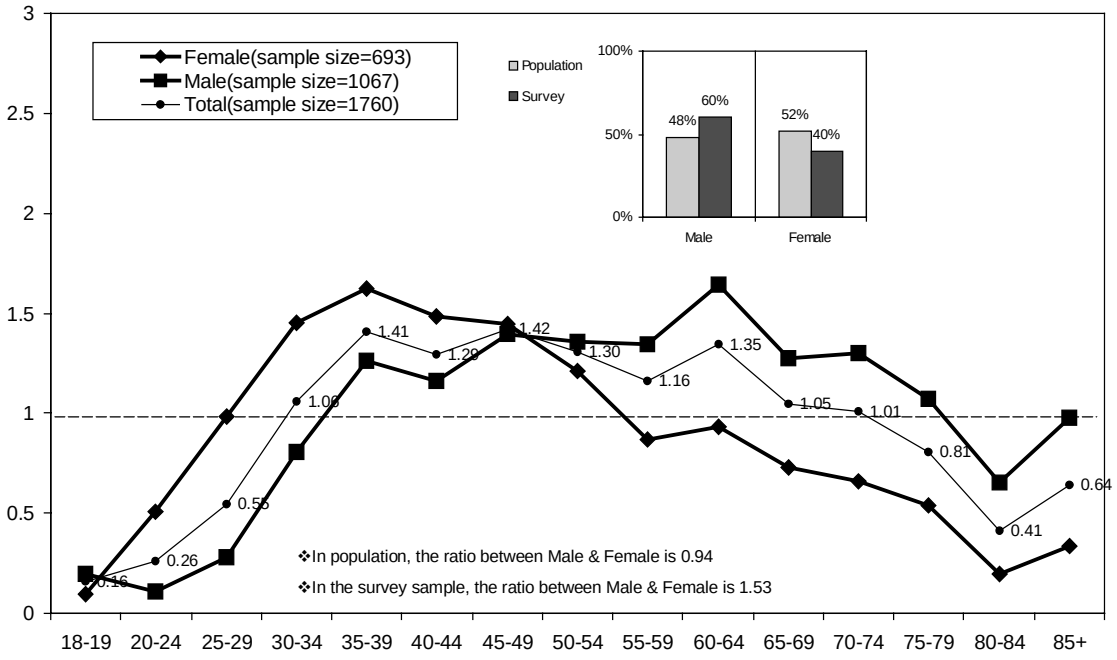
5,000 named individuals non institutionalised were randomly selected from the Greek Telephone Company catalogues. 226 entries were discarded as non eligible (business addresses etc.). The catalogue of subscribers is available on CD-ROM and is up-dated electronically every six months. Total subscribers number approximately 5,500,000. Each entry includes the telephone number, full name and address of subscriber (without post-code). Subscribers are all residents of Greece who have a fixed telephone number, regardless of nationality or citizenship.

The catalogue consists of three parts, representing the Greater Athens area, the Greater Salonika area, and the rest of the country. Within each part, the catalogues are further divided into counties.

The sample was stratified by county and ordered by post code with representative numbers proportionally selected from each Region.

Survey summary characteristics

Final Sample size	% Response rate	% Missing data	Deviation from UN population Chi-square Critical Value = 49.58 (29 df, p=0.01)
1,750	35	3.3	359.6



A two-tier, proportionately stratified random sample was used. In the first stage, the settlements (sampling points) were randomly selected so that the composition of the settlements in the sample, according to several variables such as population size, infrastructure, etc., follows the composition of the entire territory of Hungary.

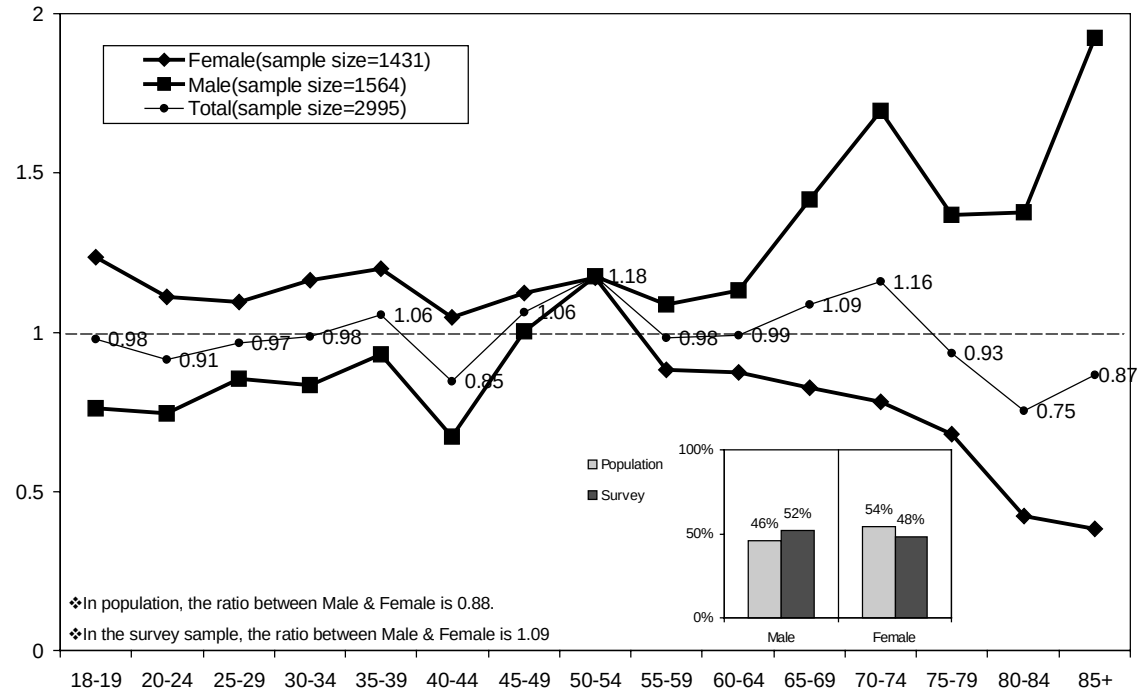
In the second stage the number of respondents to be interviewed in each settlement was determined proportionate to the population size of the settlement. Eighty settlements were used as sampling points: Budapest as one with its 23 districts and 79 rural settlements.

The sample represents the settlement network of Hungary. The composition of the persons in the sample according to sex, age and residence is identical with the composition of the population above the age of 18.

As a sampling frame Szonda Ipsos used the address database produced by the National Office for Censuses, which is a quarterly up-dated entire electronic registry of all Hungarian persons residing in the country.

Survey summary characteristics

Final Sample size	% Response rate	% Missing data	Deviation from UN population Chi-square Critical Value = 49.58 (29 df, p=0.01)
2,996	72	4.1	171.6



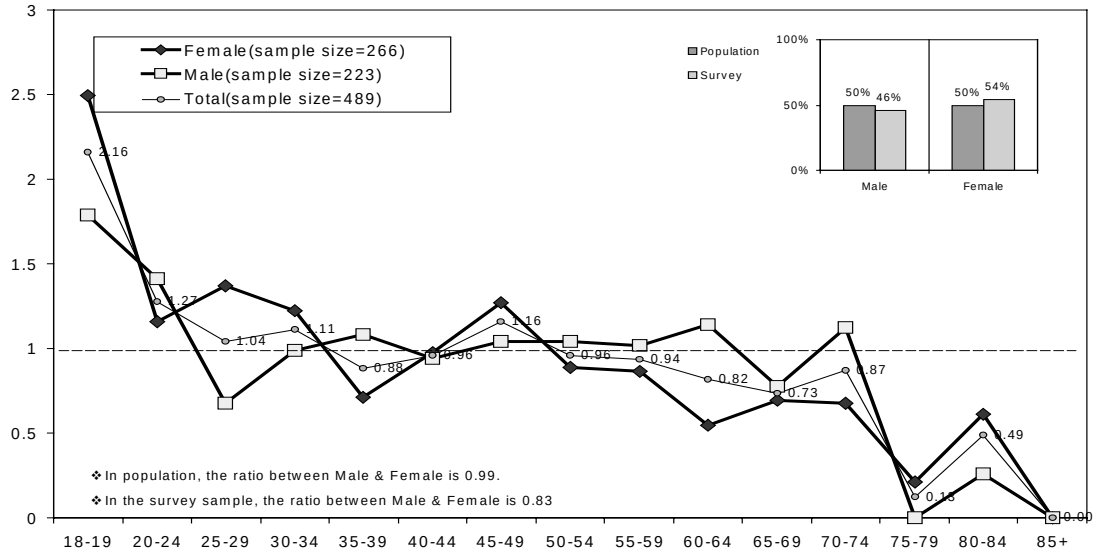
The metropolitan, urban and rural population and all "administrative regional units" as defined in Official Europe Union Statistics (NUTS 2) covered proportionately the respective population aged 18 and above. The country was divided into an appropriate number of areas, grouping NUTS regions at whatever level appropriately. The NUTS covered in Iceland were the following; Reykjavik, Near Reykjavik and Sudurnes, West-Iceland, North-Iceland, East-Iceland, South-Iceland.

The basic sample design was a multi-stage, random probability sample. 50 sampling points were drawn with probability proportional to population size, for a total coverage of the country. The sampling points were drawn after stratification by NUTS 2 region and by degree of urbanisation. They represented the whole territory of the country surveyed and are selected proportionally to the distribution of the population in terms of metropolitan, urban and rural areas. In each of the selected sampling points, one address was drawn at random. This starting address forms the first address of a cluster of a maximum of 20 addresses. The remainder of the cluster was selected as every Nth address by standard random route procedure from the initial address. In theory, there is no maximum number of addresses issued per country. Procedures for random household selection and random respondent selection are independent of the interviewer's decision and controlled by the institute responsible. They should be as identical as possible from to country, full functional equivalence being a must.

At every address up to 4 recalls were made to attempt to achieve an interview with the selected respondent. There was only one interview per household. The final sample size is 489 completed interviews.

Survey summary characteristics

Final Sample size	% Response rate	% Missing data	Deviation from UN population Chi-square Critical Value = 49.58 (29 df, p=0.01)
489	53	0.8	77.1



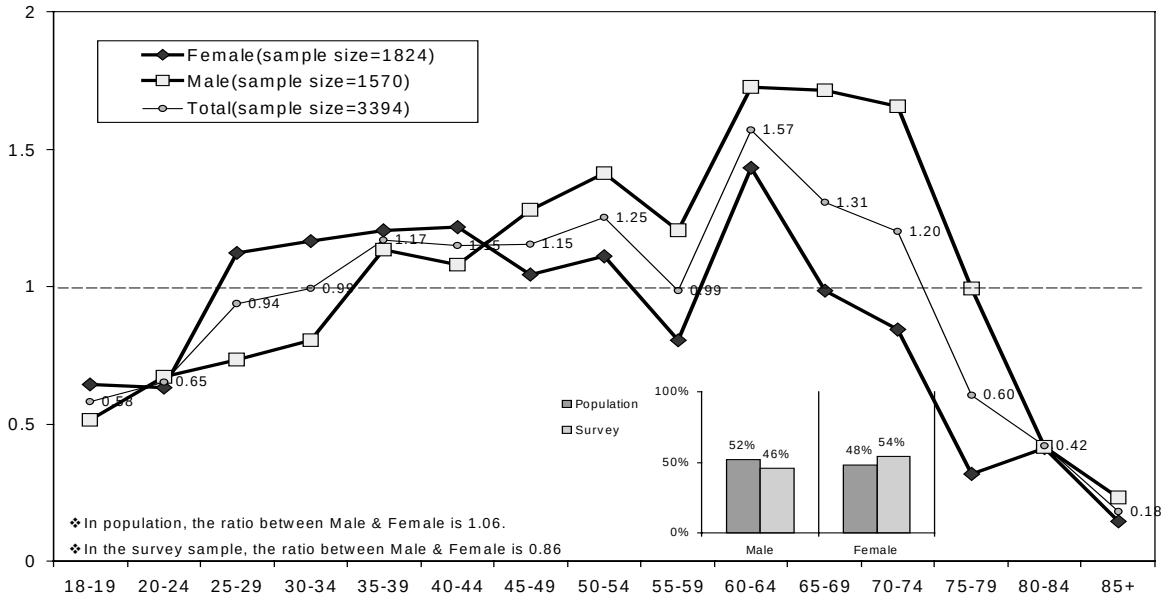
The survey was conducted in one state of India, Andhra Pradesh, and a sample of 5,000 respondents was used. The sampling procedure for the selection of clusters was a multistage, stratified and random procedure. The following strata were sampled: Rural, Urban (Municipalities), Urban (Municipal Corporations), Hyderabad.

Electoral rosters were used to select households. More females (53.3%) than males (46.7%) were interviewed.

The main problem that India faced was the floods in August, which delayed fieldwork as it affected infrastructure and communications. Some areas inland could only be reached once the rain had stopped.

Survey summary characteristics

Final Sample size	% Response rate	% Missing data	Deviation from UN population Chi-square Critical Value = 49.58 (29 df, p=0.01)	% of items with Kappa > 0.4
5,145	98	0.8	512.7	91



Indonesia is now composed of 26 provinces, which are divided into 3 regions. The region criteria have been based on the accessibility of health facilities and the total health budget provided to each province. Papua, Aceh and Maluku Province were excluded from the sampling frame due to political reforms and economic crisis. From the remaining 24 provinces, 10 provinces were sampled in 3 regions. The selection was carried out by the assistance of Central Bureau of Statistic (CBS) and used the PPS (Probability Proportionate to Size) technique. The following provinces were selected:

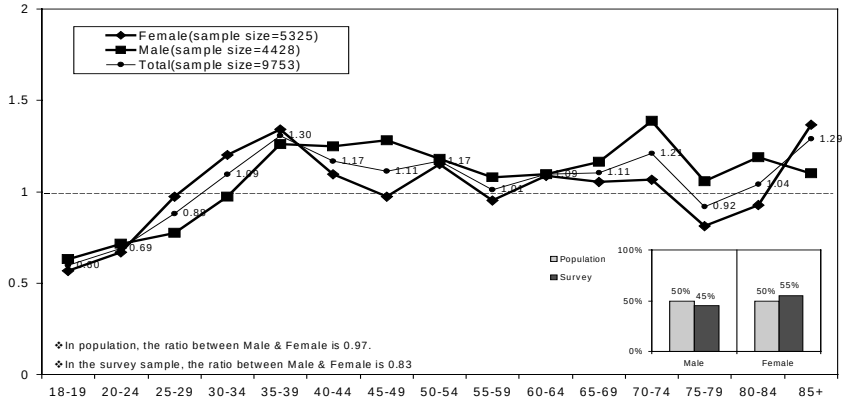
Region 1: DKI Jakarta, West Java, Central Java, East Java
Region 2: North Sumatra , South Sulawesi, South Sumatra
Region 3: West Nusa Tenggara, Central Kalimantan, South East Sulawesi

The total sample was 10,000. More females (54.8%) than males (45.2%) were interviewed.

Due to budget and time limitation (holidays between last week of November 2000 to second week of January 2001, such as Ramadhan, Christmas, New Year, Chinese New Year) the implementation phase was divided into two period of time, before the holidays and after the holidays. Additional problems experienced included: absence of respondent from home; length of questionnaire and culturally-sensitive questions, that wasted time (and even inhuman e.g. PLM) or that were difficult and misunderstood; bad political situation (e.g. ethnic conflict, president impeachment process, increase in crimes, etc.), which almost jeopardized the implementation of the household survey in several provinces; bad economic conditions, heavy rain, flood, and land slides; geographic inaccessibility; inadequate transportation and communication; survey instrument and budget delays from the WHO. In addition, Indonesia had server problems as their server depended on the Naval American Research Unit, which was adversely affected by the deterioration in its relations with the USA.

Survey summary characteristics

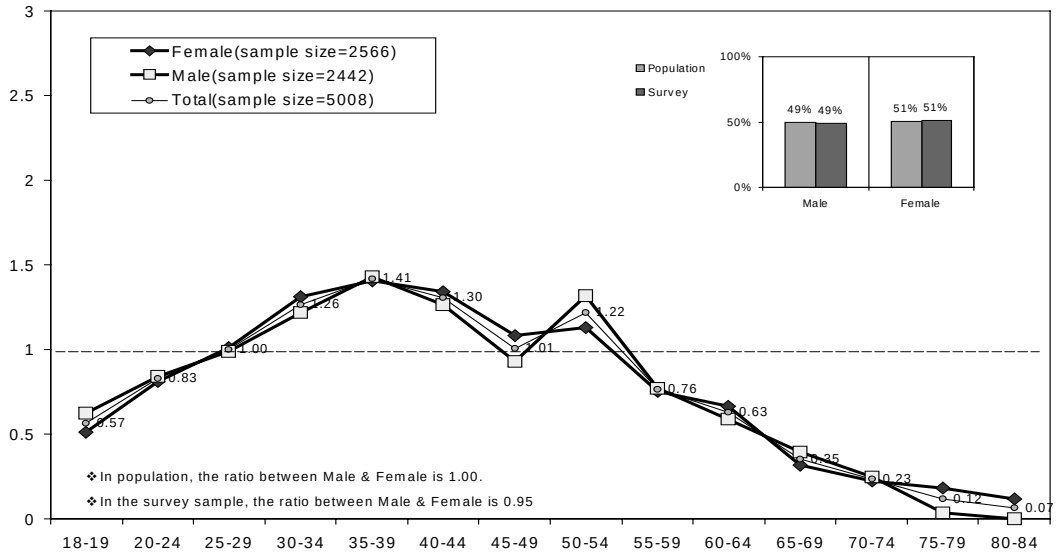
Final Sample size	% Response rate	% Missing data	Deviation from UN population Chi-square Critical Value = 49.58 (29 df, p=0.01)	% of items with Kappa > 0.4
9,995	99	6.4	309.8	98



Provinces of Indonesia have been selected according to level of development and the rural-urban distribution. The postal department maintains a list of zip codes in the country. Zip codes were selected across cities and villages in these provinces. Local post offices co-operated and picked 10 households from each of the zip code regions (approximately every 6th) and hand delivered the letters to these households. The postman wrote the name of the respondent on the letter while delivering and then collected them back after a few days.

Survey summary characteristics

Final Sample size	% Response rate	% Missing data	Deviation from UN population Chi-square Critical Value = 49.58 (29 df, p=0.01)
2,996	60	6.1	986.5



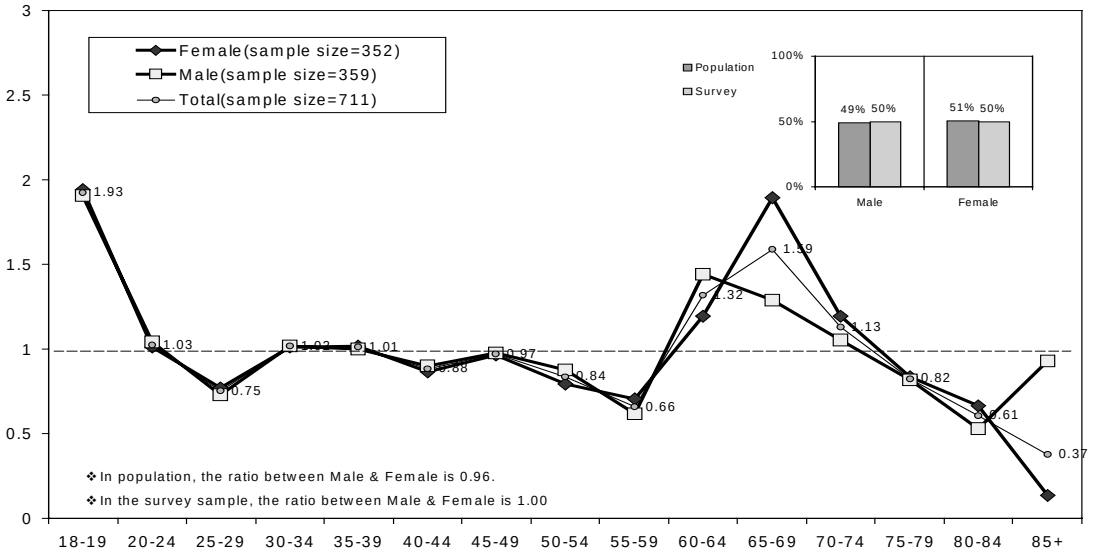
The metropolitan, urban and rural population and all "administrative regional units" as defined in Official Europe Union Statistics (NUTS 2) covered proportionately the respective population aged 18 and above. The country was divided into an appropriate number of areas, grouping NUTS regions at whatever level appropriately. The NUTS covered in Ireland were the following; Connaught/Ulster, Dublin, Munster, Rest of Leinster.

The basic sample design was a multi-stage, random probability sample. 30 sampling points were drawn with probability proportional to population size, for a total coverage of the country. The sampling points were drawn after stratification by NUTS 2 region and by degree of urbanisation. They represented the whole territory of the country surveyed and are selected proportionally to the distribution of the population in terms of metropolitan, urban and rural areas. In each of the selected sampling points, one address was drawn at random. This starting address forms the first address of a cluster of a maximum of 20 addresses. The remainder of the cluster was selected as every Nth address by standard random route procedure from the initial address. In theory, there is no maximum number of addresses issued per country. Procedures for random household selection and random respondent selection are independent of the interviewer's decision and controlled by the institute responsible. They should be as identical as possible from to country, full functional equivalence being a must.

At every address up to 4 recalls were made to attempt to achieve an interview with the selected respondent. There was only one interview per household. The final sample size is 711 completed interviews.

Survey summary characteristics

Final Sample size	% Response rate	% Missing data	Deviation from UN population Chi-square Critical Value = 49.58 (29 df, p=0.01)
711	39	0	69.8



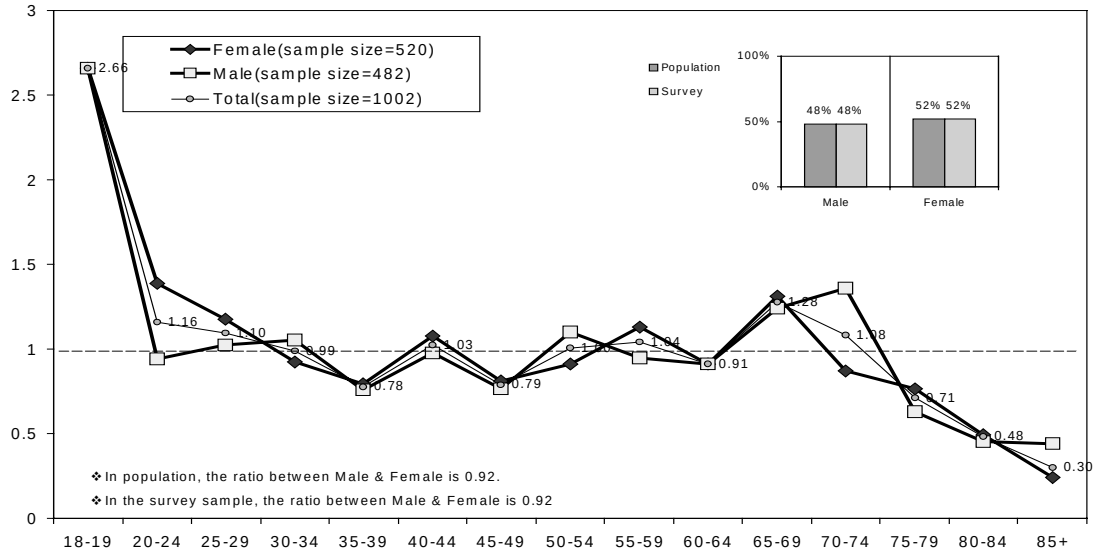
The metropolitan, urban and rural population and all "administrative regional units" as defined in Official Europe Union Statistics (NUTS 2) covered proportionately the respective population aged 18 and above. The country was divided into an appropriate number of areas, grouping NUTS regions at whatever level appropriately. The NUTS covered in Italy were the following; Basilicata, Calabria, Campania, Emilia, Friuli, Venezia, Giulia, Lazio, Liguria, Lombardia, Marche, Milano, Molise e Abbruzzi, Puglie, Sardegna, Sicilia, Toscana, Trentino, Umbria, Valle d'Aosta/Piemonte, Veneto.

The basic sample design was a multi-stage, random probability sample. 100 sampling points were drawn with probability proportional to population size, for a total coverage of the country. The sampling points were drawn after stratification by NUTS 2 region and by degree of urbanisation. They represented the whole territory of the country surveyed and are selected proportionally to the distribution of the population in terms of metropolitan, urban and rural areas. In each of the selected sampling points, one address was drawn at random. This starting address forms the first address of a cluster of a maximum of 20 addresses. The remainder of the cluster was selected as every Nth address by standard random route procedure from the initial address. In theory, there is no maximum number of addresses issued per country. Procedures for random household selection and random respondent selection are independent of the interviewer's decision and controlled by the institute responsible. They should be as identical as possible from to country, full functional equivalence being a must.

At every address up to 4 recalls were made to attempt to achieve an interview with the selected respondent. There was only one interview per household. The final sample size is 1,002 completed interviews.

Survey summary characteristics

Final Sample size	% Response rate	% Missing data	Deviation from UN population Chi-square Critical Value = 49.58 (29 df, p=0.01)
1,002	61	0.2	118



The sample was a multi-stage random probability sample representative of the population residing in urban and rural areas of Jordan.

An advanced sample design method in 2 stages was used:

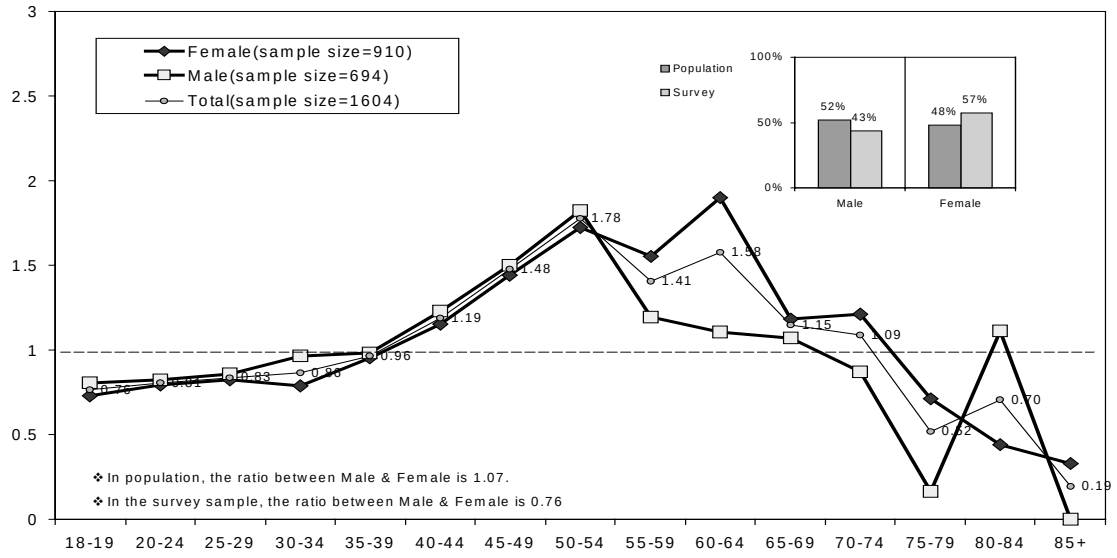
1. Jordan is administratively divided into 12 Governorates, each of which is subdivided into four regions. The survey was carried out in all four regions.
2. Selection of households within the Primary Sampling areas.

The sample structure was based on the estimated population structure elaborated on the basis of the data from the Jordan census of 1994. Statistical data acquired from the Block census had been used in the sample design of this study. The density of the population was classified into three categories: high, medium and low density areas.

The number of sampling units assigned for interviewing per Administrative Unit adequately represented the population density.

Survey summary characteristics

Final Sample size	% Response rate	% Missing data	Deviation from UN population Chi-square Critical Value = 49.58 (29 df, p=0.01)
1,604	74	0	149.4

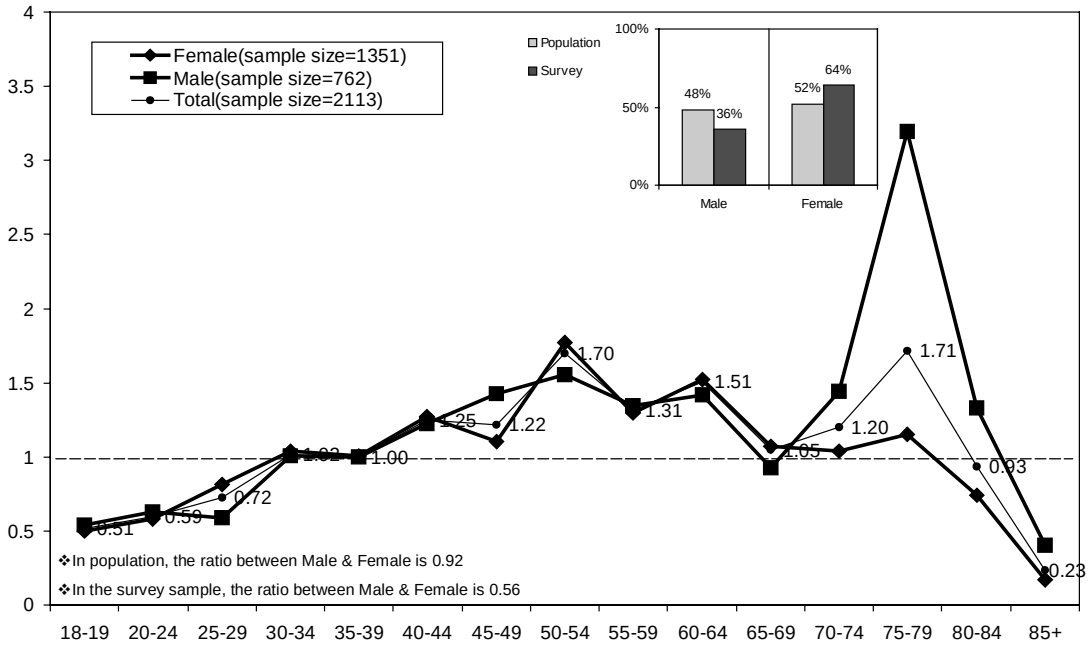


The Sample Survey Department of the National Statistic Committee of the Kyrgyz Republic randomly selected 5,000 households representative of the country's population, including all 7 regions and the capital, Bishkek, based on the National Census 1999. Since a sampling frame of individuals was not available households were selected.

The sampling represented proportionally the number of targeted population in the country including urban and rural areas. The geographical factor was taken into consideration as well. Within each region regional centre, districts, towns and villages were selected.

Survey summary characteristics

Final Sample size	% Response rate	% Missing data	Deviation from UN population Chi-square Critical Value = 49.58 (29 df, p=0.01)
2,209	44	13.2	273.1



The sample was a stratified multi-stage sample, representative of the totality of the inhabitants of Latvia aged 18 +.

At the first stage, sampling points were selected using a probability proportionate to size algorithm.

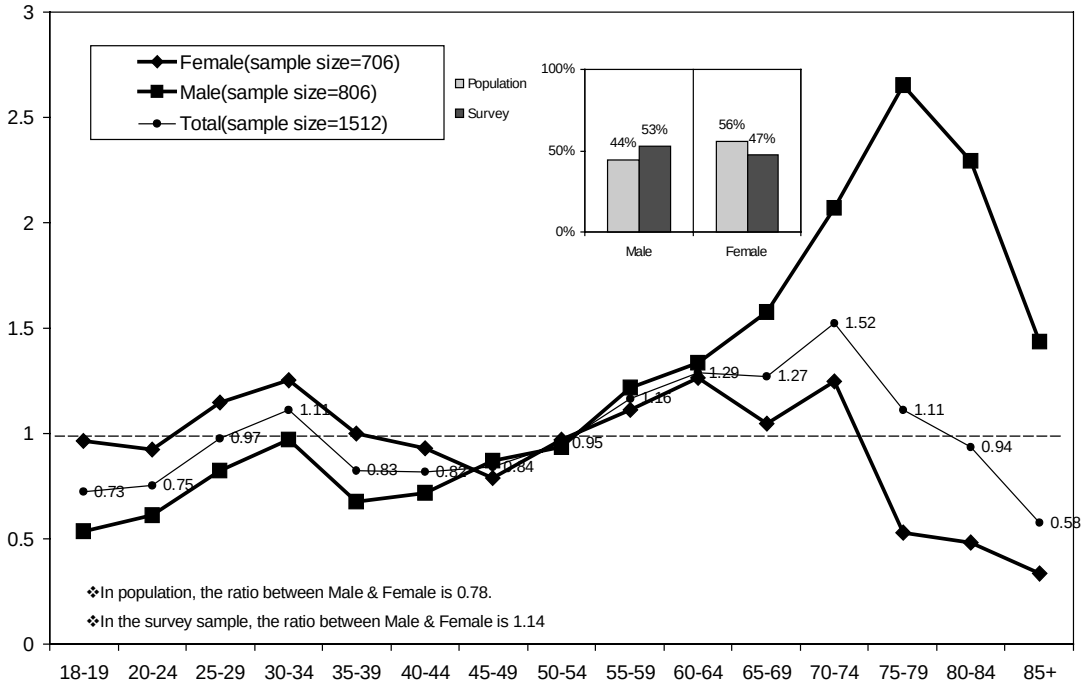
At the second stage, a random walk procedure was used. Starting addresses for the random walk were randomly selected from a register of inhabitants in the neighbourhood.

The interviewers started their work in accordance with a specified address moving forward by one side of the specified street (visiting in an ascending sequence addresses with even numbers and in a descending sequence addresses with odd numbers). Interviews were attempted in every fifth apartment or in every third private house.

The respondents within each household were chosen by applying the "latest birthday" principle.

Survey summary characteristics

Final Sample size	% Response rate	% Missing data	Deviation from UN population Chi-square Critical Value = 49.58 (29 df, p=0.01)
1,512	53	0	190.9



As there has been no census since 1930 and due to ongoing unrest, there are no available national sampling frames in Lebanon except a geographical frame by administrative districts and based on urban data such as blocks, buildings and apartments. Postal addresses remain incomplete with respect to both coverage and validity particularly for rural areas.

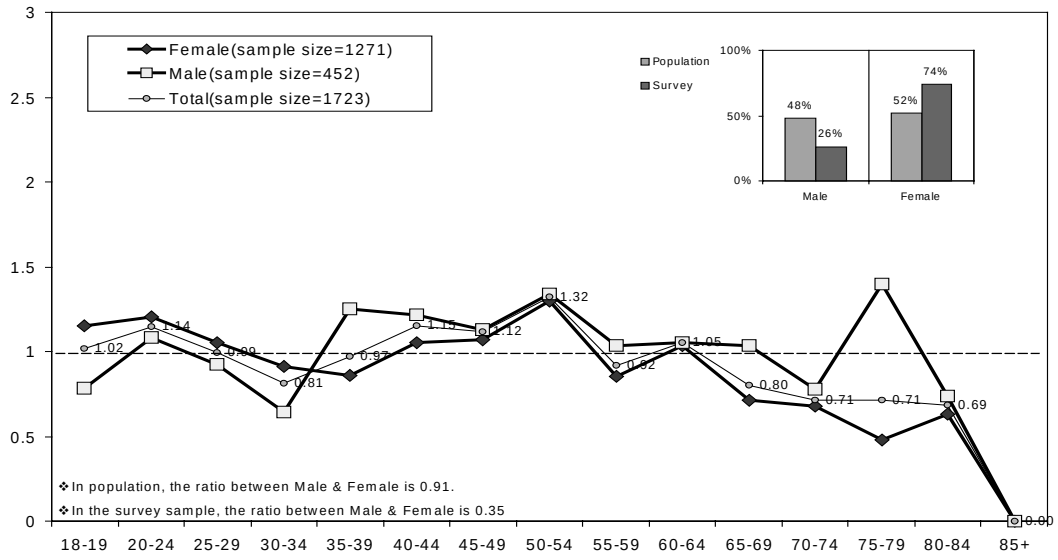
The sample selection was therefore based on local standards for nationwide samples which consist in a cluster sampling of district areas from which housing blocks were randomly selected. Households were then selected randomly from the blocks.

The individual selected from the household was 18+ years in age and the closest birthday method was used to select the respondent.

2,500 households were visited and from each one of them, two individuals were selected.

Survey summary characteristics

Final Sample size	% Response rate	% Missing data	Deviation from UN population Chi-square Critical Value = 49.58 (29 df, p=0.01)
2,224	44	1.4	66.4

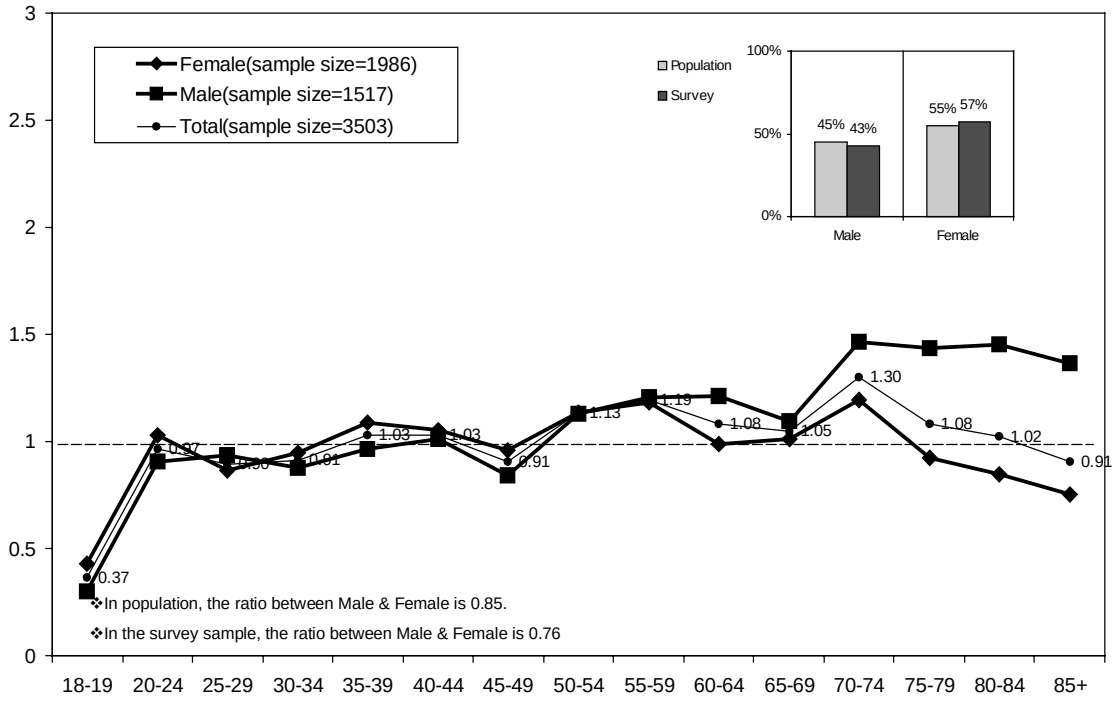


The random sample of 5000 respondents of the Lithuanian adult population, 18 years and older, was drawn by the National Statistical Department. The sampling frame was based on individual residents.

The frame was updated in 1989. The sample frame provides the following information about each respondent: name (first and family), address, data of birth and gender.

Survey summary characteristics

Final Sample size	% Response rate	% Missing data	Deviation from UN population Chi-square Critical Value = 49.58 (29 df, p=0.01)
3,498	70	4.6	120.9



The metropolitan, urban and rural population and all "administrative regional units" as defined in Official Europe Union Statistics (NUTS 2) covered proportionately the respective population aged 18 and above. The country was divided into an appropriate number of areas, grouping NUTS regions at whatever level appropriately.

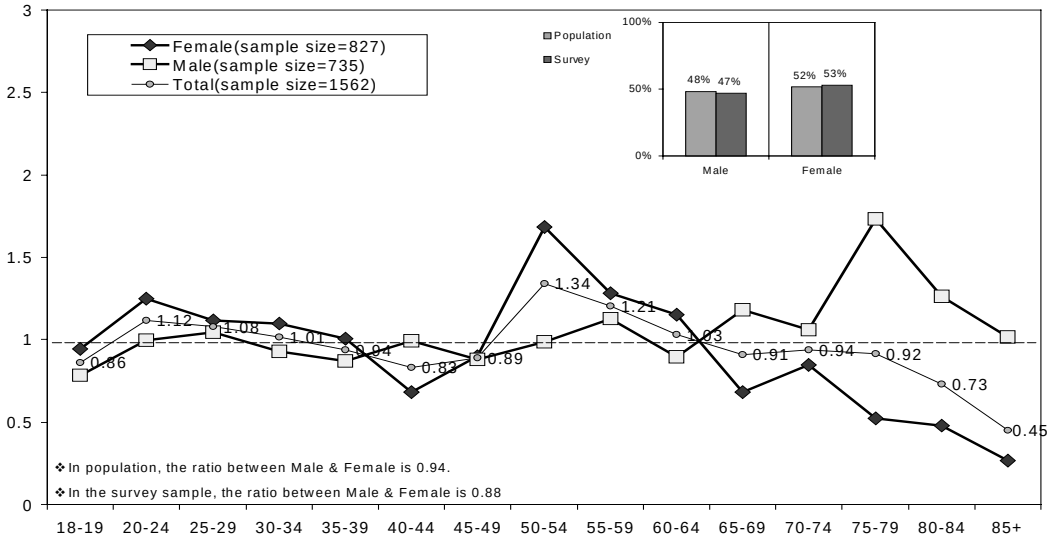
The NUTS covered in Luxembourg are the following; Centre, Est, Nord, Sud.

A CATI survey was carried out in this country using the random digit dialing method.

The final sample size is 719 completed interviews.

Survey summary characteristics

Final Sample size	% Response rate	% Missing data	Deviation from UN population Chi-square Critical Value = 49.58 (29 df, p=0.01)
719	55	1.6	85.2



The metropolitan, urban and rural population and all "administrative regional units" as defined in Official Europe Union Statistics (NUTS 2) covered proportionately the respective population aged 18 and above. The country was divided into an appropriate number of areas, grouping NUTS regions at whatever level appropriately. The NUTS covered in Malta were the following; Inner Harbour Region, Outer Harbour Region, South Eastern Region, Western Region, Northern Region, Gozo and Comino.

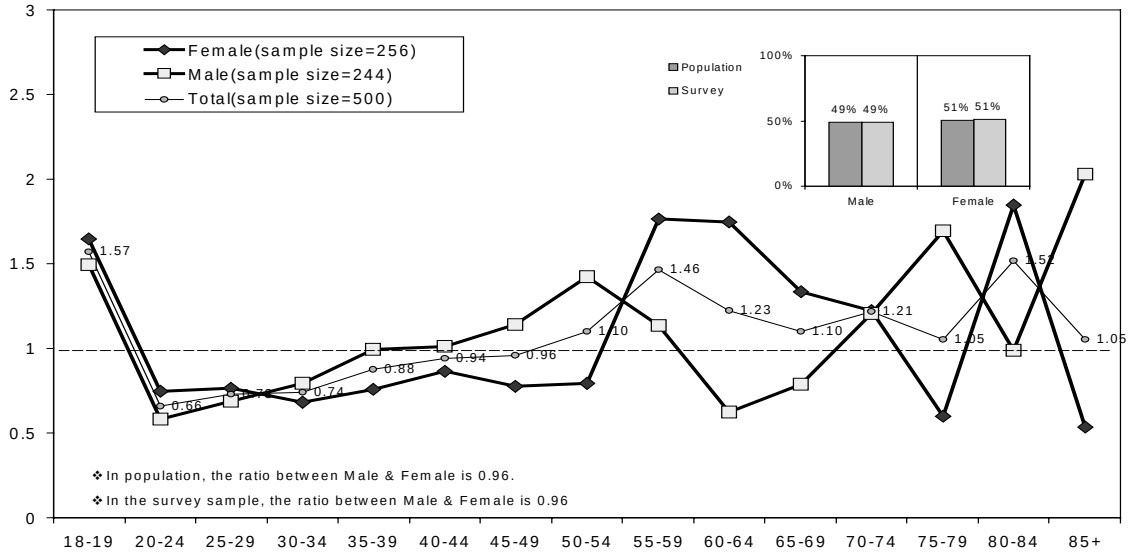
The basic sample design was a multi-stage, random probability sample. 50 sampling points were drawn with probability proportional to population size, for a total coverage of the country. The sampling points were drawn after stratification by NUTS 2 region and by degree of urbanisation. They represented the whole territory of the country surveyed and are selected proportionally to the distribution of the population in terms of metropolitan, urban and rural areas.

In each of the selected sampling points, one address was drawn at random. This starting address forms the first address of a cluster of a maximum of 20 addresses. The remainder of the cluster was selected as every Nth address by standard random route procedure from the initial address. In theory, there is no maximum number of addresses issued per country. Procedures for random household selection and random respondent selection are independent of the interviewer's decision and controlled by the institute responsible. They should be as identical as possible from to country, full functional equivalence being a must.

At every address up to 4 recalls are made to attempt to achieve an interview with the selected respondent. There was only one interview per household. The final sample size is 500 completed interviews.

Survey summary characteristics

Final Sample size	% Response rate	% Missing data	Deviation from UN population Chi-square Critical Value = 49.58 (29 df, p=0.01)
500	63	0.2	64.1

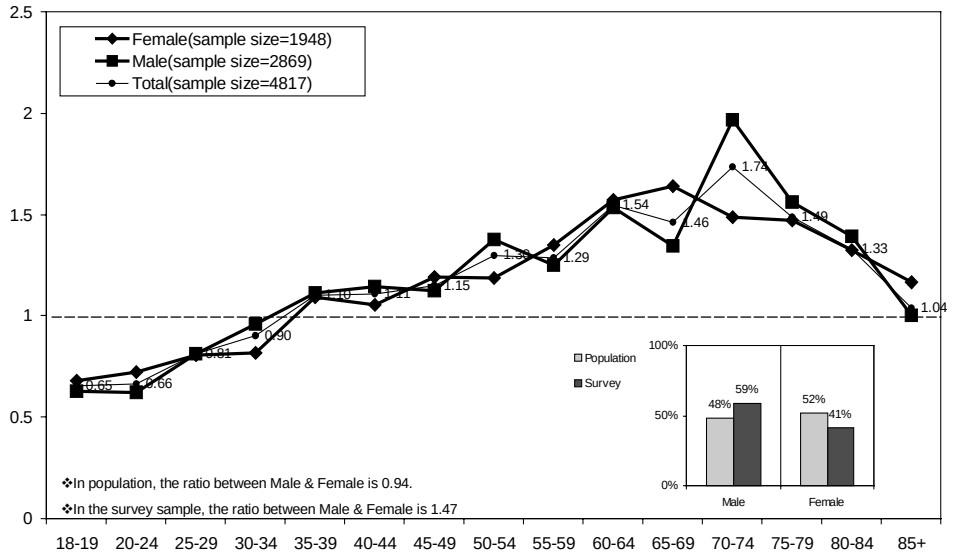


The sample used was a probabilistic, multistage, stratified and clustered sample and represented urban and rural strata. Mexico has 32 Federal States, which were divided, into 3 regions: Centre, South and North. Out of these regions, 15 were selected as follows:

Centre: Distrito Federal, Guanajuato, Jalisco, Estado de México, Michoacán, Qurétaro
South: Guerrero, Oaxaca, Puebla, Veracruz, Yucatán
North: Chihuahua, Nuevo León, San Luis Potosí, Sonora

Survey summary characteristics

Final Sample size	% Response rate	% Missing data	Deviation from UN population Chi-square Critical Value = 49.58 (29 df, p=0.01)	% of items with Kappa > 0.4
4,817	96	3.9	411.7	98



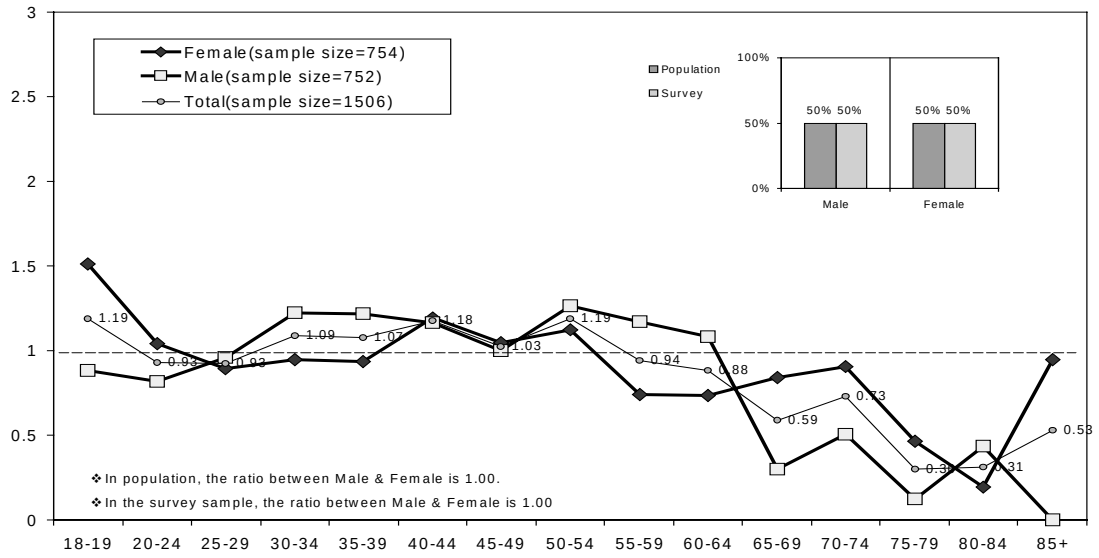
The sample was a multi-stage stratified random sampling described as follows. Twenty-two cities were selected following the main structure of Morocco composed of 5 strata and seven areas.

Once the stratum was defined, the city was chosen at random except for Casablanca and Rabat which were chosen intentionally due to their importance. At the second stage, a representative rural town from each area was chosen at random. The third stage consisted of selecting city quarters randomly. The fourth stage was the household selection according to the "step method" which meant approaching every third house.

In the case of a building, the interviewers were asked to begin from the top floor, to choose only one apartment on each floor and go downstairs every other floor with a maximum of two interviews per building. The last stage was the respondent's selection based on the Kish Method.

Survey summary characteristics

Final Sample size	% Response rate	% Missing data	Deviation from UN population Chi-square Critical Value = 49.58 (29 df, p=0.01)
1,506	69	0	72.2



The metropolitan, urban and rural population and all "administrative regional units" as defined in Official Europe Union Statistics (NUTS 2) covered proportionately the respective population aged 18 and above. The country was divided into an appropriate number of areas, grouping NUTS regions at whatever level appropriately.

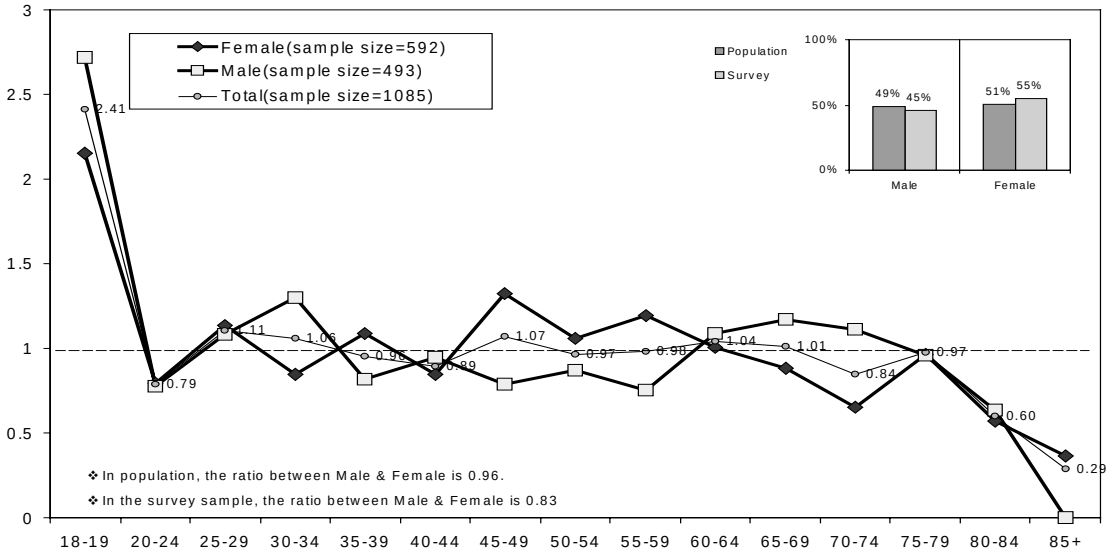
The NUTS covered in the Netherlands were the following; Drente, Flevoland, Friesland, Gelderland, Gröningen, Limburg, Noord-Brabant, Noord-Holland, Overijssel, Utrecht, Zeeland, Zuid-Holland.

The basic sample design was a multi-stage, random probability sample. 100 sampling points were drawn with probability proportional to population size, for a total coverage of the country. The sampling points were drawn after stratification by NUTS 2 region and by degree of urbanisation. They represented the whole territory of the country surveyed and are selected proportionally to the distribution of the population in terms of metropolitan, urban and rural areas. In each of the selected sampling points, one address was drawn at random. This starting address forms the first address of a cluster of a maximum of 20 addresses. The remainder of the cluster was selected as every Nth address by standard random route procedure from the initial address. In theory, there is no maximum number of addresses issued per country. Procedures for random household selection and random respondent selection are independent of the interviewer's decision and controlled by the institute responsible. They should be as identical as possible from to country, full functional equivalence being a must.

At every address up to 4 recalls were made to attempt to achieve an interview with the selected respondent. There was only one interview per household. The final sample size is 1,085 completed interviews.

Survey summary characteristics

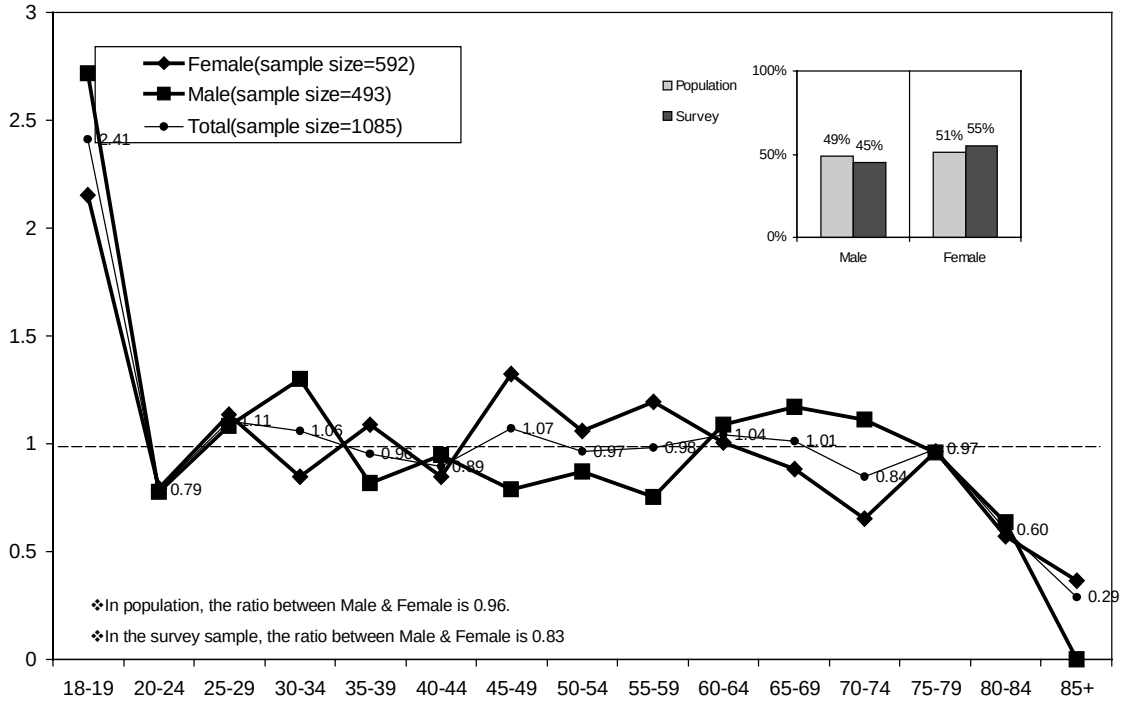
Final Sample size	% Response rate	% Missing data	Deviation from UN population Chi-square Critical Value = 49.58 (29 df, p=0.01)
1,085	59	0.4	115



The Municipal Population Registry (GBA) was used to select a representative sample of 3,000 individuals, aged 18 and over, of the Dutch population. Municipals were selected first and then the individual sample was drawn up.

Survey summary characteristics

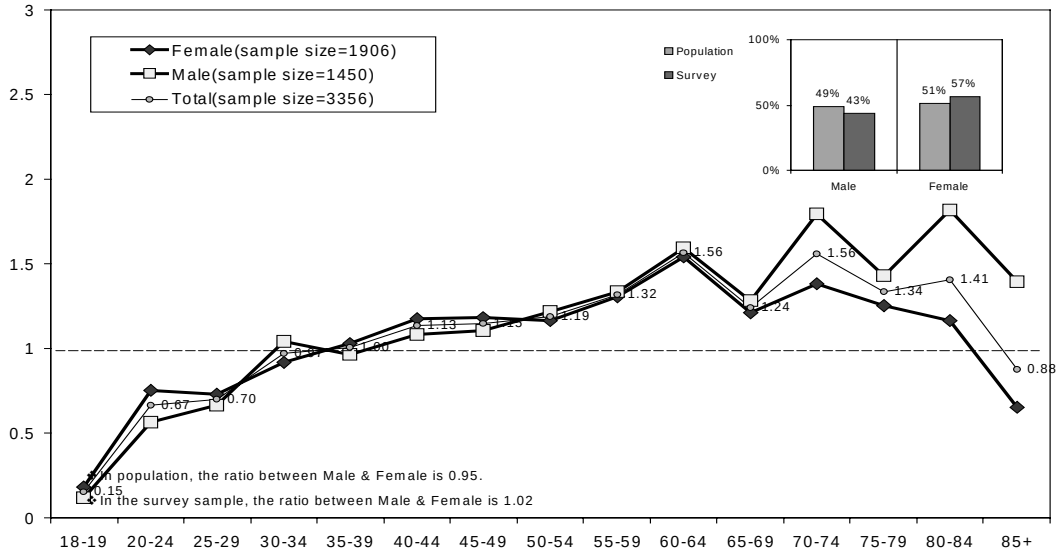
Final Sample size	% Response rate	% Missing data	Deviation from UN population Chi-square Critical Value = 49.58 (29 df, p=0.01)
1,211	40	5.0	275.2



Five thousand four hundred and fifty-five (5450) names were randomly selected from the New Zealand electoral rolls, where 90% of eligible voters are registered. In addition, an over sample of 549 was made from the Maori Rolls in order to ensure that the Maori were adequately represented in the national sample. This resulted in a total sample of 6004.

Survey summary characteristics

Final Sample size	% Response rate	% Missing data	Deviation from UN population Chi-square Critical Value = 49.58 (29 df, p=0.01)
3,401	68	5.5	474.2



The sample was not nationally representative and only covered the south western part of Nigeria (Oyo State). This region is composed of 33 local government areas (LGAs). Two rural, one semi-urban, and two urban LGAs were randomly selected. The pooled population of the selected LGAs was similar to the national population in terms of age and sex distribution, literacy rate, proportion unemployed, and in terms of the broad cadres among those employed.

The regions sampled were the following:

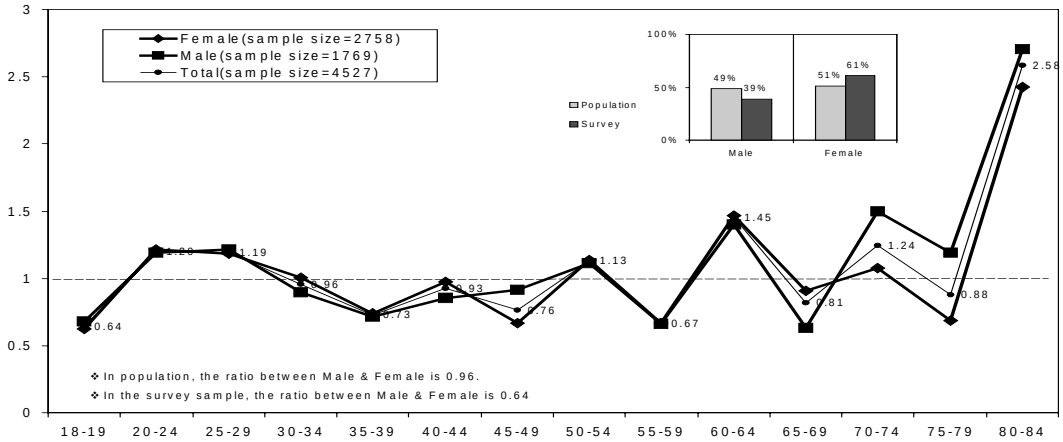
Region 1: Ibadan South West (urban)
Region 2: Ibadan North West (urban)
Region 3: Iseyin (semi-urban)
Region 4: Ido (rural)
Region 5: Ogo Oluwa (rural)

The total sample was 5,000. More females (61.4%) than males (38.6%) were interviewed.

Agreement to participate was excellent both the household level and at the respondent level even though, on many occasions, interviewers had to go to the farm or other places of work to interview selected respondents.

Survey summary characteristics

Final Sample size	% Response rate	% Missing data	Deviation from UN population Chi-square Critical Value = 49.58 (29 df, p=0.01)	% of items with Kappa > 0.4
5,047	98	11.5	28.8	45



The sample was a multi-stage random probability sample representative of the population residing in urban and rural areas of Oman. The sample structure was based on the estimated population structure elaborated on the basis of the data from the Oman census of 1993. The estimates for various categories which were excluded from research were compiled using statistical information from the Oman census. The categories that were excluded included:

- Domestic helpers (e.g. servants, maids, household helpers) and building watchmen, which tend to be Yemenis, Egyptians, Sudanese, Pakistanis, and Indians;
- And, blue collar service staff (e.g. technicians on sites, hospital support staff, etc.) and others who are difficult to reach due to the shift-work nature of their occupation and residence in specially designated centres.

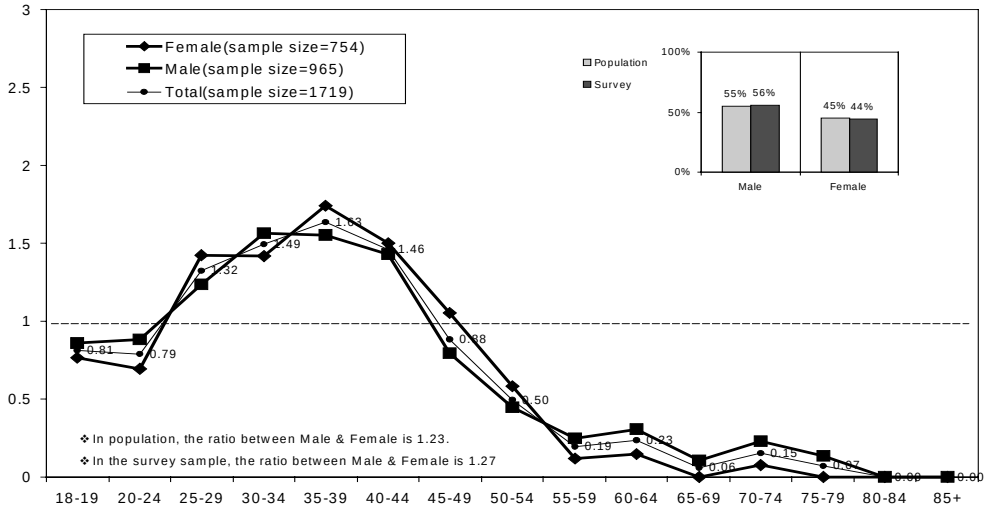
In Oman, an advanced sample design method in 2 stages was used:

1. Three main regions from Oman's 8 Regions / Governorates were selected.
2. Households were randomly selected within the Primary Sampling areas.

Within the main urban centres, the sampling system employed was that each urban centre was divided into Administrative Units, which in turn were further subdivided into clusters (consisting of an agglomeration within a determined polygon of roads and streets). Each cluster has a certain number of blocks, which are defined as the smallest tract of land outlined by streets or roads that contain houses and buildings. In each block, buildings and houses were identified and counted. The selection of respondent was done using the Kish Method.

Survey summary characteristics

Final Sample size	% Response rate	% Missing data	Deviation from UN population Chi-square Critical Value = 49.58 (29 df, p=0.01)
1,719	71	0	437.9



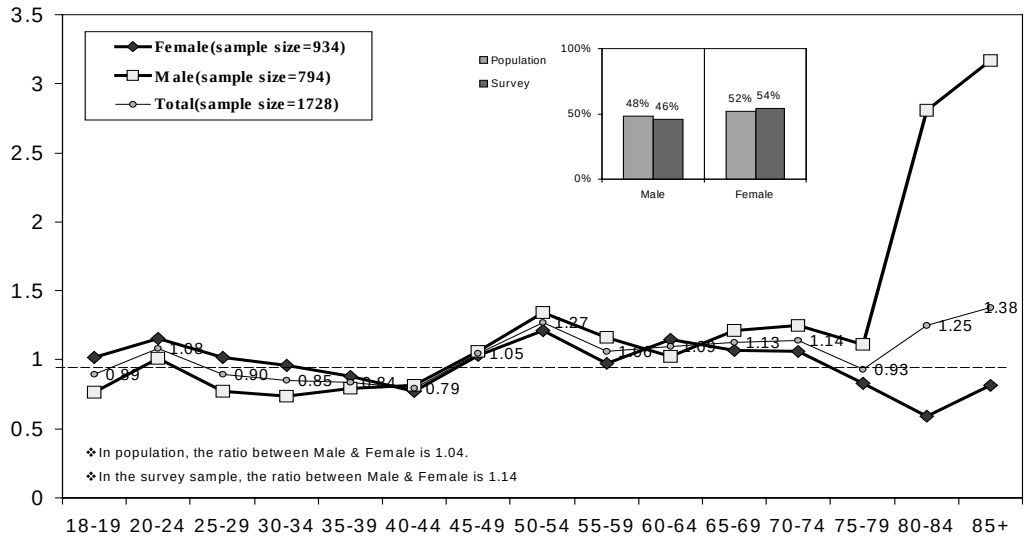
In first step of sampling, the places from where the respondents were to be selected were chosen.

The sample included 5 strata (categories): rural, urban up to 20 000 inhabitants, urban from 20 000 to 100 000 inhabitants, urban from 101 000 to 500 000 inhabitants and urban more than 501 000 inhabitants.

Respondents were randomly selected from the PESEL (General Electronic Registry of Population - Governmental Information Centre). Each adult Polish resident has his own number in the PESEL.

Survey summary characteristics

Final Sample size	% Response rate	% Missing data	Deviation from UN population Chi-square Critical Value = 49.58 (29 df, p=0.01)
1,707	34	4.7	370.9



The metropolitan, urban and rural population and all "administrative regional units" as defined in Official Europe Union Statistics (NUTS 2) covered proportionately the respective population aged 18 and above. The country was divided into an appropriate number of areas, grouping NUTS regions at whatever level appropriately. The NUTS covered in Portugal were the following; Alentejo, Algarve, Azores, Centro, Lisboa e Vale do Tejo, Madeira, Norte.

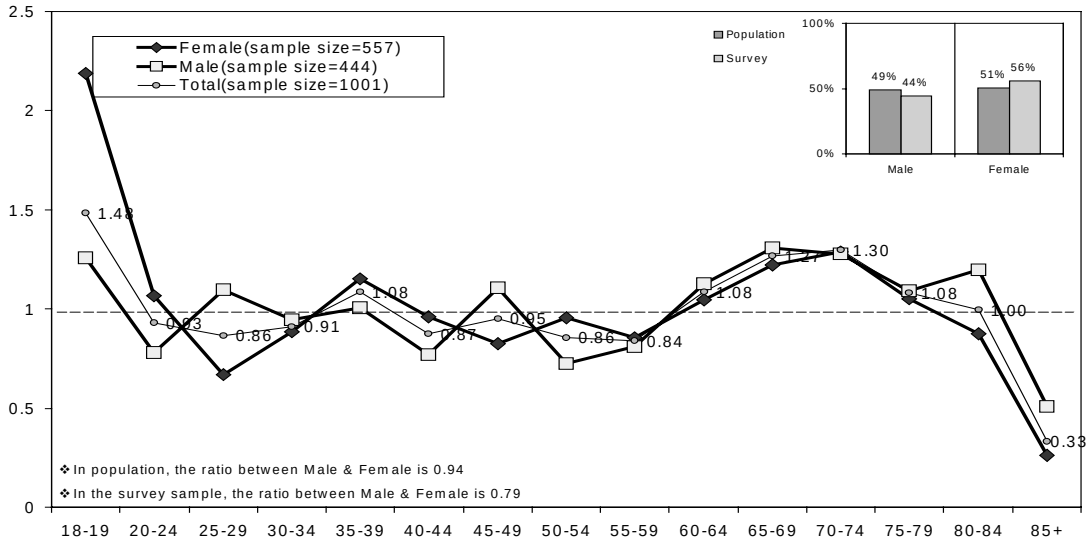
The basic sample design was a multi-stage, random probability sample. 100 sampling points were drawn with probability proportional to population size, for a total coverage of the country. The sampling points were drawn after stratification by NUTS 2 region and by degree of urbanisation. They represented the whole territory of the country surveyed and are selected proportionally to the distribution of the population in terms of metropolitan, urban and rural areas.

In each of the selected sampling points, one address was drawn at random. This starting address forms the first address of a cluster of a maximum of 20 addresses. The remainder of the cluster was selected as every Nth address by standard random route procedure from the initial address. In theory, there is no maximum number of addresses issued per country. Procedures for random household selection and random respondent selection are independent of the interviewer's decision and controlled by the institute responsible. They should be as identical as possible from to country, full functional equivalence being a must.

At every address up to 4 recalls were made to attempt to achieve an interview with the selected respondent. There was only one interview per household. The final sample size is 1,001 completed interviews.

Survey summary characteristics

Final Sample size	% Response rate	% Missing data	Deviation from UN population Chi-square Critical Value = 49.58 (29 df, p=0.01)
1,001	61	0	62.5

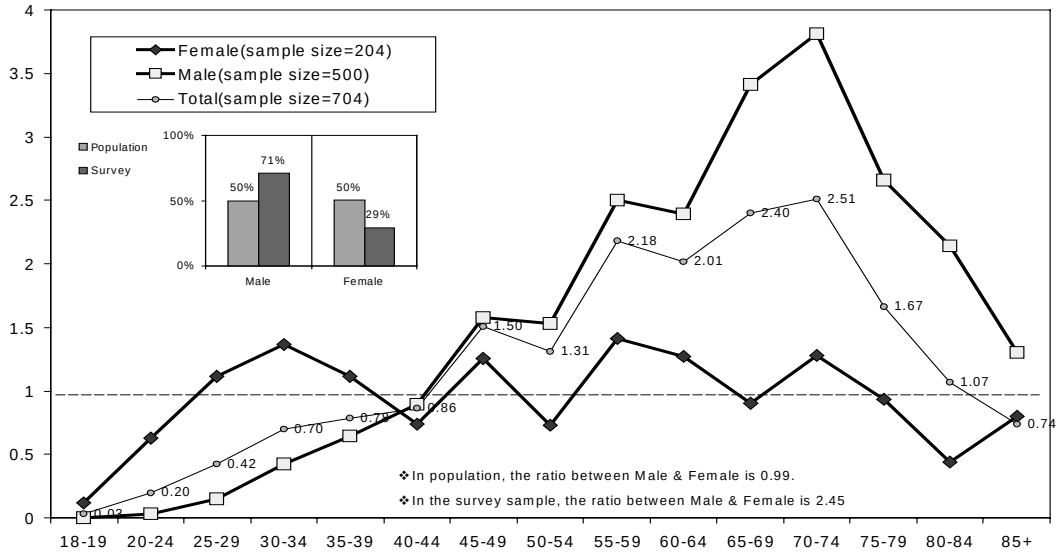


3,000 named individuals were randomly selected from the Telephone Directory published in December 1999. The latter is stratified by local authority.

Addresses were selected from the 171 Telephone Directory Registries.

Survey summary characteristics

Final Sample size	% Response rate	% Missing data	Deviation from UN population Chi-square Critical Value = 49.58 (29 df, p=0.01)
705	24	0	506.4



The metropolitan, urban and rural population and all "administrative regional units" as defined in Official Europe Union Statistics (NUTS 2) covered proportionately the respective population aged 18 and above. The country was divided into an appropriate number of areas, grouping NUTS regions at whatever level appropriately. The NUTS covered in Romania were the following; Nord-Est, Sud-Est, Sud, Sud-Vest, Vest, Nord-Vest, Centru, Bucuresti.

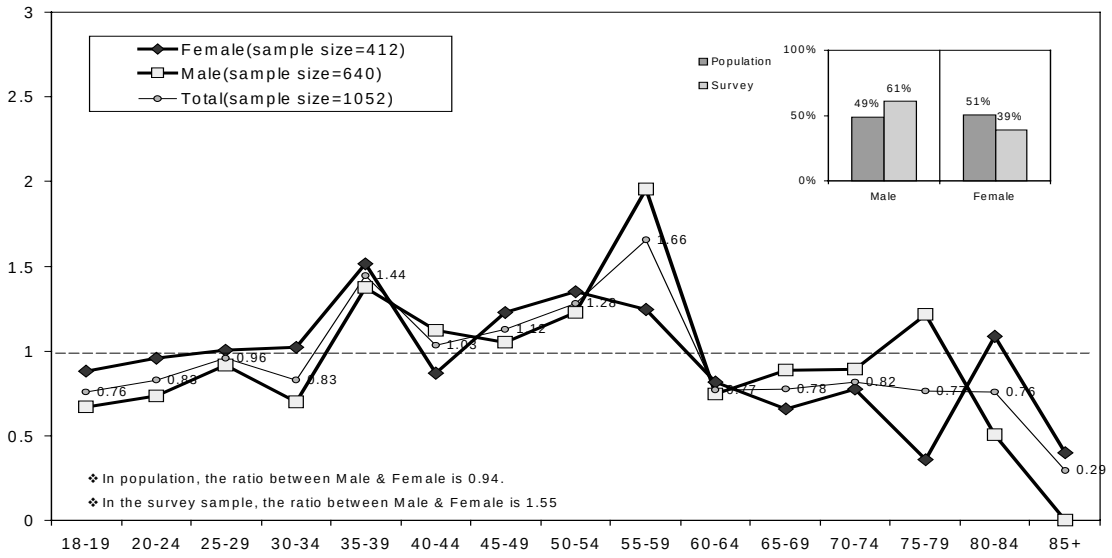
The basic sample design was a multi-stage, random probability sample. 100 sampling points were drawn with probability proportional to population size, for a total coverage of the country. The sampling points were drawn after stratification by NUTS 2 region and by degree of urbanisation. They represented the whole territory of the country surveyed and are selected proportionally to the distribution of the population in terms of metropolitan, urban and rural areas.

In each of the selected sampling points, one address was drawn at random. This starting address forms the first address of a cluster of a maximum of 20 addresses. The remainder of the cluster was selected as every Nth address by standard random route procedure from the initial address. In theory, there is no maximum number of addresses issued per country. Procedures for random household selection and random respondent selection are independent of the interviewer's decision and controlled by the institute responsible. They should be as identical as possible from to country, full functional equivalence being a must.

At every address up to 4 recalls were made to attempt to achieve an interview with the selected respondent. There was only one interview per household. The final sample size is 1,053 completed interviews.

Survey summary characteristics

Final Sample size	% Response rate	% Missing data	Deviation from UN population Chi-square Critical Value = 49.58 (29 df, p=0.01)
1,053	52	1.0	99.5



The metropolitan, urban and rural population and all "administrative regional units" as defined in Official Europe Union Statistics (NUTS 2) covered proportionately the respective population aged 18 and above. The country was divided into an appropriate number of areas, grouping NUTS regions at whatever level appropriately.

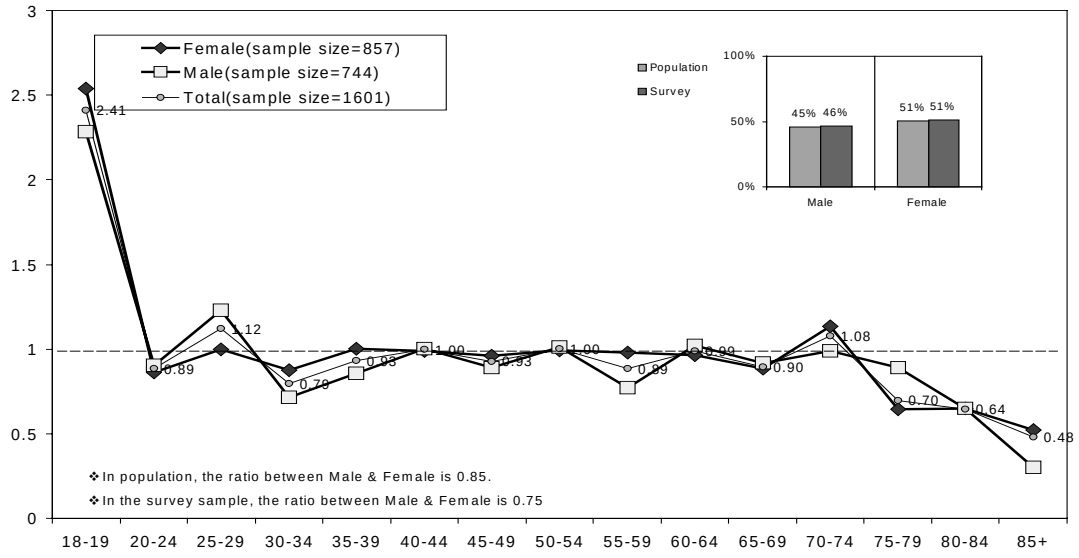
The NUTS covered in the Russian Federation were the following; Severny, Severo-Zapad, Central, Volgo-Vatsky, Centralno-Chernoz, Povolsky, Severo-Kavkaz, Uralsky, Zapadno-Sibir, Vostochno-Sibir, Dalnevost.

The basic sample design was a multi-stage, random probability sample. 100 sampling points were drawn with probability proportional to population size, for a total coverage of the country. The sampling points were drawn after stratification by NUTS 2 region and by degree of urbanisation. They represented the whole territory of the country surveyed and are selected proportionally to the distribution of the population in terms of metropolitan, urban and rural areas. In each of the selected sampling points, one address was drawn at random. This starting address forms the first address of a cluster of a maximum of 20 addresses. The remainder of the cluster was selected as every Nth address by standard random route procedure from the initial address. In theory, there is no maximum number of addresses issued per country. Procedures for random household selection and random respondent selection are independent of the interviewer's decision and controlled by the institute responsible. They should be as identical as possible from to country, full functional equivalence being a must.

At every address up to 4 recalls were made to attempt to achieve an interview with the selected respondent. There was only one interview per household. The final sample size is 1,601 completed interviews.

Survey summary characteristics

Final Sample size	% Response rate	% Missing data	Deviation from UN population Chi-square Critical Value = 49.58 (29 df, p=0.01)
1,601	25	0.1	166.8



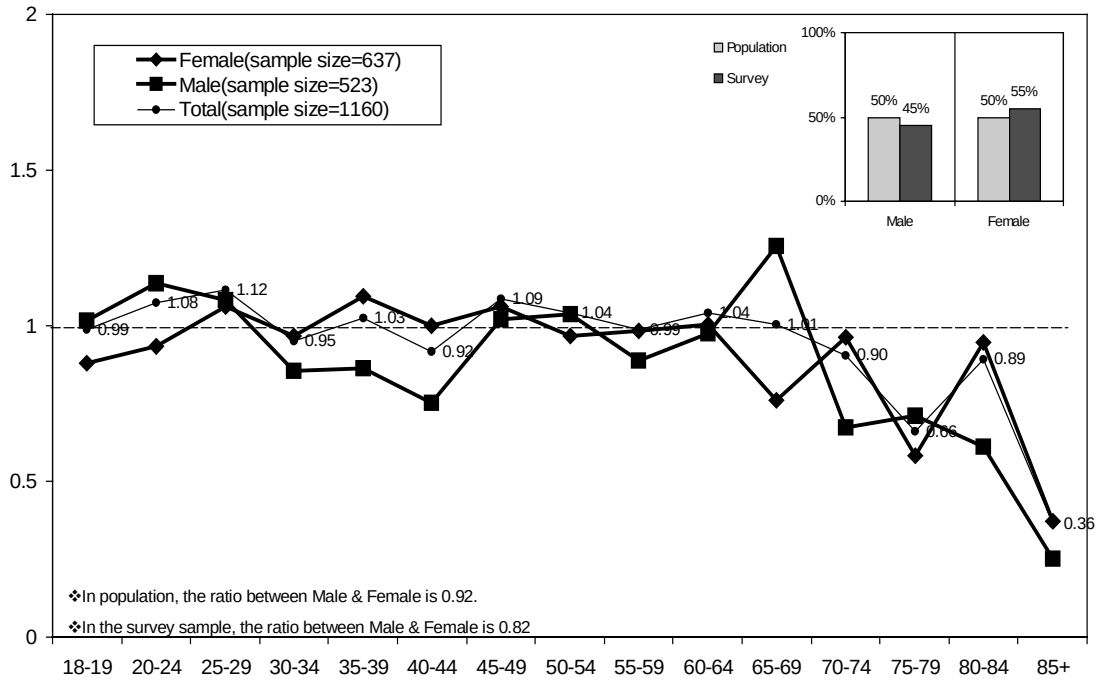
A sample of 1,200 was taken and respondents were selected from the population register. Eight regions were sampled:

Region 1: Bratislava
Region 2: Banska Bystrica
Region 3: Košice
Region 4: Nitra
Region 5: Prešov
Region 6: Trenčín
Region 7: Trnava
Region 8: Žilina

More females (54.9%) than males (45.1%) were interviewed.

Survey summary characteristics

Final Sample size	% Response rate	% Missing data	Deviation from UN population Chi-square Critical Value = 49.58 (29 df, p=0.01)	% of items with Kappa > 0.4
1,183	84	16.5	717.5	97



The metropolitan, urban and rural population and all "administrative regional units" as defined in Official Europe Union Statistics (NUTS 2) covered proportionately the respective population aged 18 and above. The country was divided into an appropriate number of areas, grouping NUTS regions at whatever level appropriately. The NUTS covered in Spain were the following; Andalucia, Aragón, Asturias, Baleares, Canarias, Cantabria, Castilla-La Mancha, Castilla-León, Cataluña, Extremadura, Galicia, La Rioja, Madrid, Murcia, Navarra, País Vasco, País Valenciano.

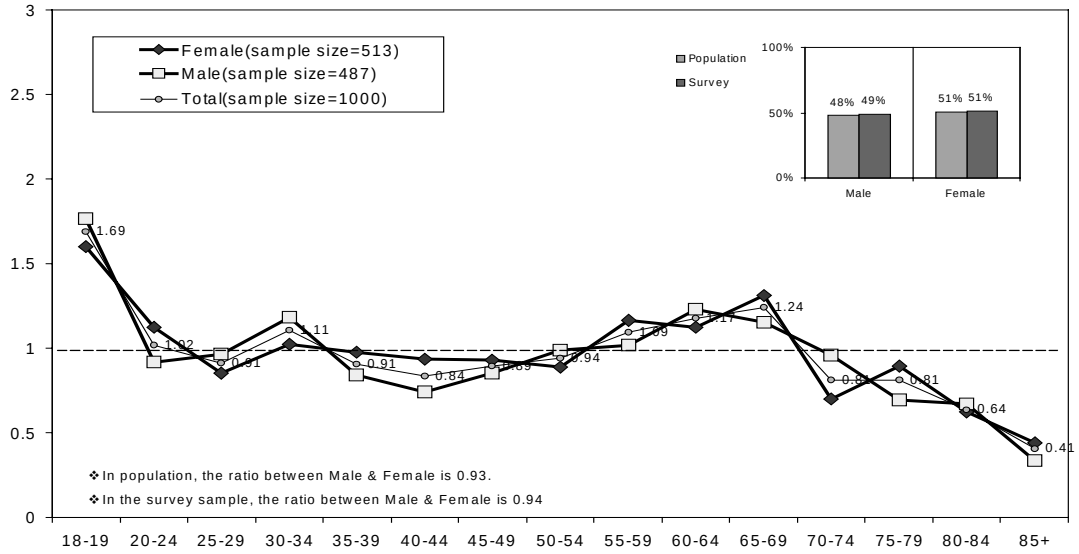
The basic sample design was a multi-stage, random probability sample. 100 sampling points were drawn with probability proportional to population size, for a total coverage of the country. The sampling points were drawn after stratification by NUTS 2 region and by degree of urbanisation. They represented the whole territory of the country surveyed and are selected proportionally to the distribution of the population in terms of metropolitan, urban and rural areas.

In each of the selected sampling points, one address was drawn at random. This starting address forms the first address of a cluster of a maximum of 20 addresses. The remainder of the cluster was selected as every Nth address by standard random route procedure from the initial address. In theory, there is no maximum number of addresses issued per country. Procedures for random household selection and random respondent selection are independent of the interviewer's decision and controlled by the institute responsible. They should be as identical as possible from to country, full functional equivalence being a must.

At every address up to 4 recalls were made to attempt to achieve an interview with the selected respondent. There was only one interview per household. The final sample size is 1,000 completed interviews.

Survey summary characteristics

Final Sample size	% Response rate	% Missing data	Deviation from UN population Chi-square Critical Value = 49.58 (29 df, p=0.01)
1,000	75	1.6	59.4



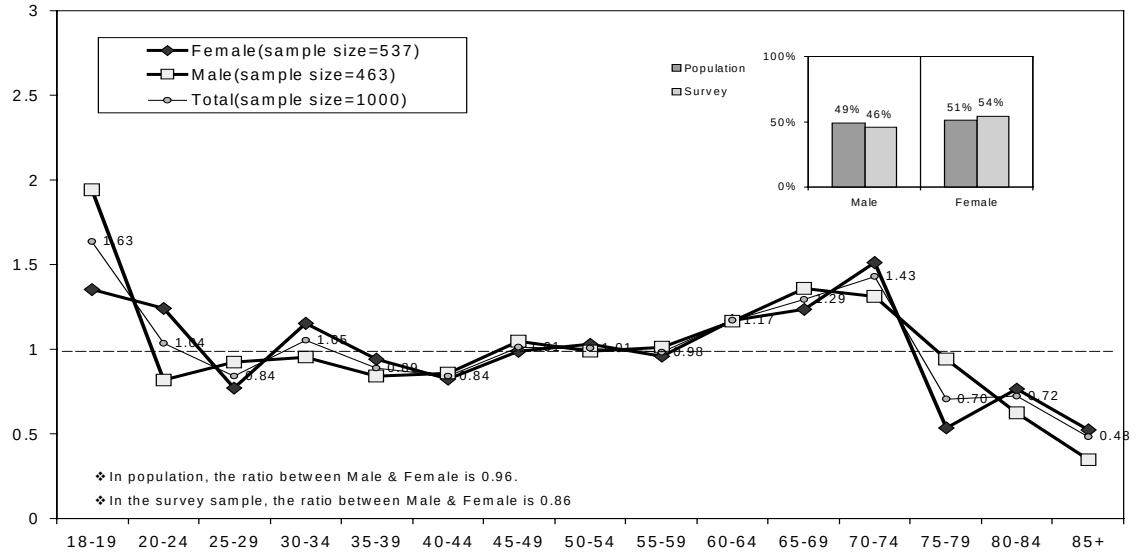
The metropolitan, urban and rural population and all "administrative regional units" as defined in Official Europe Union Statistics (NUTS 2) covered proportionately the respective population aged 18 and above. The country was divided into an appropriate number of areas, grouping NUTS regions at whatever level appropriately. The NUTS covered in Sweden were the following; Stockholm/Södertäjle A-Region, Gothenburgs A-Region, Malmö/Lund/Trelleborgs A-region, Semi urban area, Rural area.

The basic sample design was a multi-stage, random probability sample. 100 sampling points were drawn with probability proportional to population size, for a total coverage of the country. The sampling points were drawn after stratification by NUTS 2 region and by degree of urbanisation. They represented the whole territory of the country surveyed and are selected proportionally to the distribution of the population in terms of metropolitan, urban and rural areas. In each of the selected sampling points, one address was drawn at random. This starting address forms the first address of a cluster of a maximum of 20 addresses. The remainder of the cluster was selected as every Nth address by standard random route procedure from the initial address. In theory, there is no maximum number of addresses issued per country. Procedures for random household selection and random respondent selection are independent of the interviewer's decision and controlled by the institute responsible. They should be as identical as possible from to country, full functional equivalence being a must.

At every address up to 4 recalls were made to attempt to achieve an interview with the selected respondent. There was only one interview per household. The final sample size is 1,000 completed interviews.

Survey summary characteristics

Final Sample size	% Response rate	% Missing data	Deviation from UN population Chi-square Critical Value = 49.58 (29 df, p=0.01)
1,000	53	0.1	59.6



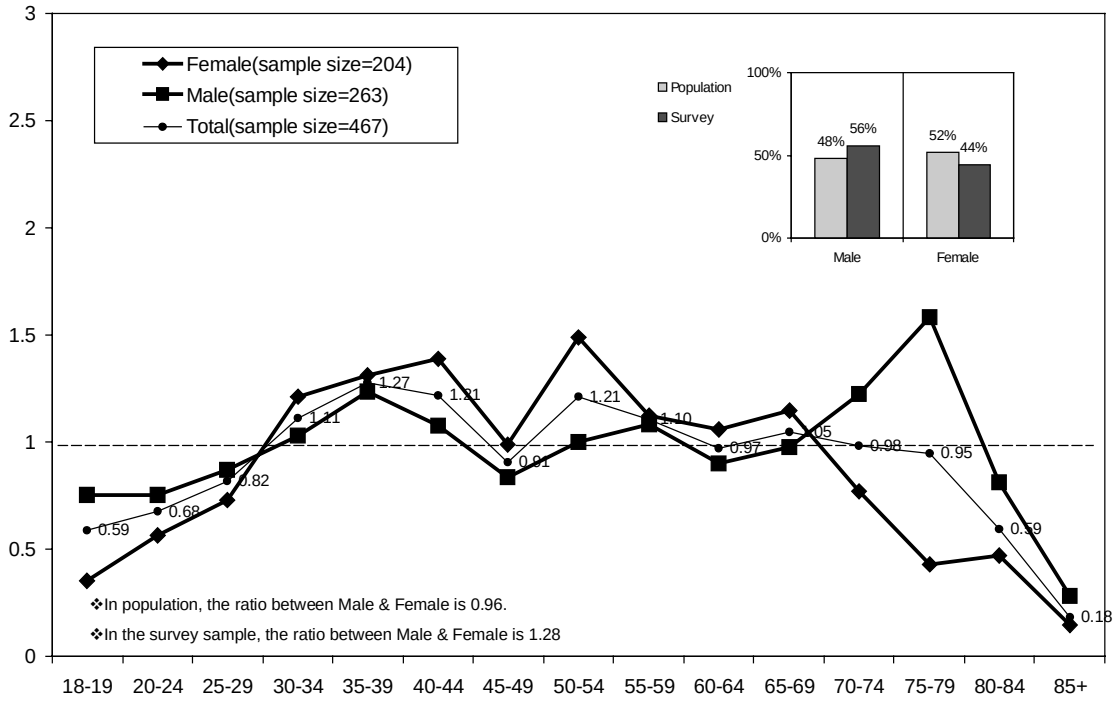
The sampling frame used was the electronic telephone directory, available from Swisscom for the Federal Office of Statistics as well as the large social research agencies. The directory contains all registered land-line and mobile telephone numbers.

2,500 randomly selected households were contacted by telephone. The size of the household was recorded and a list of the members of the household noted (with first name, age and gender).

The target respondent, aged 18 years and over, was randomly selected by the computer during the telephone interview.

Survey summary characteristics

Final Sample size	% Response rate	% Missing data	Deviation from UN population Chi-square Critical Value = 49.58 (29 df, p=0.01)
962	38	9.7	63.4



A multistage sampling procedure was used to sample the nationally representative households. Thailand is divided into 76 provinces, which are divided into approximately 700 districts.

The district was the first stage unit of selection in the rural areas. The second stage of selection was the villages and the third stage was the household.

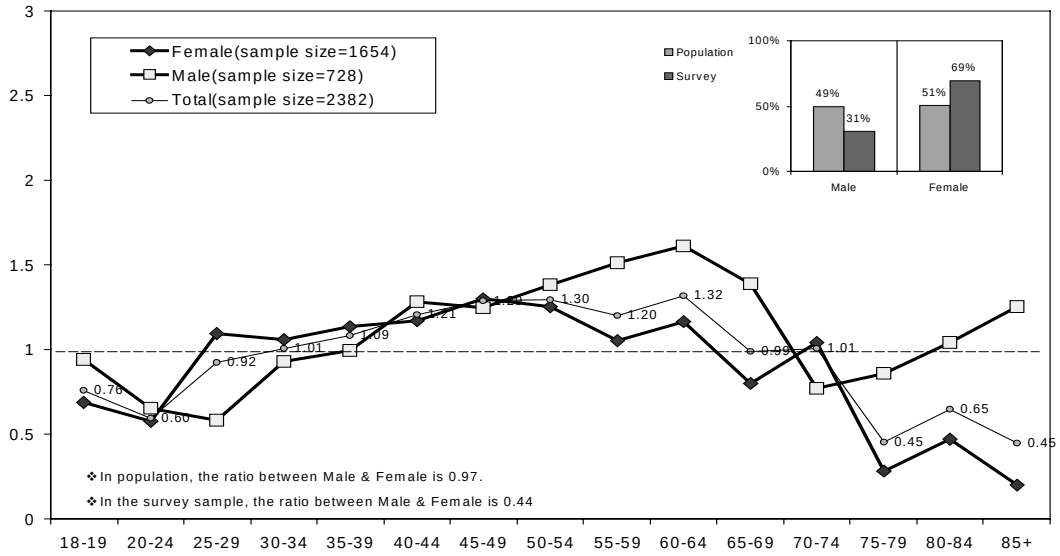
The first and second stages of selection for rural areas and the first stage of the urban areas were based on the probability of selection being proportional to the population size (PPS) of the area.

Individuals were randomly selected from the list of eligible voters for each sampled unit. The department of Local Administration, Ministry of Interior, compiles this list which contains names and addresses of eligible voters that are 18 years and older.

The sample included 500 households in Bangkok, 1,000 provincial urban households, and 3,500 rural households.

Survey summary characteristics

Final Sample size	% Response rate	% Missing data	Deviation from UN population Chi-square Critical Value = 49.58 (29 df, p=0.01)
2,288	46	0	179.4

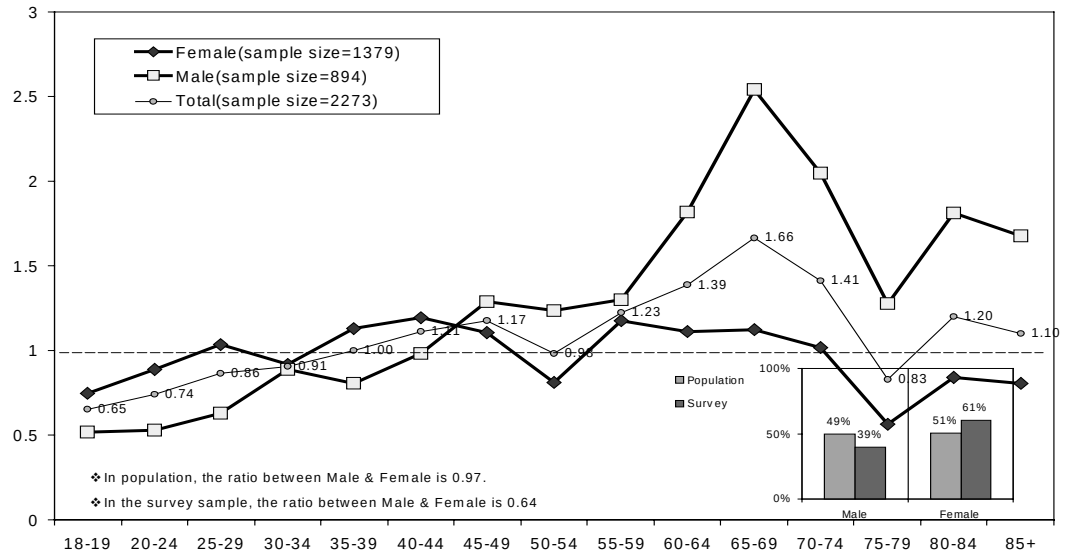


The sample was selected from a national sample frame that is used by the Central Statistical Office for the purpose of its Continuous Sample Survey of the Population (CSSP). The CSSP has been used for generating samples related to surveys that have addressed topics such as housing, family life, youth, crime, living conditions, household consumption and health.

Using a two stage sampling design, enumeration districts were selected in each of the 14 administrative domains of the country and a total of 10 households were selected in each enumeration district. Altogether, 5,520 households were selected nationwide in 552 enumeration districts. In each enumeration district, the 10 households constituted a compact cluster in which the first household was selected randomly.

Survey summary characteristics

Final Sample size	% Response rate	% Missing data	Deviation from UN population Chi-square Critical Value = 49.58 (29 df, p=0.01)
2,582	52	22.3	231.7



The sample was a nationally representative quota sampling of 5,000 respondents. The country was divided into strata provided by the State Planning Organization (SPO). The selection of sampling units was done by demographic variables such as SES, gender, and dwelling.

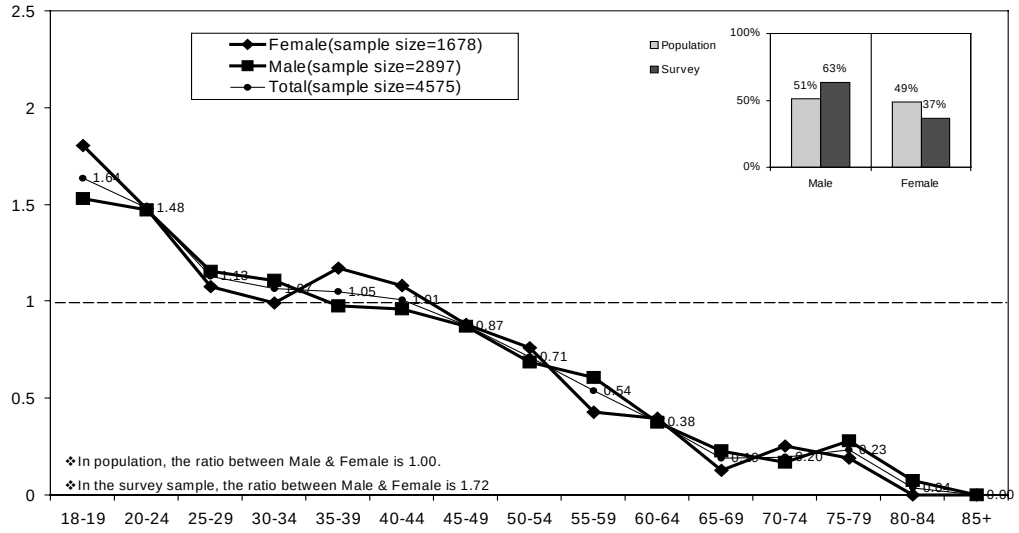
The sampling frame of the survey corresponded to the index of development of the cities in five strata of SPO; Istanbul, Antalya, Manisa, Trabzon, Yozgat, Adiyaman.

The sampling frame considered gender, dwellings and socioeconomic status. All respondents were identified in terms of socioeconomic status, phone numbers and addresses

More males (57.2%) than females (42.8%) were interviewed.

Survey summary characteristics

Final Sample size	% Response rate	% Missing data	Deviation from UN population Chi-square Critical Value = 49.58 (29 df, p=0.01)	% of items with Kappa > 0.4
5,194	90	22.6	408.6	93

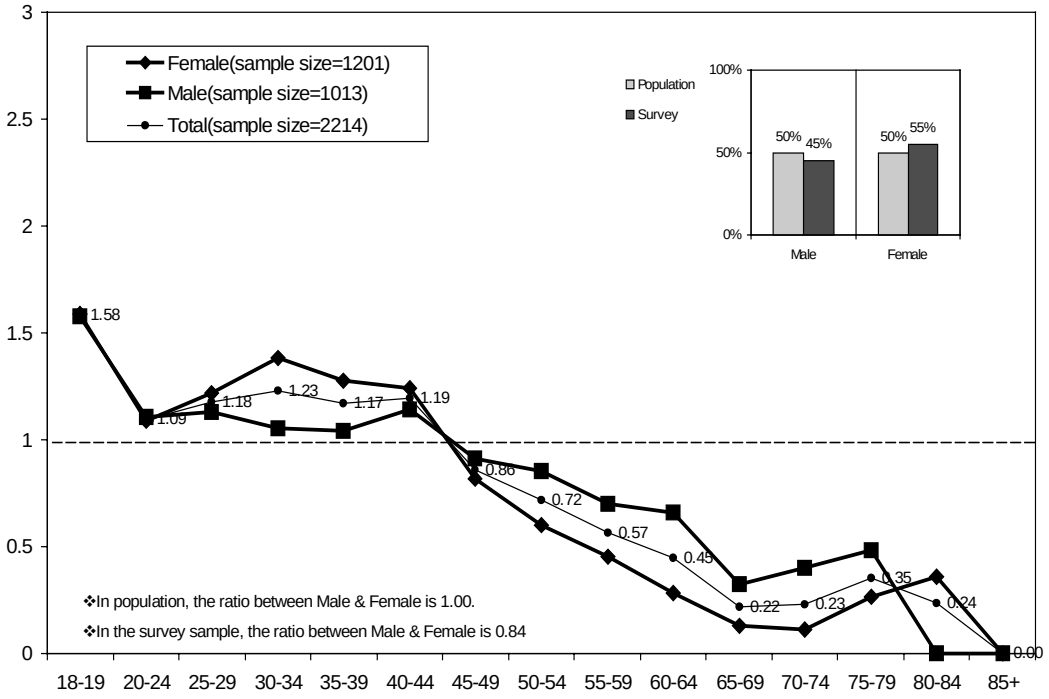


A sample of 5,000 persons throughout Turkey were selected by vote registers (1999) obtained from the Republic of Turkey, Higher Election Committee and available in Panajans which is the only document containing household and individual addresses in Turkey.

The sample was nationally representative and the survey was carried out in İstanbul, Bursa, Adana, Diyarbakır, Samsun, Gaziantep, Aydın. These provinces, except İstanbul, were selected as they were not a part of the full face-to-face household survey sample of the WHO.

Survey summary characteristics

Final Sample size	% Response rate	% Missing data	Deviation from UN population Chi-square Critical Value = 49.58 (29 df, p=0.01)
4,524	90	23.9	511.4



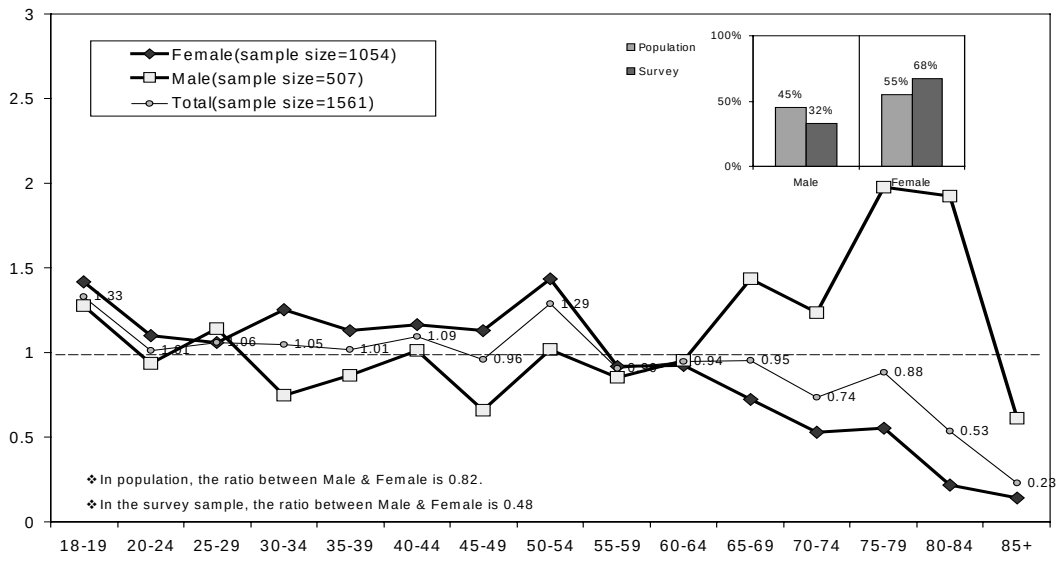
The sample contained 5,000 addresses from a postal register that the company created with the information they could gather based on their annual assessment survey.

Telephone directories could not be used as these have a low penetration rate (around 40% for all population) and the electoral register is updated only a few months before the election. Hence, these two registers are not representative of the Ukrainian population.

A multi-level stratified sample was used.

Survey summary characteristics

Final Sample size	% Response rate	% Missing data	Deviation from UN population Chi-square Critical Value = 49.58 (29 df, p=0.01)
1,562	31	4.5	129.9



The sample was a multi-stage random probability sample representative of the population residing in the U.A.E. The sample structure was based on the estimated population structure from the UAE Census data of 1995.

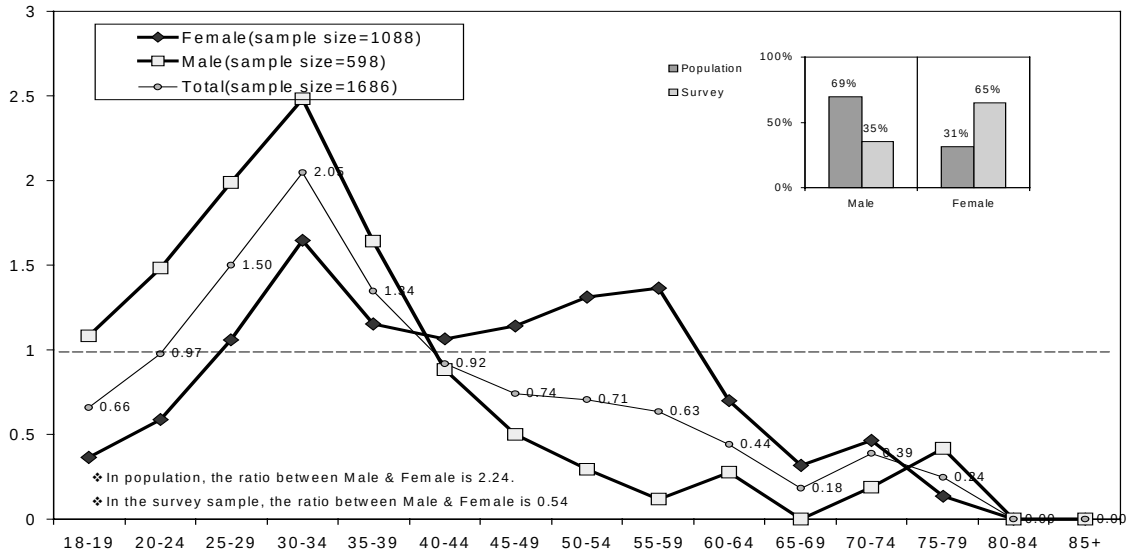
The estimates for various categories that are excluded from the sample were compiled using information from the UAE census. The categories of workers excluded from the sample and from projected population estimates include:

- Domestic helpers, servants, drivers etc., which are principally of Asian ethnic origin.
- Building watchmen, usually consisting of Egyptians, Sudanese, as well as some Pakistanis and Indians.
- Laborers in camps, or on sites such as construction workers, and service workers in facilities such as public transport, airport cleaning services, hospital cleaning services, etc. They are predominantly men, the majority of whom are of a low literacy level, and mostly unskilled.

The sampling system design employed in all the Emirates was an advanced sample design method. The survey did not cover any irregular, casual or provisional shacks or shelters. Interviews were conducted with people living only within legal households.

Survey summary characteristics

Final Sample size	% Response rate	% Missing data	Deviation from UN population Chi-square Critical Value = 49.58 (29 df, p=0.01)
1,686	72	17.6	469.5



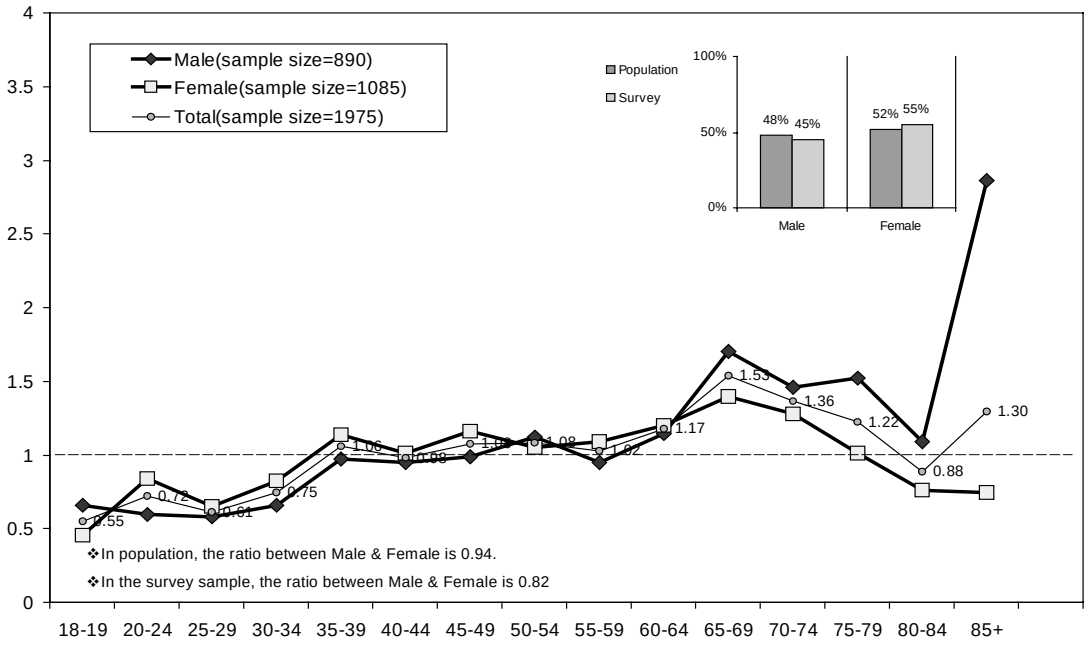
5,350 named individuals in the United Kingdom were systematically selected from the Electoral Register, which was stratified by local authority, and ordered by postcode.

Addresses were checked against Laing & Bussion's Care Home and Hospital Information database and 14 addresses were removed. A further 336 named individuals were systematically selected and removed from the remaining sample, using a random start and fixed interval method on the sample sorted by local authority and postcode, leaving 5,000 addresses for the usable sample.

The 5,000 sampled individuals were sorted by local authority and postcode.

Survey summary characteristics

Final Sample size	% Response rate	% Missing data	Deviation from UN population Chi-square Critical Value = 49.58 (29 df, p=0.01)
1,996	40	26.7	182.68



A sample of 5,000 households across the US was purchased from Survey Sampling, Inc. located in Connecticut. This sample is based on Random Digit samples.

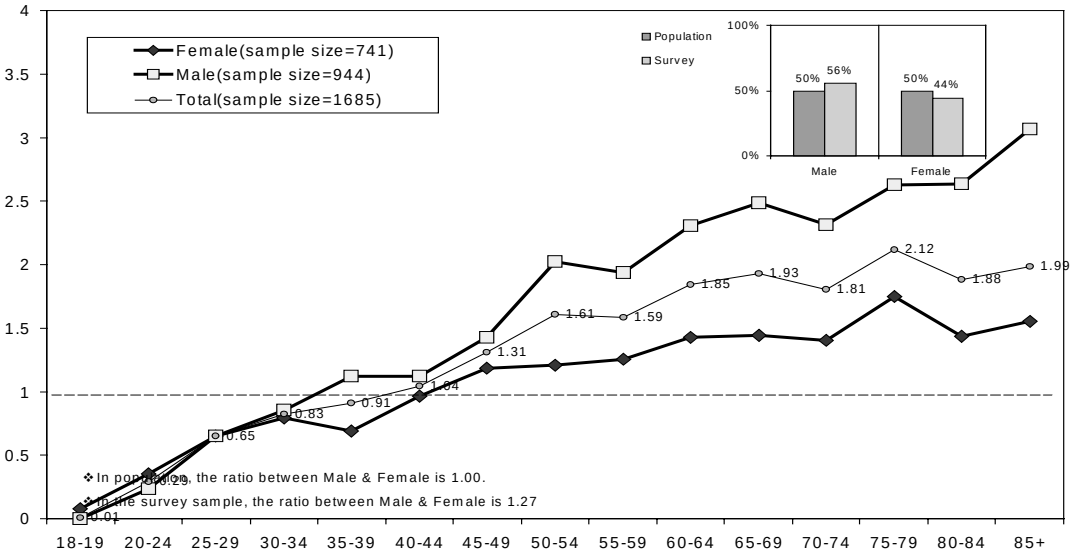
This sample was stratified by state to match the percentage of U.S. residents living in each of the fifty states.

The 5,000 sampled households were randomly assigned to one of three different experimental treatments (normal, personalized and personalised plus 2$ incentive)

The experiment was done for purposes of evaluating response rate effects of alternative means of contacting US residents.

Survey summary characteristics

Final Sample size	% Response rate	% Missing data	Deviation from UN population Chi-square Critical Value = 49.58 (29 df, p=0.01)
1,736	35	10.9	857.5



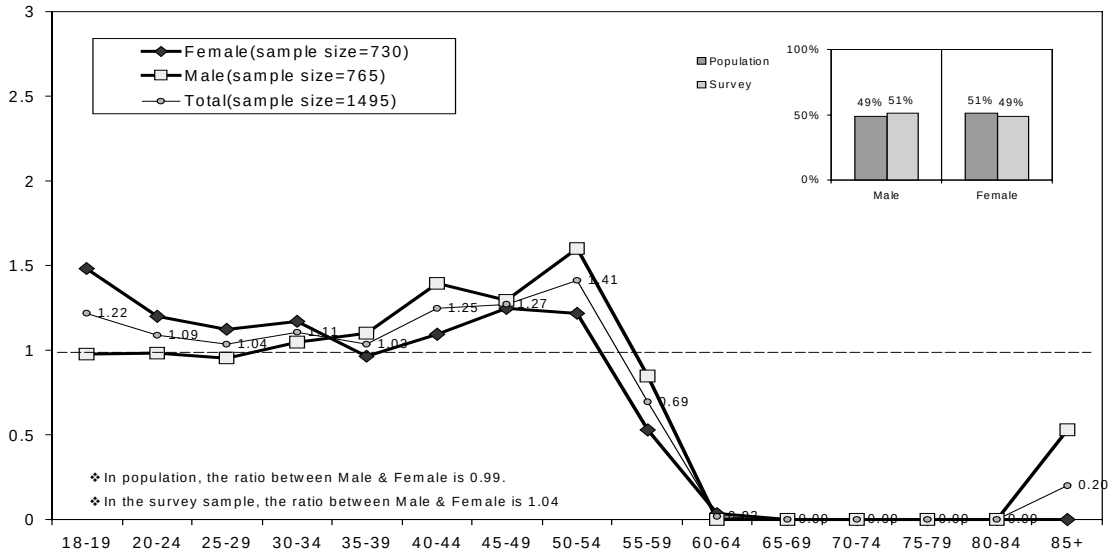
This was a stratified random multistage sample based on the following criteria:

1. By geographical population distribution, proportional to the population of the municipalities of the urban areas covered.
2. The sample distribution in each of the municipalities was based on the population size, the maximum number of interviews per municipality was 80 in the most inhabited municipality of Caracas and the minimum was two in several municipalities surrounding urban areas in other cities of the country.
3. Sample unit selection was dependent on social class.
4. The selected respondent within each sample unit was the one whose birthday was most recent to the date of the interview, with one call back.
5. Social class was dependent on the profile of each parish/municipality. In addition the following social class classification criteria were used:

    - Type of neighbourhood
    - Education of head of household and spouse
    - Occupation of head of household

Survey summary characteristics

Final Sample size	% Response rate	% Missing data	Deviation from UN population Chi-square Critical Value = 49.58 (29 df, p=0.01)
1,495	66	0.4	232.3



Appendix 3: Deviation of Study Sample from UN Population Database

Sample Population Deviation Index (study sample/UN population data base)

Males

	18-19	20-24	25-29	30-34	35-39	40-44	45-49	50-54	55-59	60-64	65-69	70-74	75-79	80-84	85+
China	0.88	1.00	0.84	0.89	1.07	0.99	1.42	1.29	0.90	0.94	0.85	0.87	0.64	0.68	0.34
Colombia	0.85	0.82	0.80	0.80	1.03	0.93	1.20	1.09	1.23	1.63	1.47	1.71	1.65	1.95	1.49
Egypt	0.62	0.60	0.90	0.84	1.31	1.25	1.20	1.31	1.13	1.47	1.13	1.05	1.40	1.47	0.94
Georgia	0.71	0.81	0.95	0.90	0.96	1.08	1.05	1.16	1.22	1.08	1.16	1.07	1.18	1.03	0.57
Indonesia	0.63	0.72	0.77	0.97	1.26	1.25	1.28	1.18	1.08	1.09	1.16	1.39	1.06	1.19	1.10
India	0.51	0.67	0.73	0.80	1.13	1.08	1.28	1.41	1.20	1.73	1.71	1.66	0.99	0.41	0.23
Mexico	0.63	0.62	0.81	0.96	1.11	1.15	1.12	1.38	1.25	1.53	1.34	1.97	1.56	1.39	1.00
Nigeria	0.68	1.19	1.22	0.90	0.72	0.85	0.91	1.11	0.66	1.40	0.63	1.49	1.19	2.71	7.03
Slovakia	1.02	1.14	1.08	0.86	0.86	0.75	1.02	1.04	0.89	0.97	1.26	0.67	0.71	0.61	0.25
Turkey	1.53	1.47	1.15	1.11	0.97	0.96	0.87	0.68	0.61	0.37	0.23	0.17	0.28	0.07	0.00

Females

	18-19	20-24	25-29	30-34	35-39	40-44	45-49	50-54	55-59	60-64	65-69	70-74	75-79	80-84	85+
China	0.88	1.16	0.87	0.96	1.19	1.25	1.28	1.19	0.72	0.80	0.64	0.61	0.35	0.37	0.23
Colombia	1.00	0.80	0.93	1.05	1.12	1.13	0.95	1.03	1.07	1.24	1.09	1.14	0.72	0.82	0.58
Egypt	0.61	0.80	0.98	1.27	1.28	1.15	0.95	1.02	0.90	1.36	0.94	0.61	0.64	0.41	1.06
Georgia	0.70	0.73	0.97	1.07	1.10	1.14	0.99	1.07	1.00	1.18	1.11	1.06	0.84	0.81	0.46
Indonesia	0.57	0.67	0.97	1.20	1.34	1.10	0.97	1.15	0.95	1.09	1.05	1.07	0.81	0.93	1.37
India	0.64	0.63	1.12	1.17	1.21	1.22	1.05	1.11	0.80	1.43	0.98	0.84	0.32	0.41	0.14
Mexico	0.68	0.72	0.81	0.82	1.09	1.06	1.19	1.18	1.35	1.57	1.64	1.49	1.47	1.32	1.17
Nigeria	0.62	1.21	1.18	1.00	0.74	0.97	0.67	1.14	0.67	1.47	0.91	1.07	0.69	2.42	2.77
Slovakia	0.88	0.93	1.06	0.97	1.10	1.00	1.06	0.97	0.98	1.00	0.76	0.96	0.58	0.94	0.37
Turkey	1.80	1.48	1.08	0.99	1.17	1.08	0.88	0.76	0.43	0.39	0.13	0.25	0.19	0.00	0.00