Survey Methodology
University of Helsinki
Introduction

Fall 2016

Seppo Laaksonen

Course Information
Introduction

Something about me

- Working 4 weeks at least in addition to the University of Helsinki: Statistics Finland, Eurostat, Stakes (now THL), Finnish Labor Research Institute, Southampton University, Surrey University, PISA Helsinki University, The European Social Survey
- Survey consulting in Hungary, Moldova, Slovenia, UK, Ethiopia, OKM Finland
- 204 research publications, 65 of these have been cited according to the Google Citations (714 citations altogether)

Chapters

1. Introduction to surveys and survey terms	Mainly Week 1
2. Questionnaire designing and survey modes	Week 2
3. Sampling principles and missingness mechanisms	Week 3
4. Design effects at sampling phase	Week 3 (not everything)
5. Sampling design data file	Week 3 (Basic ideas)
6. Missingness, its reasons and treatment	Week 4
7. Weighting adjustments due to unit missingness	Week 4 (Basics only)

Chapters (cont.)

8. Special cases in weighting	Week 4 (a bit only)
13. Basic survey data analysis using survey instruments	Week 5
9. Statistical editing	Week 5 (Basics)
10. Introduction to statistical imputation	Week 6 (Basics)
11. Imputation methods for single variables	Week 6 (Some ideas)
12. Summary and key tasks of survey data cleaning	Week 6 and earlier
Bibliography	All the time

Course practice

Language in delivered material is mainly in English. Finnish language is also used but I cannot say exactly when and how much. My e-book is larger than required in the course. The same is concerned my draft English version being emailed to you. This is newer and includes some recent ideas as well. It cover material for other survey courses as well.

In general, the survey world has been changing all the time, including:

- Problems in survey climate.
- New data collection tools.
- Web is growing, old tools are possibly disappearing.
- Smart phones are used.
- Media is abusing survey information.
- Social media is here.
- International surveys are most interesting while small scale surveys may still be needed like as pilots for real surveys, or for a specific topic.

Course practice

1. Lectures including discussion and debating
Wednesdays 16-19 (since 16:15 as long as about 3x45 minutes are spent)
2. Computer Class Training with real data sets (European Social Survey and the PISA, your own data are possible for extra credits), Thursdays 14-16 (14:15-15:45) or 16-18 (16.15-17:45). These two alternatives can be chosen freely but I will ask your preference soon.
SPSS is the main package but SAS is OK as well. R is not easily possible because its meta data tools are poor but the meta data of our data sets is fine; you thus can loose too much when using R.
Excel is possible to use in summarizing and for graphics.
The training tasks are sent by e-mail. The reporting can be made with WORD or POWER POINT or both. The template for reporting has been sent already and explained in the first training event.

Course practice

I think that I will use emails in our conversation as well and you can submit your comments by email as well. Naturally, I am most happy if I will get your feedback any time face-to-face.

The credits from the course:

- Main option = 8 if the exam has been passed successfully and the training and its basic report is reasonably done.
- Minimum option = 5, if the course has been made 60% (exam + training)
- Intermediate options = 6 to 7 credits as agreed mutually
- More than 8 credits if additional work is done (max =12). Agreed with me.
- The first exam event is obviously 19th October, the next in December.

Questions?

Survey is a methodology and a practical tool used to collect, handle, and analyze in a systematic way information from individuals. These individuals or micro units can be of various types, such as people, households, hospitals, schools, businesses, or other corporations. The units can be simultaneously available from two or more levels, such as from households and their members. Information in surveys may be concerned various topics such as people's personal characteristics, their behavior, health, salary, attitudes and opinions, incomes, poverty and housing environments, or characteristics and performance of businesses. Survey research is unavoidably inter-disciplinary, although the role of statistics is most influential since the data for surveys is constructed in a quantitative form. Correspondingly, many survey methods are special statistical applications. However, surveys exploit substantially many other sciences such as informatics, mathematics, cognitive psychology, and theories of subject-matter sciences of each survey topic.

Five populations in surveys after target group

1. *Population of interest* is the population that a user would like to get or estimate ideally but it is not possible always to completely reach and hence she/he determines the second population:

1. *Target population* which is such a population that is realistic. Naturally, this population should be exactly determined including its reference period (a point of time or a time period).

Examples about target populations used in this book. We do not mention any year since it varies in the first two cases but it should be mentioned.

- The European Social Survey (ESS): ”Persons 15 years or older who are residents within private households in the country in the first of November.”
- The EFSS (European Finnish Security Survey): 15 to74 years old non-Swedish speaking residents in Finland 1st of October.
- The Programme for International Student Assessment (PISA) survey: fifteen-year-old school students (this is specified so that the full calendar year is covered).
- The grid-based study of Finland: People from 25 to 74 years of age living in southern Finland.

3. *Frame population and the frame* from which the statistical units for the survey can be found. Usually, the frame is not exactly from the same period as the target population (delay in population surveys is rather short i.e. 1-5 months, but for enterprise surveys much more, even years).

The frame is not always at element level available as in the case of the central population register based surveys.

Instead, the frame population can be created:

From several frames, often from the three ones and so that they are not available when starting the survey fieldwork. However, the first frame is necessary to be able to begin. This consists often of regions or areas or schools or addresses, but it is needed later other frames, fortunately only of those who have selected in the first stage.

4. *Updated frame population* is useful for estimating the results better. Usually, the initial frame population has been used for estimation too. This may lead to biased estimates. Fortunately, this bias is not severe in most human surveys. At contrast, old frames can lead to dramatic biases in business surveys, if this is concerned large businesses.

Finally, we will have the fifth population when we also know how well the fieldwork has been succeeded.

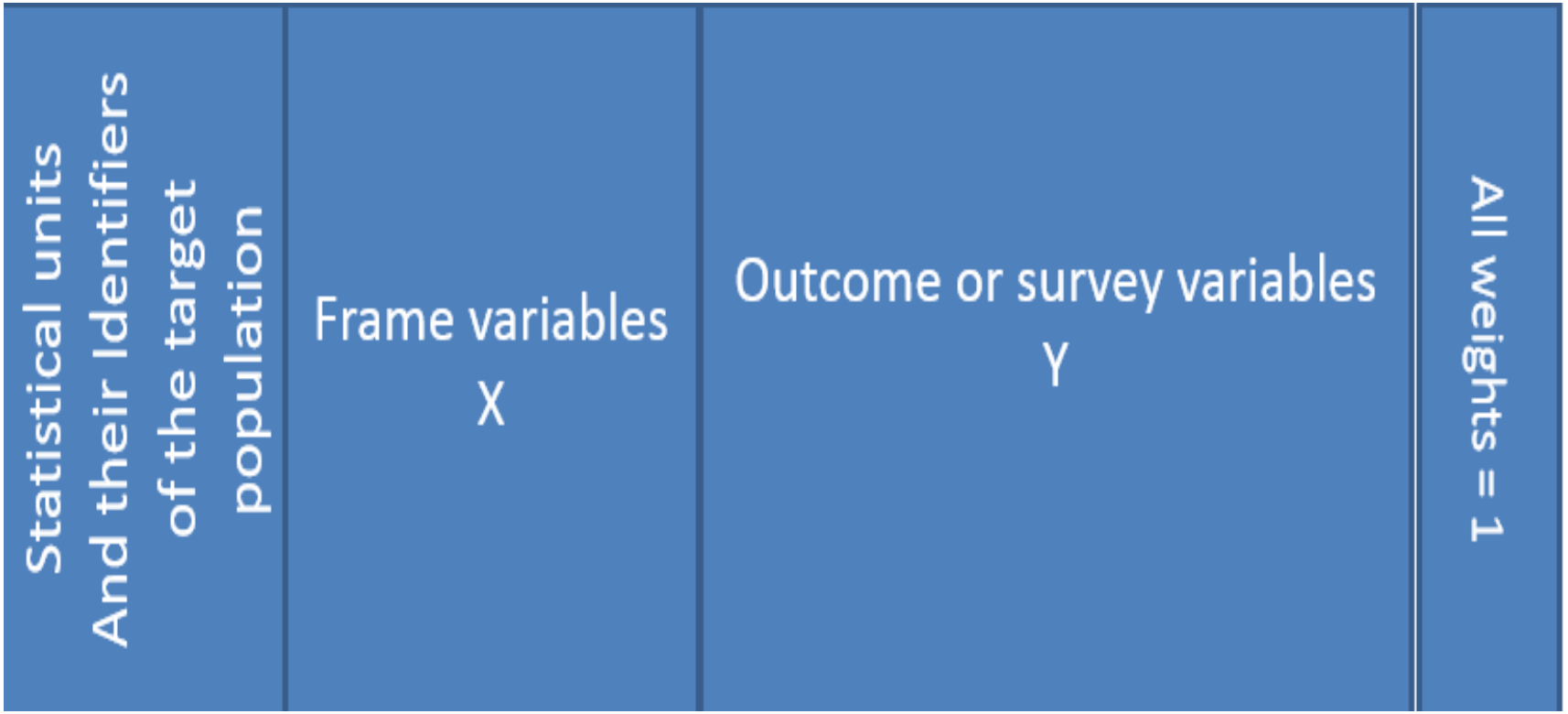
5. *Survey population or study population.*
It is ideal if this fifth population corresponds to our target population or even the population of interest. But if not, the estimates are somewhat biased. If clear gaps are in the final data, this should be informed for users, thus how much the survey population differs from the target population?

Purpose of populations

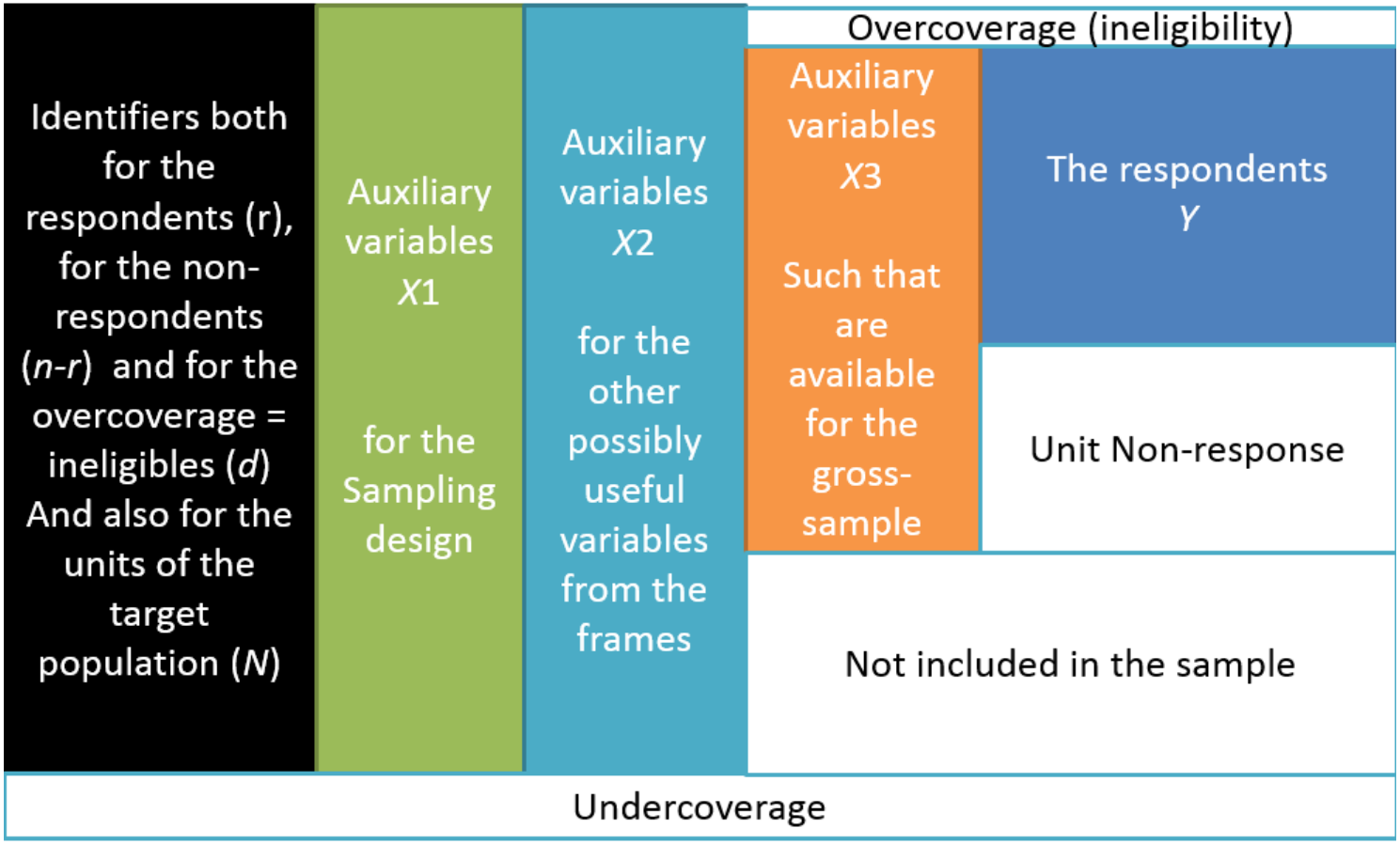
Before continuing with survey terms it is good to discuss about the purpose of these populations. Naturally, the first point is to approach to the targets of the survey as well as possible, and hence it is needed to know all steps and possible gaps passed or hopefully solved.

The final target is to estimate the desired parameters, such as averages, standard deviations, medians, distributions, ratios and statistical model parameters. This can be made just calculating whatever ways but such figures cannot be generalized at any population level without using survey instruments that are explained in this book. If all coverage and related problems are solved, the estimates can be generalized at the target population level.

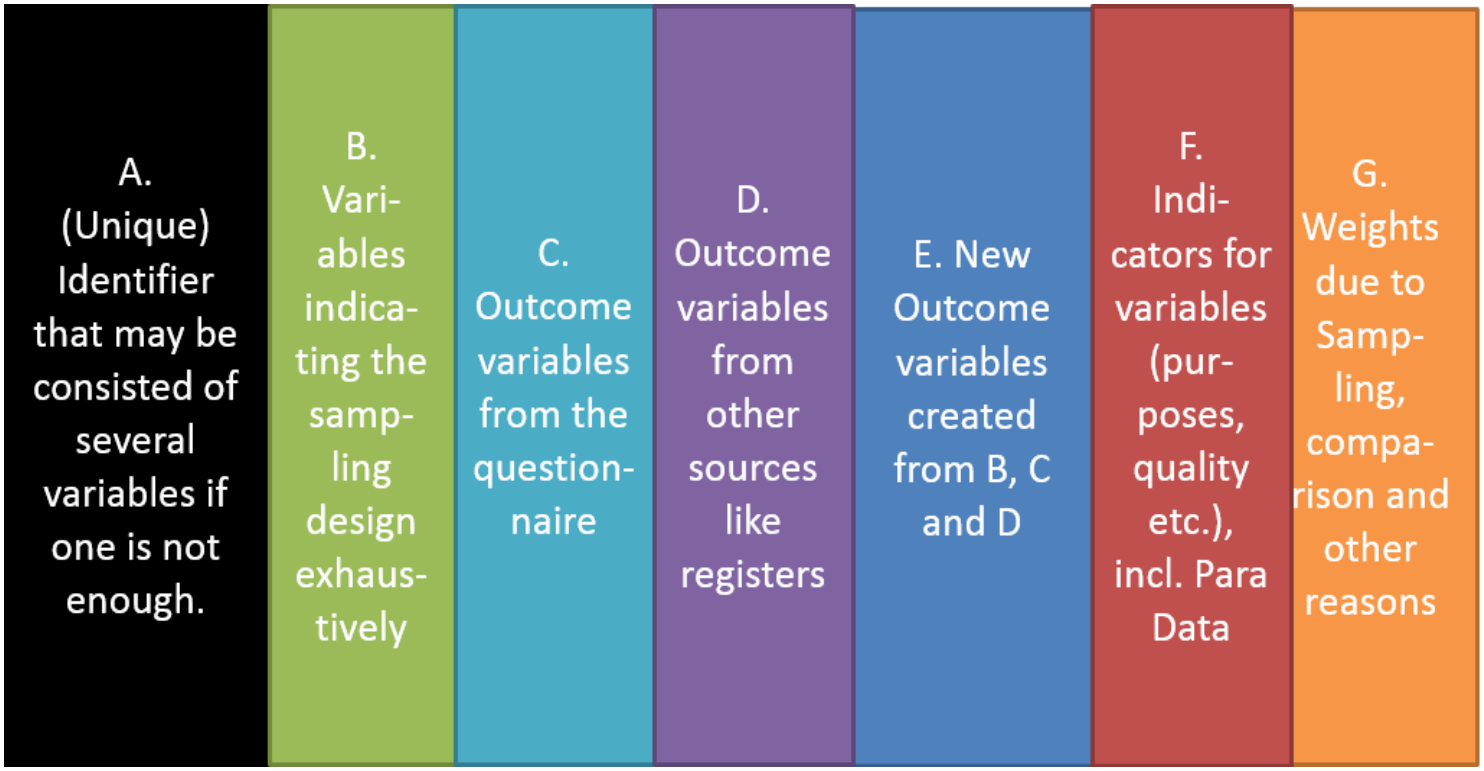
Scheme 1.1 Micro data for the entire target population



Scheme 1.2 General structure of a micro level cross-sectional survey data file. The weights variables are not here. See Scheme 1.3.



Scheme 1.3 General structure of a micro level cross-sectional survey data file that consists of r respondents (rows of the matrix). It is possible that there are outside this scheme other data, e.g. more para data, and content data.



X variables, auxiliary variables in more details
These variables can be found, collected and/or downloaded from the different sources, as follows:

- Population register (e.g. age, gender, members living in the same address, house type and size, kitchen type, many regional or areal options are available including grids).
- Other registers such as tax register or job seekers' register, formal education register (e.g. tax income, unemployed, education)
- Other administrative sources, often at aggregate level (e.g. % owner occupation, % social renting, % detached housing, % divorced, % under crowding, % 2 or more cars, 1 or more cars, % owner occupation, % unemployed, % long term unemployed, % social renting, % highly educated); the aggregate here may vary, being e.g. municipality, postal area code, grid square, block, village).

- In panels and longitudinal analysis, the variables of preceding points of time can be used as auxiliary variables if their values are believed to remain correct.

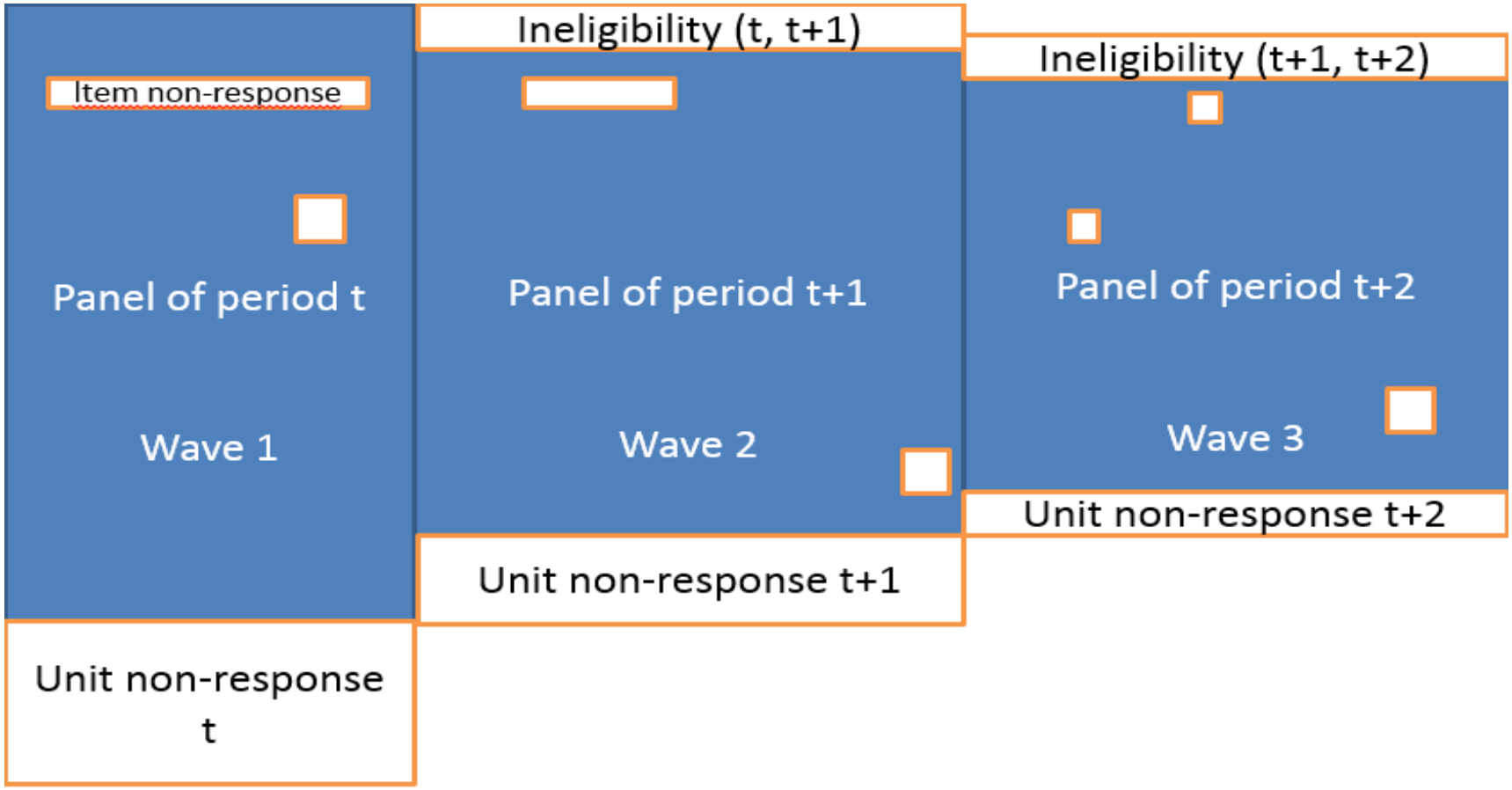
- Using interviewer observations of the immediate vicinity of the houses of sample units about visible signs of neighbourhood disorder. These observations of neighbourhood disorder or decay can be linked to the ‘broken windows’ hypothesis. The neighbourhood has been classified into one or more variables by an interviewer using harmonised rules. This type of X variables is becoming more common but it is difficult to get, regularly in particular.

Sampling weights are of two types:

- Their average is for each target population = 1 and hence their sum = the number of the respondents. They are called analysis weights. These thus are relative weights and good to use in comparing the weights of different surveys. The ESS weights are of this type but the Pisa weights are of the second type.

- Their sum = the number of the target population units (e.g. households or individuals) and each weight indicates how many units one unit represents in the target population; thus these weights are for generalizing (estimating) the results?

Scheme 1.4 General structure of cohort *type of panel or longitudinal data*



Summary of the terms and symbols of Chapter 1

U = target population

D = over-coverage or ineligibles

N = size of the target population (under-coverage may be a problem)

d = number of the ineligibles in the gross sample

r = number of the (unit) respondents = net sample size

n = number of units of the target population in gross sample

$n+d$ = gross sample size

k = statistical unit, e.g. for the respondents k=1, ..., r

$r(y)$ = number of responses to the variable y

Transformations

This annex of Chapter 1 pays attention to transformations, thus concerning Variables E of Scheme 1.3. Each single variable can be transformed into another scale or into different categories than initially. The simplest transformation is linear that only changes the scaling even though the results are similar. The purpose of the linear transformation is to make the results easier to interpret. Typical other transformations that lead to a new (and hopefully better) interpretation or going to satisfy the model conditions for example, are

- logarithmic for ratio-scale variables such as income and wage in which case the outcomes are relative (log-percentages).
- exponential that is most often used to return from logarithmic to initial but this transformation can be in a few cases possible as such if the distribution is 'peculiar'
- categorization into two or more categories, most common is binary (dichotomous) variable.

In the end of this chapter two most commonly used transformations for several initial variables are given. The outcome of such transformation is called 'summary variable' or 'compound variable.' These two most common technics for constructing this variable are:

1. Linear transformations of each initial variable into the equal scale, and then taking the average of them.
2. Exploratory factor analysis of the variables with the same phenomenon that leads to a smaller number of variables.

Question B 31

How about people from the poorer countries outside Europe?

Variable name and label: IMPCNTR Allow many/few immigrants from poorer countries outside Europe

Values and categories

1 Allow many to come and live here
2 Allow some
3 Allow a few
4 Allow none
7 Refusal
8 Don't know
9 No answer

The variables `IMSMETN and IMDFETN` are respectively concerned either same or different ethnic groups.

Question B 32

Would you say it is generally bad or good for [country]'s economy that people come to live here from other countries?

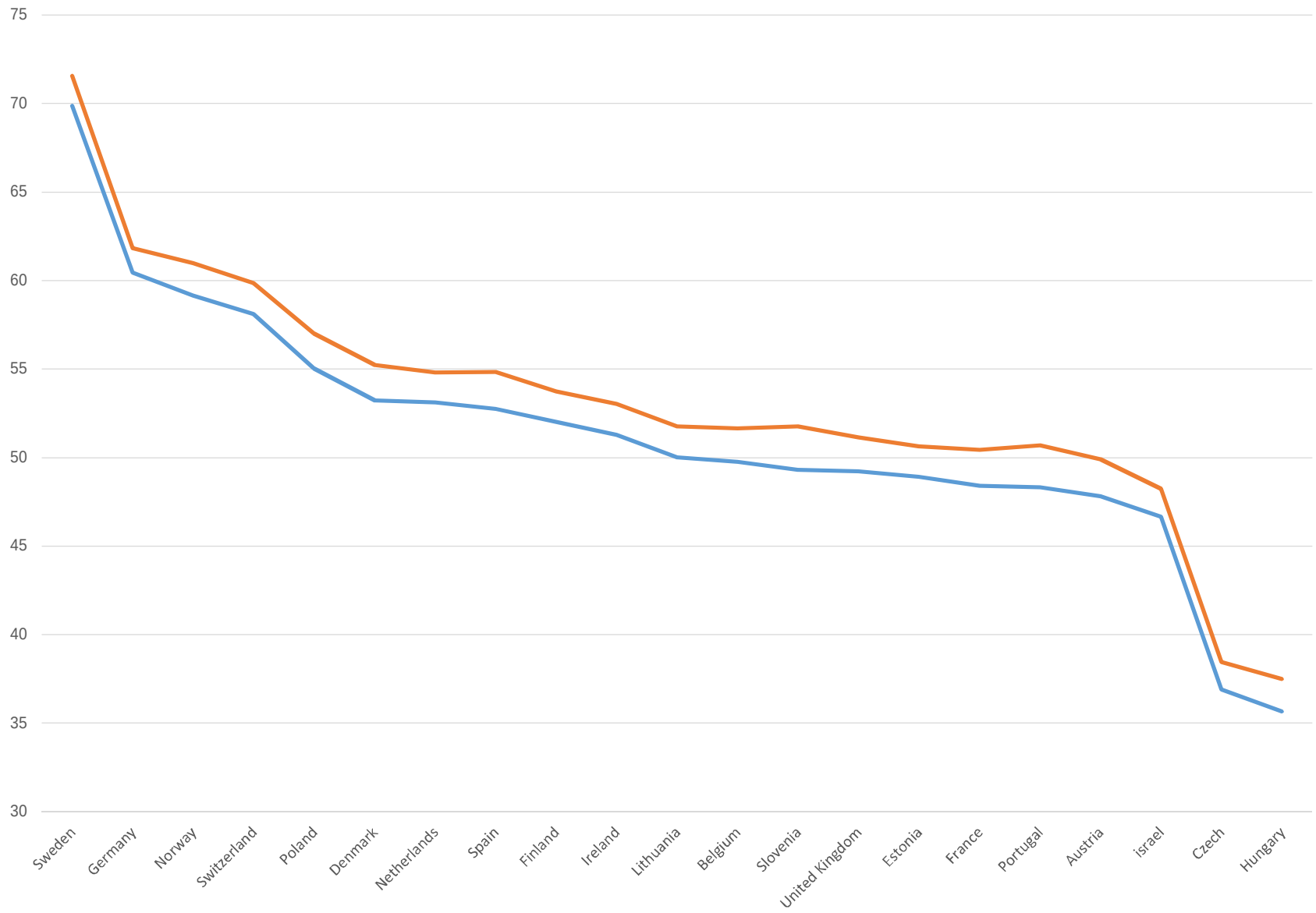
Variable name and label: IMBGECO Immigration bad or good for country's economy

Values and categories

00 Bad for the economy
01 1
02 2
03 3
04 4
05 5
06 6
07 7
08 8
09 9
10 Good for the economy
77 Refusal
88 Don't know
99 No answer

This transformation for the first group is obtained with the function (100/3) x (4-IMBGECO) and the analogously to the two other variables, whereas by multiplying with 10 for the second group.

Finally, the average of all these linearly transformed variables is our new summary (compound) variable.


75
70
65
60
55
50
45
40
35
30
Sweden
Germany
Norway
Switzerland
Poland
Denmark
Netherlands
Spain
Finland
Ireland
Lithuania
Belgium
Slovenia
United Kingdom
Estonia
France
Portugal
Austria
israel
Czech
Hungary

Example 1.2 Summary/compound variable using exploratory factor analysis and factor scores

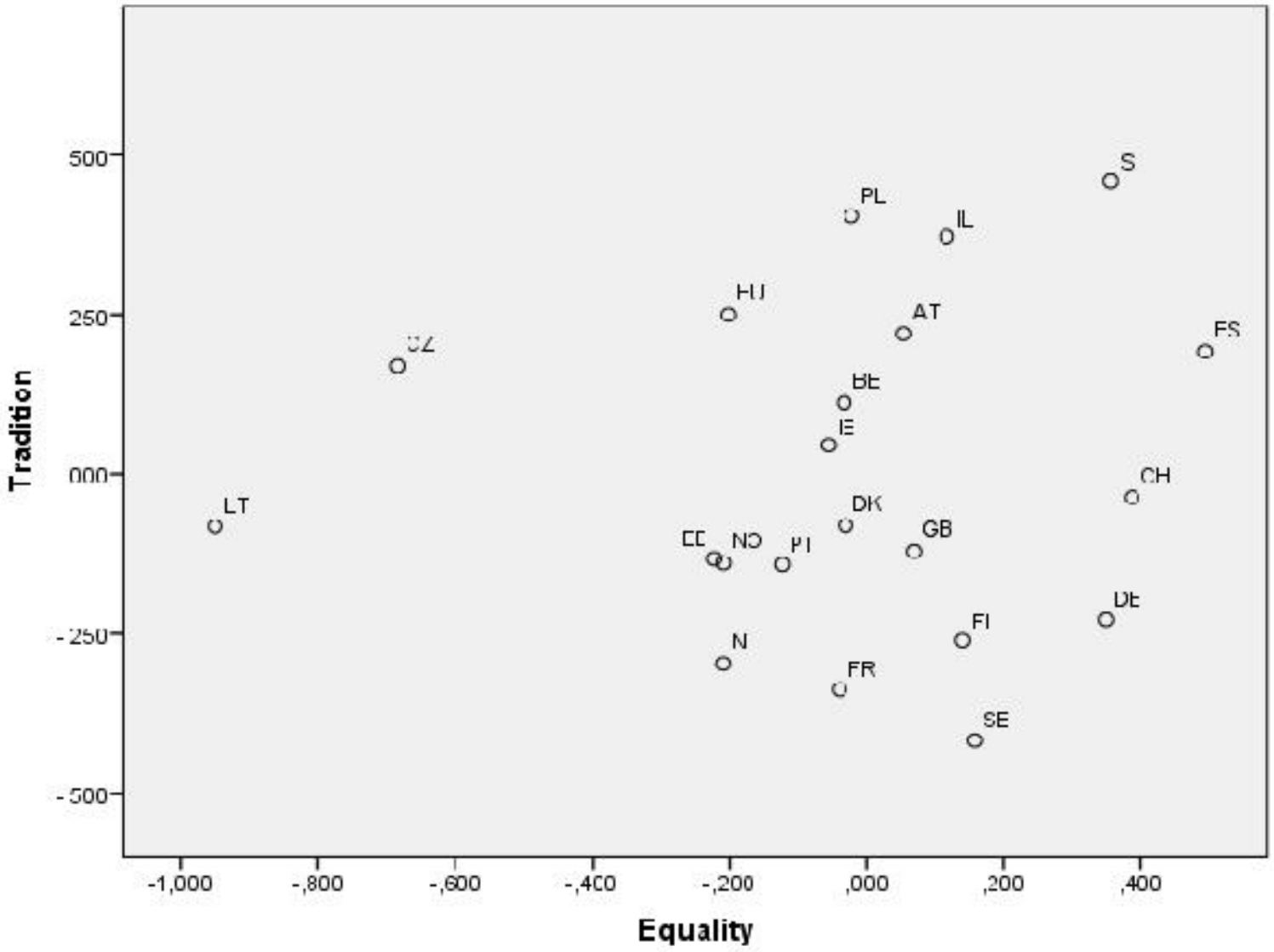
The initial factorized variables are Shalom Schwartz' Human values of the ESS, 21 altogether Schwartz (2012). There are six alternatives in the questionnaire to answer but so that any completely neutral category does not exist (see Chapter 2 as well). We thus wish to find a much smaller number of dimensions of human values by exploratory factor analysis. The step here is to omit values with missingness codes. Fortunately, their number is fairly small. Table 1.1 shows the VARIMAX rotated factor pattern of all 21 questions for four factors. This same number has been obtained in each round but the order of the factors varies to some extent. The table also includes the factor loadings that helps in the interpretation of these four dimensions. The highest loadings are marked.

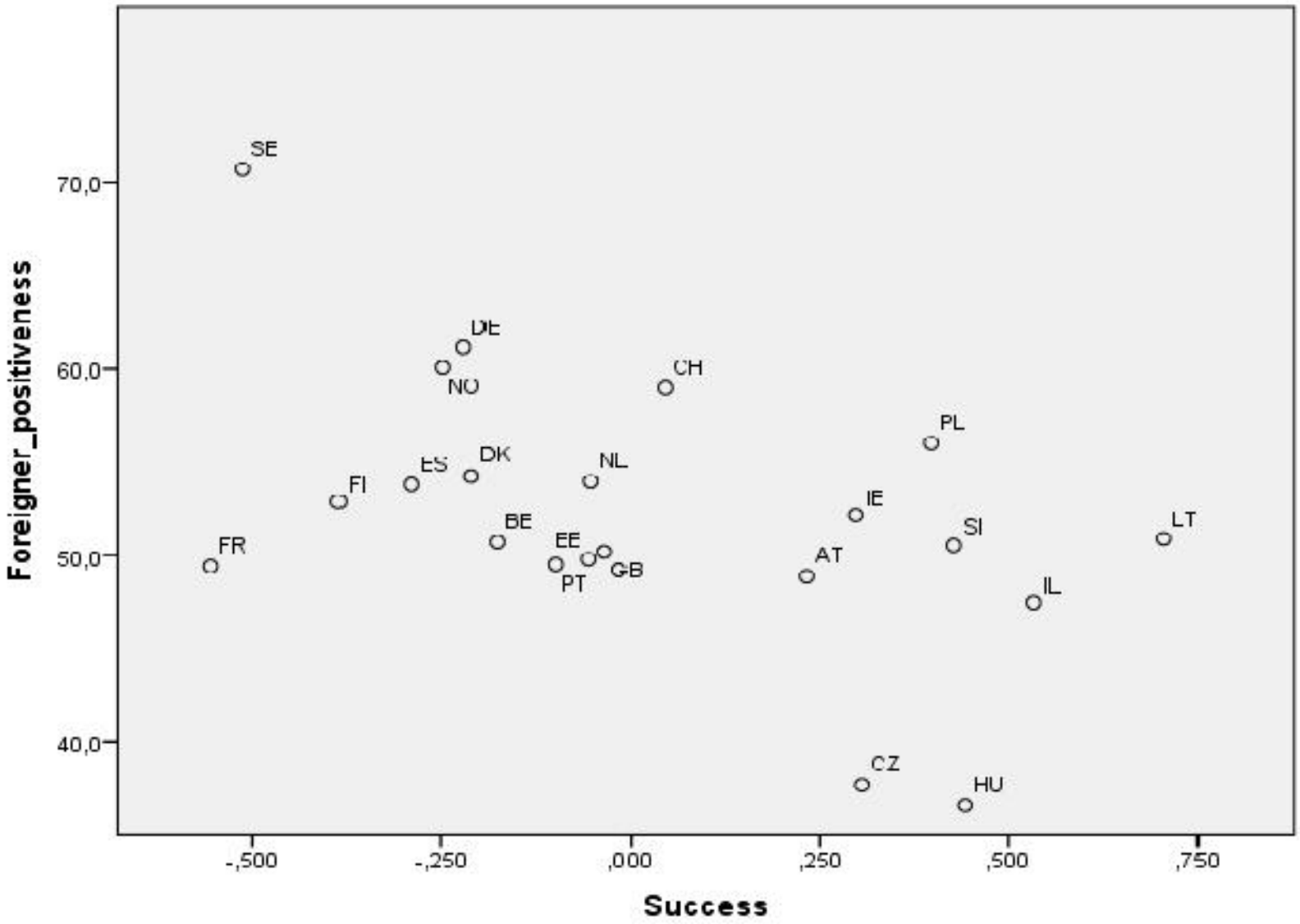
Rotated Factor Pattern

		Factor1	Factor2	Factor3	Factor4
ipcrtiv	Important to think new ideas and being creative	0.50039	-0.19926	0.36534	0.19092
imprich	Important to be rich, have money and expensive things	-0.19119	0.05636	0.67597	0.20425
ipeqopt	Important that people are treated equally and have equal opportunities	0.68608	0.03273	0.04745	-0.06398
ipshabt	Important to show abilities and be admired	0.16097	0.10197	0.72829	0.17537
impsafe	Important to live in secure and safe surroundings	0.19957	0.57004	0.27976	-0.17525
impdiff	Important to try new and different things in life	0.32647	-0.05826	0.26154	0.57487
ipfrule	Important to do what is told and follow rules	-0.01832	0.63355	0.09810	0.03434
ipudrst	Important to understand different people	0.65828	0.12926	-0.08800	0.16965
ipmodst	Important to be humble and modest, not draw attention	0.28233	0.49916	-0.26870	0.02342
ipgdtim	Important to have a good time	0.14466	0.04653	0.16798	0.71054
impfree	Important to make own decisions and be free	0.47514	-0.01510	0.25188	0.21841
iphlppl	Important to help people and care for others well-being	0.62314	0.27703	-0.04811	0.17742
ipsuces	Important to be successful and that people recognize achievements	0.10018	0.17724	0.70252	0.27157
ipstrgv	Important that government is strong and ensures safety	0.24328	0.54689	0.21960	-0.05419
ipadvnt	Important to seek adventures and have an exciting life	0.05952	-0.16632	0.31919	0.67231
ipbhprp	Important to behave properly	0.14602	0.70628	0.06306	-0.01267
iprspot	Important to get respect from others	-0.00533	0.42353	0.53092	0.14803
iplylfr	Important to be loyal to friends and devote to people close	0.56952	0.27867	-0.05466	0.24730
impenv	Important to care for nature and environment	0.57448	0.26809	0.00743	-0.01528
imptrad	Important to follow traditions and customs	0.04373	0.64219	0.02138	0.04902
impfun	Important to seek fun and things that give pleasure	0.09437	0.08629	0.10504	0.79910

Factor name	Interpretation
Factor1 = (Equality)	People are treated equally, understand different people, help people
Factor2 = (Tradition)	Behave properly, follow traditions and rules
Factor3 = (Success)	Show abilities and be admired, be successful and rich
Factor4 = (Enjoy)	Seek fun and things that give pleasure, have a good time

The final factor score variables
Equality = - Factor1
Tradition = -Factor2
Success = -Factor3
Enjoy = -Factor4


Tradition
500
250
000
-250
-500
-1,000
-,800
-,600
-,400
-,200
,000
,200
,400
Equality
S
PL
IL
HU
AT
FS
LT
DK
GB
FI
N
FR
SE
CH


Foreigner_positiveness
70,0
60,0
50,0
40,0
SE
DE
NO
CH
PL
DK
ES
NL
FI
IE
BE
SI
LT
EE
FR
AT
GB
PT
IL
CZ
HU
-,500
-,250
,000
,250
,500
,750
Success

Chapter 2

Questionnaire Designing and Survey Modes