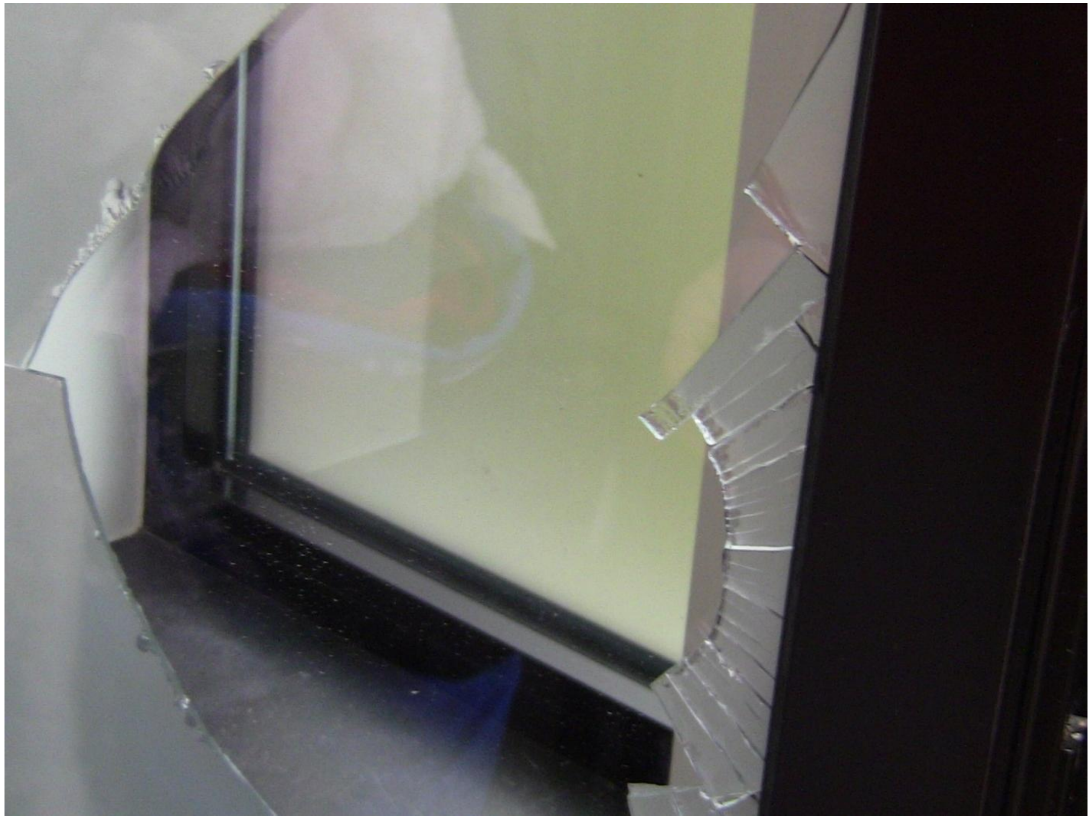
Chapter 6

Missingness, its reasons and treatment

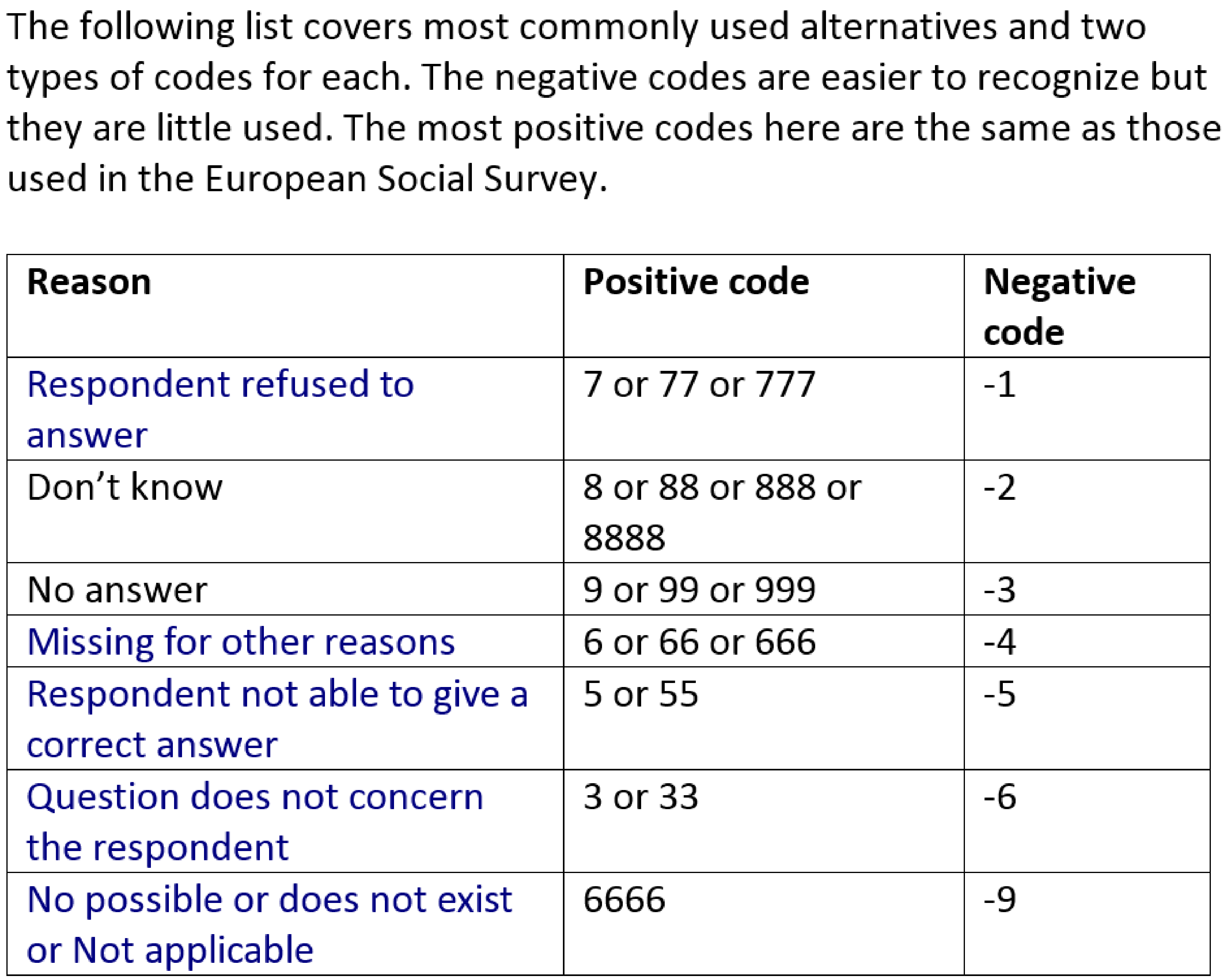


Reason	Unit nonresponse	Item nonresponse
Non-contact due to incorrect data	Possible	Not possible since the data is concerned respondents
Inability to answer correctly	Due to general disability, is possible	Information difficult to get due to disability or hard to get a correct answer for various reasons.
Hard refusal	Don’t participate at all, maybe in any other surveys either	Not possible since this is concerned respondents
Soft refusal	Reply to most questions but classified as respondent	Does not reply to all questions for various reasons (item nonresponse), see below
Screening question	Not necessarily any problem	Second stage answers missing but the first can be used in analysis
Lost data	Possible	Should not be possible
Other or unknown reason	Possible	Possible

Table 6.1 Response rates and reasons for unit nonresponse in some countries of Round 3 of the European Social Survey. The last column includes the response rate of a later round, i.e. either Round 6 or 7.

	Ineligibility rate (%)	Response rate (%)	Noncontact rate (%)	Refusal rate (%)	Response rate of a later round
Austria	1.7	62.5	7.8	28.6	51.9
Belgium	4.9	61.5	7.1	22.7	57.4
Denmark	6.4	65.1	5.6	23.9	51.9
Finland	1.5	70.8	2.8	21.2	62.9

Codes for missing-ness



The following list covers most commonly used alternatives and two types of codes for each. The negative codes are easier to recognize but they are little used. The most positive codes here are the same as those used in the European Social Survey.

Reason	Positive code	Negative code
Respondent refused to answer	7 or 77 or 777	-1
Don't know	8 or 88 or 888 or 8888	-2
No answer	9 or 99 or 999	-3
Missing for other reasons	6 or 66 or 666	-4
Respondent not able to give a correct answer	5 or 55	-5
Question does not concern the respondent	3 or 33	-6
No possible or does not exist or Not applicable	6666	-9

Missingness indicator and missingness rate

In the case of **unit missingness**, two missingness indicators can be created:

(i) **The nonresponse indicator** that gets either the value = 0 if the unit responds or the value = 1 if the unit does not respond but belongs to the target population.

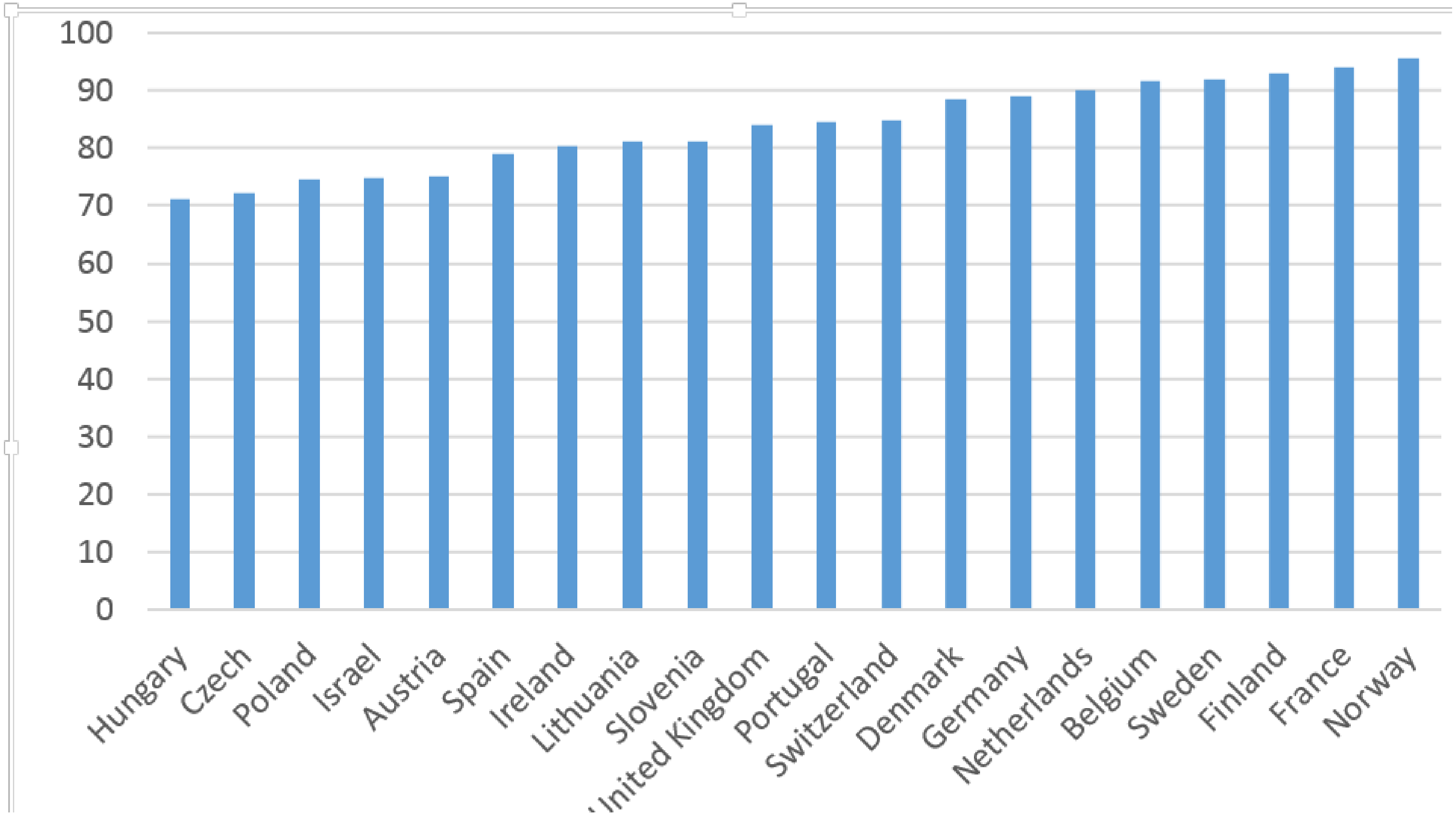
(ii) **The ineligibility indicator** that gets the value = 1 if the unit is ineligible, and the value = 0 if the unit responds or does not respond.

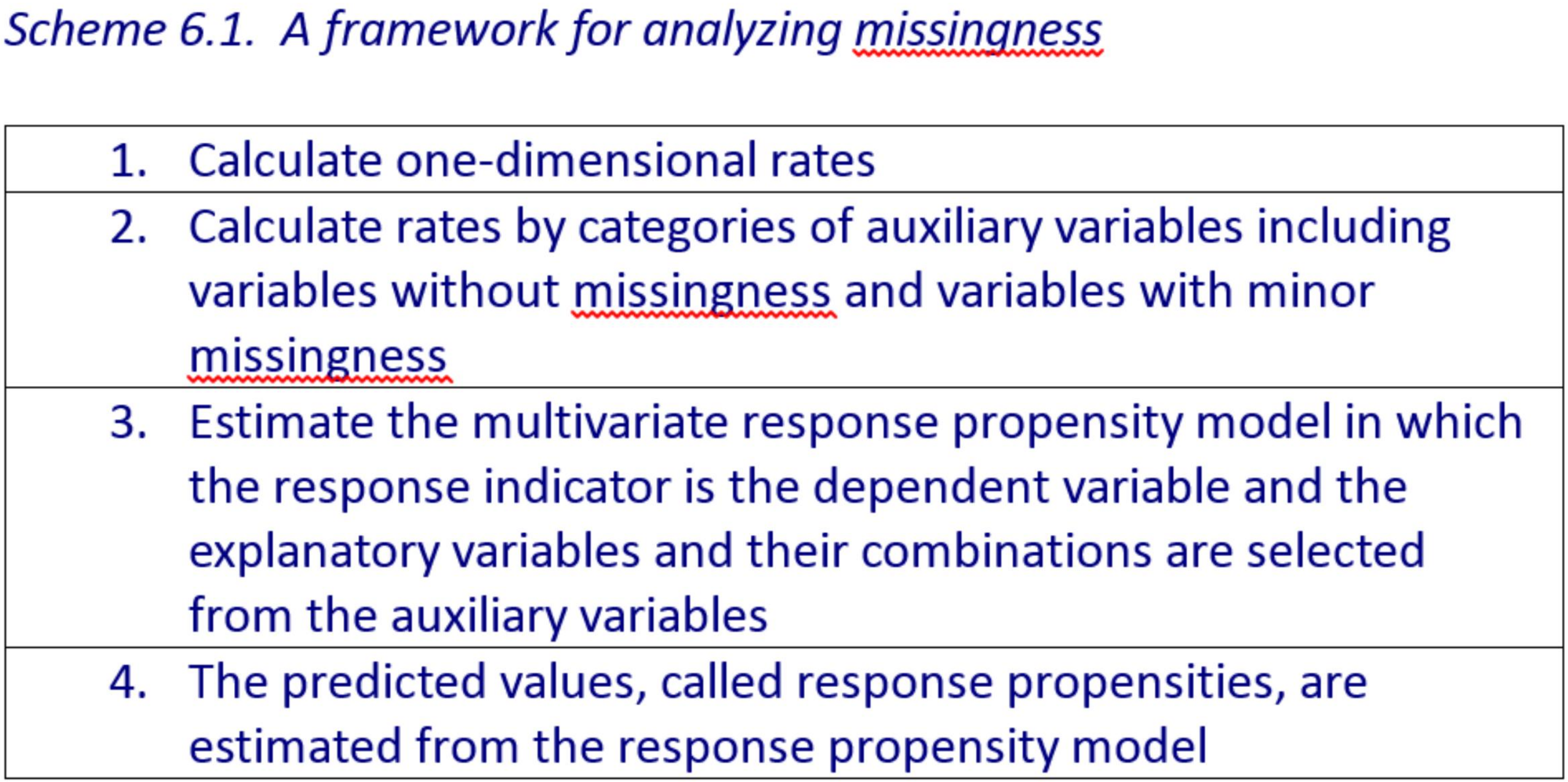
These indicators can be considered as complements as well, changing the values 1 and 0. Now the first indicator is called (unit) **response indicator.** The same label has sometimes given to the second case, in which case the zero category covers both the non-respondents and ineligibility units. This is the only alternative if there has been difficulties to distinguish these two alternatives; this occurs if the unit cannot be contacted at all.

In the case of **item missingness**, one type of the indicator only is needed although this might be determined either from the positive or negative direction. We present it from the positive direction:

(iii) **The item response indicator** that gets the value = 1 if a valid answer into the question *y* is obtained, and the value = 0 if the respondent has not given any valid value. Note that this is concerned only the unit respondents.

Figure 6.1 Item response rates of income by 20 ESS countries, round 7

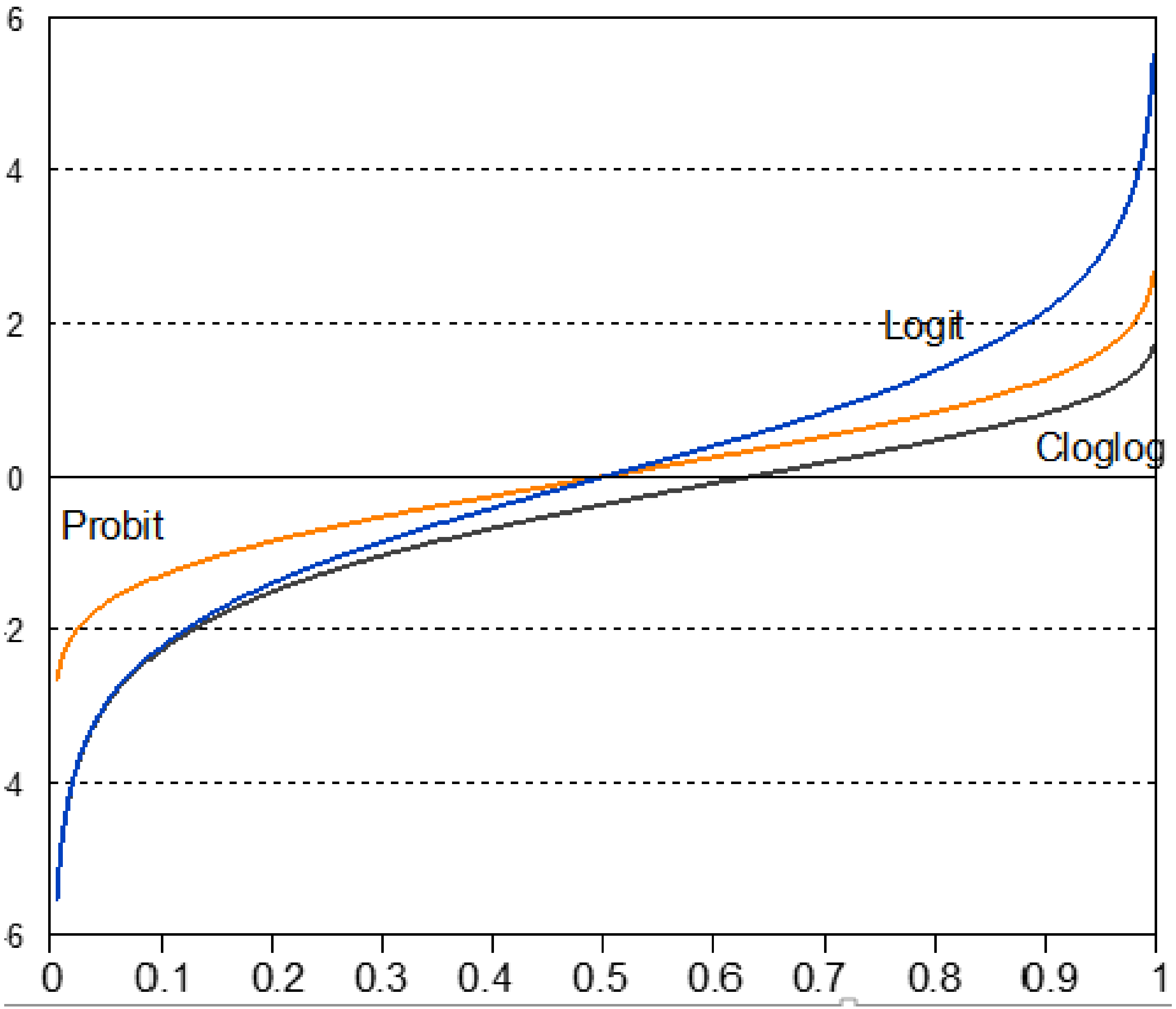




Scheme 6.1. A framework for analyzing missingness

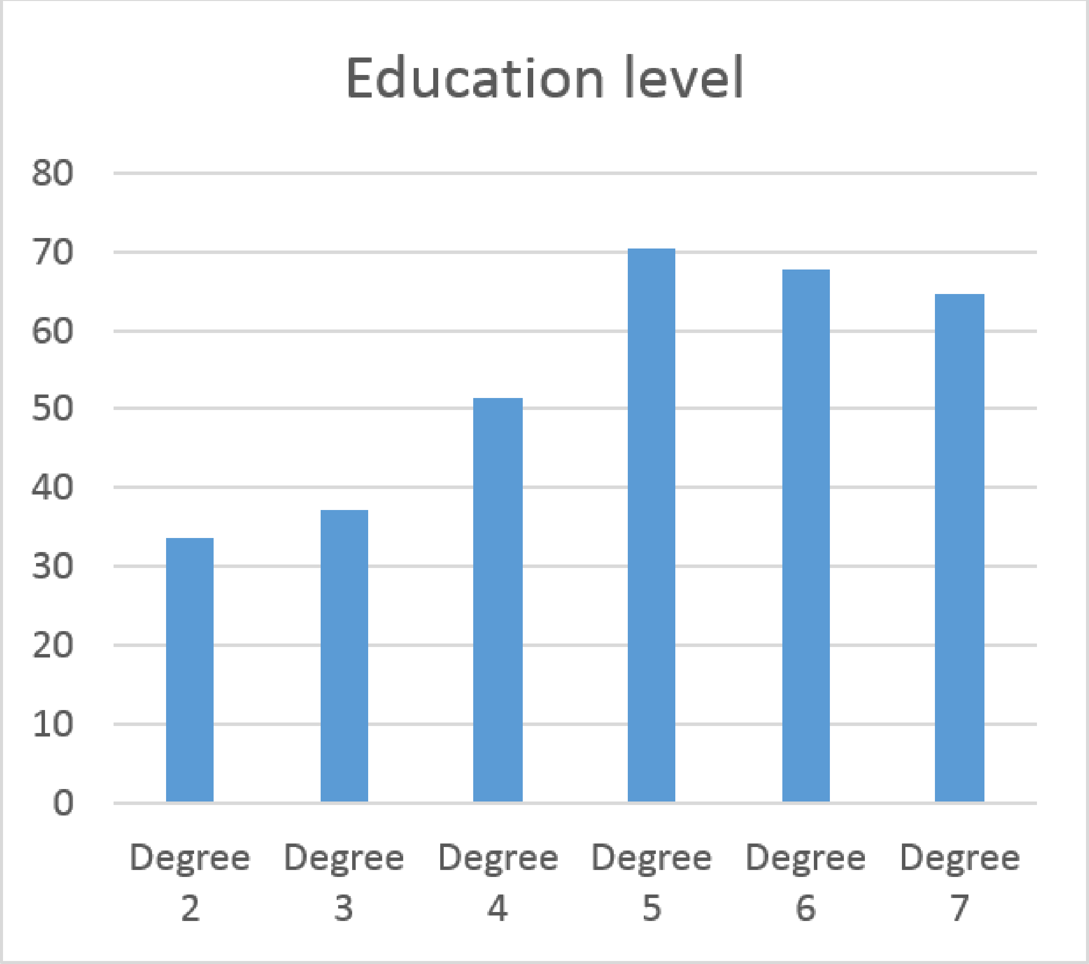
1. Calculate one-dimensional rates
2. Calculate rates by categories of auxiliary variables including variables without missingness and variables with minor missingness
3. Estimate the multivariate response propensity model in which the response indicator is the dependent variable and the explanatory variables and their combinations are selected from the auxiliary variables
4. The predicted values, called response propensities, are estimated from the response propensity model

Three most common link functions, in addition log-log is a mirror curve for complementary log-log = Cloglog



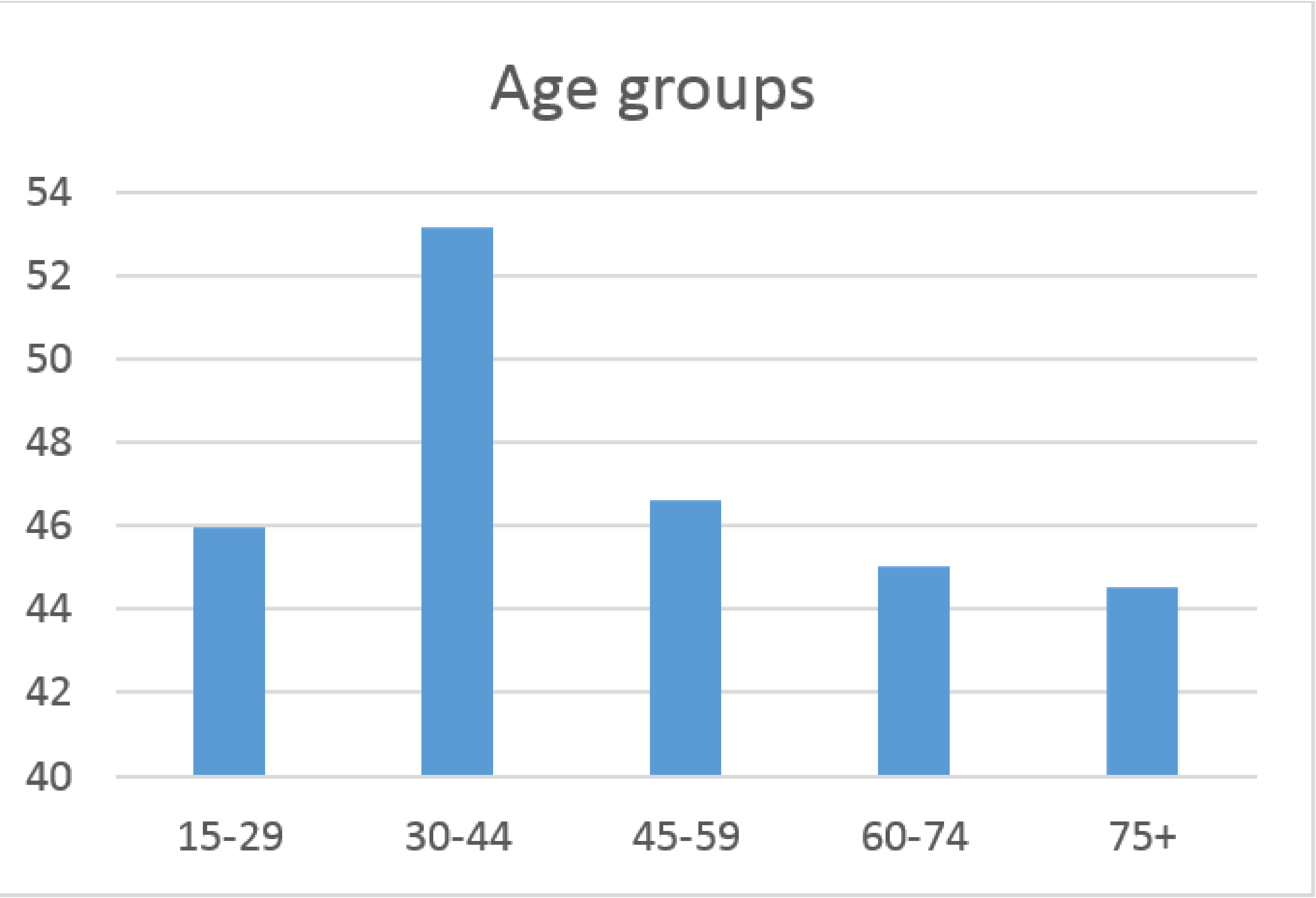
From our test data

Figure 6.1 The unit response rates by Education level, Age group and Marital status in the test data



Another example

From our
test data



Example 6.1 Propensity model for item response.

This is from the same ESS data as Figure 6.1 in which the response rates by 14 countries for income are given. Now we go forward by estimating a logit regression for this response indicator, taken a number of explanatory variables in the model. Best variables are such without missing values. This number is not big, but in addition to country, the following ones can be used:

- Age
- Gender
- Household size
- Interviewing time.

On the other hand, it is possible to test such variables whose missingness is minor. We here find the two such variables

- Happiness (item nonresponse rate = 0.3%)
- Marital status (1.0 %)
- Subjective income (0.9%)

The estimation of the model indicates that the one explanatory variable only is not significant. This is subjective income. It thus means that the item nonresponse of objective income is not depending on subjective income. We use this information in an example of Chapter 10.

At contrast the other explanatory variables thus are significant. We do not present all details of this result but some main lines:

- Males were able to tell the income of their households better than females (odds ratio = 1.12)
- Middle age groups gave their income much more often than the oldest and youngest age groups
- When the household size increases, the item nonresponse declines
- Income was most difficult to get from ‘Never married’ but the differences between marital status groups are not big.
- Happiness and interviewing time are positively related to the income response.

Figure 6.2 Unit response rates by education and survey modes in the Finnish Security Survey (Laaksonen & Heiskanen 2014)

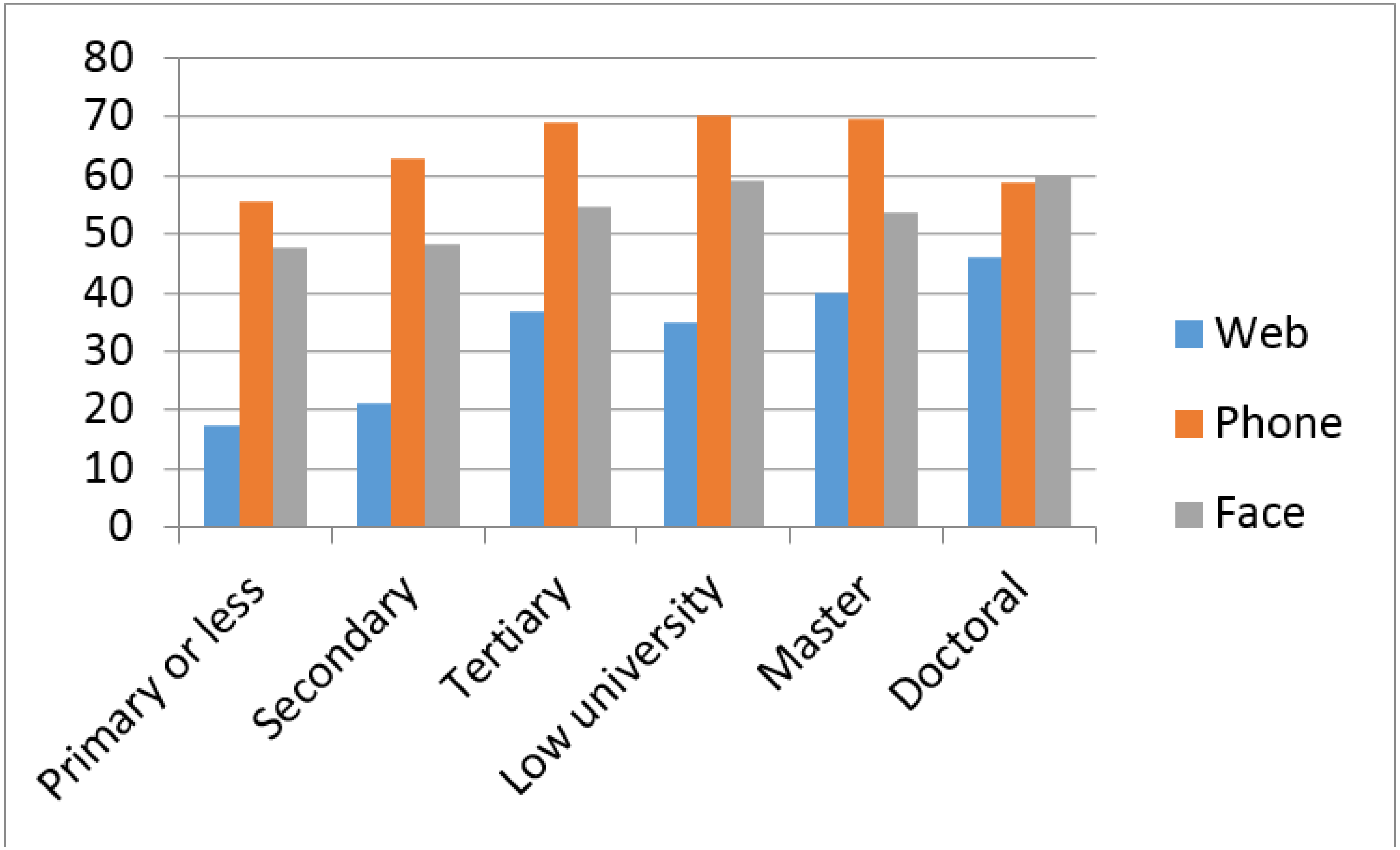


Table 6.3 Probit estimates of the unit response model for the Finnish Security Survey. The following variables are not included: Living Area, Formal Education and Age but the age is in the Figure 6.3

Auxiliary Variable	*Web*	*Phone*	*Face-to-face*
Male vs. female	-0.0565	-0.0256	0.1038
Native language			
Finnish vs. Russian	0.2599	0.2319	-0.1071
Other vs. Russian	-0.1135	-0.1085	-0.3186
Partnership			
Widowed vs. old marriage	-0.2224	0.1060	-0.1071
Single vs. old marriage	-0.3414	-0.2130	-0.3115
Many vs. old marriage	-0.1236	-0.2956	-0.1181
Recent vs. old marriage	-0.3011	-0.0946	-0.2245

A figure that is very illustrative. Four age variables are used in estimation: Age, age-squared, age-cubed, and age with power four

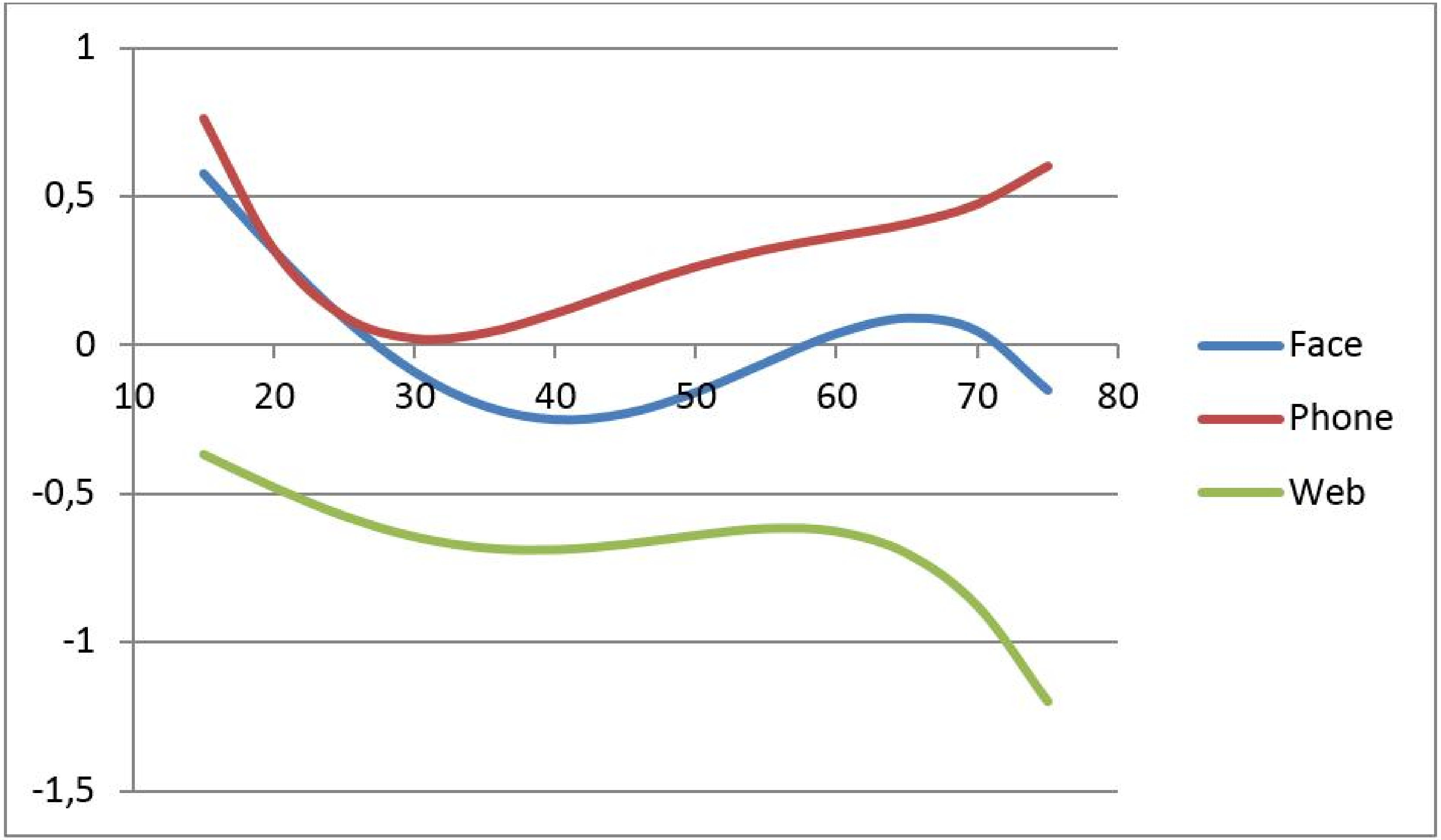


Figure 6.3. Probit estimates by age for three survey modes, The Finnish Security Survey (Laaksonen & Heiskanen 2014).

The response propensity model is applied in Chapter 7 for weighting adjustments. Now the predicted response probabilities, called response propensities, have a big role.