Università degli Studi di Palermo

AREA RICERCA E TRASFERIMENTO TECNOLOGICO
SETTORE DOTTORATI E CONTRATTI PER LA RICERCA
U. O. DOTTORATI DI RICERCA

Corso di Dottorato in Scienze Economiche e Statistiche
Dipartimento di Scienze Economiche, Aziendali e Statistiche
Settore Scientifico Disciplinare: SECS-P/05-ECONOMETRICS

An analysis of nonresponse errors in the Survey of Health, Ageing and Retirement in Europe (SHARE)

IL DOTTORE
MOSLEM RASHIDI

IL COORDINATORE
VITO MICHELE ROSARIO MUGGEO

IL TUTOR
GIUSEPPE DE LUCA

CICLO XXXV.
ANNO CONSEGUIMENTO TITOLO 2023.

An analysis of nonresponse errors in the Survey of Health, Ageing and Retirement in Europe (SHARE)

University of Palermo

Moslem Rashidi

September 2023

Acknowledgements

I would like to express my heartfelt gratitude to the following individuals and institutions for their invaluable support and assistance throughout the journey of completing this thesis:

First and foremost, I am profoundly thankful for the exceptional guidance, expertise, and unwavering support provided by my thesis advisor, Professor Dr. Giuseppe De Luca. Their mentorship and insightful feedback have been instrumental in shaping this research. Professor Dr. Giuseppe De Luca delved into the intricate domain of econometrics and broadened my perspective, offering invaluable insights that significantly contributed to this work. His limitless patience, allowing me to learn from my mistakes and grow as a researcher, is deeply appreciated.

Reviewers: Professor Dr. Omar Paccagnella from the Department of Statistical Sciences at the University of Padua and Professor Dr. Danilo Cavapozzi from the Department of Economics at Ca' Foscari University of Venice. I appreciate my thesis committee members for their valuable insights, constructive critiques, and thoughtful recommendations that greatly improved the quality of this work.

I would also like to express my sincere appreciation to Professor Dr. Consiglio and Professor Dr. Muggeo, Ph.D. directors, whose timely assistance proved invaluable when needed. Additionally, my gratitude extends to all the professors at the Department of Economics and Statistics at Palermo University. Attending their lectures enriched my knowledge and was instrumental in my academic journey.

Beyond academia, my deepest sense of gratitude goes to my colleagues. Sharing these years' experiences with them has been an incredible privilege, and they are, without a doubt, the best companions one could wish for.

I want to acknowledge my fiancée, whose unwavering inspiration throughout my Ph.D. program was a constant source of motivation. I sincerely thank my parents and my family for their selfless support and encouragement.

Thank you.

Moslem Rashidi

September, 2023

Contents

1 Preface **1**

2 Weights and imputations in SHARE Wave 8 **2**
1 Introduction . . . 3
2 Composition of the Sample in Wave 8 . . . 3
3 Calibrated weights . . . 6
4 Calibration procedure . . . 7
5 Calibrated cross-sectional weights for the CAPI subsample . . . 9
6 Calibrated longitudinal weights for the CAPI subsample . . . 10
7 Supplementary material and user guide on calibrated weights . . . 11
8 Imputations of missing values in the CAPI data . . . 12
9 Hot-deck imputations . . . 12
10 FCS imputations . . . 13
11 Appendix . . . 18
11.1 Appendix A: Tables . . . 18

3 Effects of interviewers on response to income and wealth items **25**
1 Introduction . . . 25
2 SHARE data . . . 27
2.1 Main survey data of SHARE wave 6 . . . 28
2.2 Interviewer survey data of SHARE wave 6 . . . 31
3 Choice of predictors and missing data patterns . . . 33
3.1 Regressor of interest . . . 33
3.2 Control variables . . . 36
3.3 Missing data patterns. . . . 38
4 Methodology . . . 40
4.1 Complete-case analysis . . . 40
4.2 Fill-in approach . . . 41
4.3 Generalized missing-indicator (GMI) and model averaging (MA) . . . 43
5 Results . . . 46
6 Conclusions . . . 47
7 Appendix . . . 48

7.1 Appendix A: Tables . . . 48
7.2 Appendix B: Technical questions in regular and interviewer SHARE wave 6 questionnaires . . . 52

4 Analysis response propensity score using WALS in SHARE-HCAP 59
1 Introduction . . . 53
2 SHARE-HCAP Data and sample design . . . 55
2.1 SHARE-HCAP data . . . 55
3 Weighted-Average Least Squares (WALS) estimator . . . 59
3.1 Statistical framework . . . 59
3.2 One-step ML estimators . . . 61
3.3 WALS estimation of linear regression models . . . 63
3.4 Implications for WALS . . . 66
4 Choice of focus and auxiliary regressors . . . 68
5 Results . . . 69
5.1 Estimation of Marginal Effects . . . 70
5.2 Predicted response probabilities . . . 71
6 Conclusion . . . 73
7 Appendix . . . 75
7.1 Appendix A: Tables . . . 75
7.2 Appendix B: Figures . . . 80

List of Tables

2.1 Number of individual interviews of Wave 8 by country and type of interview 4
2.2 Number of household interviews of Wave 8 by country and type of interview 5
2.3 Gender-Age National Calibration Margins for the Calibrated Cross-sectional Weights of Wave 8 . . . 19
2.4 Gender-Age National Calibration Margins for the Calibrated Cross-sectional Weights of Wave 8 . . . 20
2.5 Gender-Age National Calibration Margins for the Calibrated Cross-sectional Weights of Wave 8 . . . 21

3.1 Response rate of answers on the financial variables of SHARE wave 6 in the eligible respondent's sample . . . 30
3.2 Number of interviewers, number of participants, and participation rate to the interviewer survey (IWS) of SHARE wave 6 by country . . . 31
3.3 Definitions and summary statistics of the control variables in the eligible respondent's sample . . . 37
3.4 Complete-case subsample and missing data patterns by country . . . 39
3.5 Estimates of the average marginal effects of interviewer's expected response rate over complete-case subsample by country through CCA approach . . . 48
3.6 Estimates of the average marginal effects of interviewer's expected response rate over complete-case subsample by country through FI_MI approach . . . 49
3.7 Estimates of the average marginal effects of interviewer's expected response rate over complete-case subsample by country through BBMA_BIC approach 50
3.8 Estimates of the average marginal effects of interviewer's expected response rate over complete-case subsample by country through BBMA_AIC approach 51

4.1 Number of respondents, participants, and participation rates in the gross sample SHARE-HCAP by country . . . 57
4.2 Description of statistics of explanatory variables in response and nonresponse samples . . . 58
4.3 Marginal effects of key socio-demographic variables and cognitive ability variables on the participation probability in CZ . . . 75
4.4 Marginal effects of key socio-demographic variables and cognitive ability variables on the participation probability in DE . . . 76

4.5 Marginal effects of key socio-demographic variables and cognitive ability variables on the participation probability in DK 77
4.6 Marginal effects of key socio-demographic variables and cognitive ability variables on the participation probability in FR 78
4.7 Marginal effects of key socio-demographic variables and cognitive ability variables on the participation probability in IT 79

List of Figures

2.1 NUTS1 Population Margins for the Calibrated Cross-sectional Weights of Wave 8 (Millions of People) . . . 10
2.2 Item non-response rates for value of the house and amount in bank account by country . . . 13

3.1 Distribution of expectations of the interviewer on response rate to income by country . . . 34
3.2 the country-level response rate and the country-level average interviewers' confidence level . . . 35

4.1 Predicted response probability across different ages and selected countries . . . 80
4.2 Predicted response probability across different ages and selected countries in three different word recall scores groups . . . 81
4.3 Predicted response probability across different ages and selected countries in two different memory test scores groups . . . 82
4.4 Predicted response probability across different ages and selected countries in two different numeracy 1 test scores groups . . . 83
4.5 Predicted response probability across different ages and selected countries in two different numeracy 2 test scores groups . . . 84
4.6 Predicted response probability across different ages and selected countries in two different orientations in time scores groups . . . 85
4.7 Predicted response probability across different fluency scores and selected countries . . . 86
4.8 Predicted response probability across different fluency scores and selected countries in three different word recall scores groups . . . 87
4.9 Predicted response probability across different fluency scores and selected countries in two different memory scores groups . . . 88
4.10 Predicted response probability across different fluency scores and selected countries in two different numeracy 1 scores groups . . . 89
4.11 Predicted response probability across different fluency scores and selected countries in two different numeracy 2 scores groups . . . 90
4.12 Predicted response probability across different fluency scores and selected countries in two different orientations in time scores groups . . . 91

Chapter 1

Preface

The challenges posed by unit and item nonresponse are pervasive in the survey research. Unit nonresponse denotes the failure of a sample unit to participate in the survey as a whole, while item nonresponse signifies the failure of a unit respondent to answer one or more survey items for which they are eligible. Nonresponses create doubt about the reliability of survey data and introduce errors that undermine the accuracy of statistical conclusions drawn from that data. Even increasing the sample size cannot eliminate nonsampling errors that undermine the representativeness of sample parameters in relation to their population parameter. Survey researchers have long grappled with these formidable challenges, resorting to weightings for unit nonresponse and imputation techniques for item nonresponse as indispensable tools in their methodological arsenal. These techniques aim to alleviate the bias introduced by nonresponse errors, ensuring that the resulting estimates remain reliable and valid.

In this thesis, consisting of three interconnected papers, we embark on a journey to unravel the intricacies of nonresponse errors within the context of the Survey of Health, Ageing, and Retirement in Europe (SHARE). The aim of this thesis is to analyze and handle missing data in different contexts comparatively and effectively and with new methods.

The first chapter of this thesis is dedicated to a rigorous exploration of weighting and imputation methods employed to minimize the impact of unit and item nonresponse errors. We will use these methods in constructing weights and generating imputations that reduce nonsampling errors within the SHARE Wave 8 dataset.

The second chapter investigates whether the interviewer's characteristics affect the response probabilities of respondents. We will also delve into a comparative analysis of various approaches to address missing covariate data issues, such as complete-case analysis, fill-in, and generalized missing indicators approaches. Notice that we will focus on their application in estimating item nonresponse concerning sensitive financial information, namely income and assets, within the SHARE Wave 6 dataset.

In the final chapter, we diverge from conventional paths and introduce a novel technique—the weighted average Least Squares (WALS) model averaging—to address nonresponse errors and analyze response propensity scores. WALS offers a powerful means to address the model uncertainty associated with constructing response propensity scores.

Chapter 2

Weights and imputations in SHARE Wave 8

Abstract

The paper describes the weighting and imputation strategies employed to tackle issues of unit non-response, sample attrition, and item non-response in Wave 8 of the Survey of Health, Ageing and Retirement in Europe (SHARE). We describe the procedure used to construct calibrated cross-sectional and longitudinal weights for addressing issues arising from unit nonresponse and attrition in the CAPI subsample. Subsequently, we describe the model used to obtain multiple imputations of the missing values due to item nonresponse in the CAPI data.

Keywords: unit non-response, item non-response, calibrated weighting, multiple imputation

1 Introduction

This chapter provides a description of the weighting and imputation strategies used for dealing with problems of unit non-response, sample attrition and item non-response in the eighth wave of the Survey of Health, Ageing and Retirement in Europe (SHARE). As discussed in the previous chapters (i.e., SHARE Wave 8 methodology book), the data collection process of Wave 8 was suddenly interrupted in March 2020 by the COVID-19 outbreak and the subsequent lockdowns enforced by the national governments of the various countries. SHARE reacted promptly to this deep pandemic shock through the design of a special COVID-19 questionnaire, which was fielded between June and July 2020. We expect that the data collected in the regular Wave 8 will become an extraordinary source of information for studying health and socio-economic implications of the shock for the elderly population. To best exploit the available data, it is important for the user to have a basic understanding of the fieldwork rules adopted for the standard interview and the specific COVID-19 interview of Wave 8, the different types of non-response errors that occurred in the implementation of these two interview instruments and the basic strategies adopted to cope with these errors. In the following, we first use the different patterns of participation to define three subsamples of primary interest for the analysis of the data collected in Wave 8: CAPI, CATI and CAPI & CATI. We then describe the procedure used to construct calibrated cross-sectional and longitudinal weights for handling problems of unit non-response and attrition in the CAPI subsample. Afterward, we describe the model used to obtain multiple imputations of the missing values due to item non-response in the CAPI data. The construction of calibrated weights and multiple imputations for the CATI data is discussed in Chapter 11 of the SHARE Wave 8 methodology book.

2 Composition of the Sample in Wave 8

The data collection process of Wave 8 started regularly in October 2019 by means of a face-to-face Computer-Assisted Personal Interview (CAPI) administered in 28 countries. As usual, the sample in Wave 8 consisted of a longitudinal subsample and a refreshment subsample. The longitudinal subsample includes all respondents already interviewed in any previous wave of the study. The refreshment subsample, on the other hand, includes the new sample units drawn in Wave 8 to maintain the representation of the younger cohorts of the target population that were not age-eligible in the previous waves (i.e. people born between 1967 and 1969) and to compensate for the reduction of sample size due to attrition across waves of the SHARE panel.

The fieldwork activities of Wave 8 were suddenly interrupted in March 2020 due to the COVID-19 outbreak. To study the impact of the pandemic on the health and socio-economic conditions of SHARE respondents, a new COVID-19 questionnaire was promptly fielded between June and July 2020 by a Computer-Assisted Telephone Interview (CATI). By design, this new survey instrument was administrated to the longitudinal part of the sample

only (not to the refreshment sample). Tables 2.1 and 2.2 provide, respectively, a breakdown of the number of individual interviews and the number of household interviews by country and type of interview (CAPI and/or CATI) based on SHARE Wave 8, Release 0 as well as SHARE Wave 8, Release 0.0.1 beta (Börsch-Supan, 2020a, 2020b). In total, 23 per cent of respondents answered the CAPI only, 28 per cent answered the CATI only, and 49 per cent answered both the CAPI and CATI instrument. For the type of data collected in Wave 8 one can then distinguish three subsamples of primary interest: CAPI, CATI and CAPI & CATI. The CAPI subsample consists of 51,018 respondents in 35,914 households who have answered the CAPI questionnaire irrespective of whether they have also answered the CATI questionnaire. The CATI subsample consists of 54,600 respondents in 37,222 households who have answered the CATI irrespective of whether they have also answered the CAPI. The CAPI & CATI subsample consists of 34,916 respondents in 24,191 households who have answered both interviews.

Table 2.1: Number of individual interviews of Wave 8 by country and type of interview

Country	CAPI only	CATI only	CAPI & CATI	Total CAPI	Total CATI
AT	607	1,204	1,265	1,872	2,469
BE	518	2,095	1,687	2,205	3,782
BG	170	171	640	810	811
CH	364	246	1,640	2,004	1,886
CY	123	416	374	497	790
CZ	884	579	2,040	2,924	2,619
DE	1,406	378	2,278	3,684	2,656
DK	854	530	1,453	2,307	1,983
EE	577	1,836	2,706	3,283	4,542
ES	961	1,037	1,011	1,972	2,048
FI	119	457	1,006	1,125	1,463
FR	1,189	316	1,727	2,916	2,043
GR	184	1,039	2,595	2,779	3,634
HR	862	961	1,048	1,910	2,009
HU	666	513	483	1,149	996
IL	640	763	687	1,327	1,450
IT	171	1,860	1,846	2,017	3,706
LT	269	179	1,086	1,355	1,265
LU	193	202	726	919	928
LV	490	322	656	1,146	978
MT	104	200	628	732	828
NL	1,400	276	504	1,904	780
PL	1,055	1,300	1,628	2,683	2,928
PT	0	1,118	0	0	1,118
RO	73	378	1,101	1,174	1,479
SE	1,367	238	1,121	2,488	1,359
SI	819	983	2,129	2,948	3,112
SK	37	87	851	888	938
Total	16,102	19,684	34,916	51,018	54,600

Note. SHARE Wave 8, Release version: 0.

Table 2.2: Number of household interviews of Wave 8 by country and type of interview

Country	CAPI only	CATI only	CAPI & CATI	Total CAPI	Total CATI
AT	467	843	913	1,380	1,756
BE	396	1,543	1,258	1,654	2,801
BG	130	111	437	567	548
CH	275	135	1,236	1,511	1,371
CY	72	249	270	342	519
CZ	602	414	1,442	2,044	1,856
DE	1,120	241	1,515	2,635	1,756
DK	614	379	1,073	1,687	1,452
EE	423	1,260	1,981	2,404	3,241
ES	723	665	679	1,402	1,344
FI	95	268	684	779	952
FR	881	230	1,254	2,135	1,484
GR	152	666	1,680	1,832	2,346
HR	600	598	670	1,270	1,268
HU	475	336	335	810	671
IL	482	506	490	972	996
IT	125	1,149	1,160	1,285	2,309
LT	193	121	786	979	907
LU	152	123	513	665	636
LV	355	226	471	826	697
MT	52	124	388	440	512
NL	950	182	357	1,307	539
PL	696	859	1,090	1,786	1,949
PT	0	725	0	0	725
RO	52	236	719	771	955
SE	1,021	161	819	1,840	980
SI	601	629	1,430	2,031	2,059
SK	19	52	541	560	593
Total	11,723	13,031	24,191	35,914	37,222

Note. SHARE Wave 8, Release version: 0.

The distinction between these three subsamples has important implications for the information available in the analysis of Wave 8 data. Specifically, the CAPI subsample contains the data collected before the COVID-19 outbreak by the regular SHARE questionnaire of Wave 8, and its longitudinal part (about 86 per cent) can be merged with the data collected in one or more previous waves. The CATI subsample contains the data collected after the COVID-19 outbreak by the SHARE Corona Survey and can be fully merged with some of the previous waves of SHARE as it consists of longitudinal respondents only. The CAPI & CATI subsample exploits the full force of the survey instruments implemented in Wave 8 as it contains the data collected before and after the outbreak and can be fully merged with previous waves. As discussed in the next section, the SHARE weights database provides different sets of calibrated cross-sectional weights for the three subsamples. SHARE also provides different sets of imputations for the missing values due to item non-response in the CAPI and CATI data. In this chapter, we shall focus attention on calibrated weights and imputations for the standard CAPI data of Wave 8.

3 Calibrated weights

In the ideal situation of complete response, the availability of design weights allows the users to account for the randomness of the sampling process by compensating for unequal selection probabilities of the various sampling units. Unfortunately, properties of inferential procedures based on the sampling design weights depend on the assumption of a complete survey response, which is almost never satisfied in the practical implementation of surveys. SHARE is not an exception to this common situation. The baseline and refreshment samples of each wave suffer from problems of unit non-response (Groves and Peytcheva, 2008). Moreover, the longitudinal part of the sample is subject to attrition problems (Lynn, 2009). Because of these non-sampling errors, we discourage the users from relying on sampling design weights for standard analyses of the SHARE data. These weights are included in the public release of the SHARE weights database only to favour the implementation and comparison of alternative statistical procedures for handling non-response and attrition errors.

The baseline strategy adopted by SHARE to handle problems of unit non-response and attrition relies on the calibration approach proposed by Deville and Särndal (1992). This approach allows the sample and population distributions of some benchmark variables to be aligned without the need for specifying an explicit model for the non-response mechanism. Under the assumption that the missing data mechanism is missing at random (Rubin, 1987), calibrated weights may help reduce the potential selection bias generated by non-response errors. Thus, unless these sources of non-sampling errors are controlled for in other ways, these are the types of weights that we generally recommend using in standard analyses of the SHARE data. In the remainder of this section, we first discuss the key methodological advantages and limitations of the calibration procedure. Then, we describe the implementation of the calibration procedure for constructing the various types of calibrated cross-sectional and longitudinal weights available in the public release of SHARE Wave 8 data.

4 Calibration procedure

Let $U = \{1, \ldots, i, \ldots, N\}$ be a finite population of N elements, from which a probability sample $s = \{1, \ldots, i, \ldots, n\} \subseteq U$ of size $n \leqslant N$ is drawn according to a probability-based sampling design. Unless otherwise specified, we shall assume that the inclusion probability $\pi_i = \Pr(i \in s)$ is known and strictly positive for all population units. To describe the basic ideas and the key properties of the calibration approach, we consider first the ideal situation of complete response where all units in the sample s agree to participate in the survey. Then, we relax this ideal set-up to describe the key implications of non-response errors for the properties of this weighting method.

The sampling design weights $w_i = \pi_i^{-1}$ are typically used to account for the randomness of the sampling process and the variability of the inclusion probabilities across sample units due to stratification and clustering strategies (additional details can be found in Chapter 2 of the SHARE Wave 8 methodology book). For example, one can estimate the population total $t_y = \sum_{i \in U} y_i$ of a variable of interest y using the Horvitz-Thompson estimator (Horvitz and Thompson, 1952):

$$\hat{t}_y = \sum_{i \in s} w_i y_i. \tag{1}$$

Under the ideal set-up of complete response, this estimator is known to be design unbiased, that is $\mathrm{E}_p(\hat{t}_y) = t_y$, where $\mathrm{E}_p(\cdot)$ denotes the expectation with respect to the sampling design.

Let us assume now that the sampling frame or other external sources such as census data and administrative archives provide supplementary data on a q-vector of categorical auxiliary variables $\mathrm{x}_i = (x_{i1}, \ldots, x_{iq})^\top$ with known population totals $\mathrm{t}_x = \sum_{i \in U} \mathrm{x}_i$. We shall refer to the auxiliary variables x_i as calibration variables and to their population totals t_x as calibration margins. The basic idea of the calibration approach is to determine a set of calibrated weights w_i^* that are as close as possible to the design weights w_i and that satisfy the constraints:

$$\sum_{i \in s} w_i^* x_{ij} = t_{x_j}, \quad j = 1, \ldots, q \tag{2}$$

Thus, given a distance function $G(w_i^*, w_i)$ and the availability of survey data on $(w_i, \mathrm{x}_i^\top : i = 1, \ldots, n)$ and population data on the calibration margins t_x, the aim of the procedure is to determine the calibrated weights w_i^* by minimizing the aggregate distance $\sum_{i \in s} G(w_i^*, w_i)$ with respect to w_i^* subject to the q equality constraints in (2). Under some regularity conditions on the distance function $G(w_i^*, w_i)$ (see Deville & Särndal, 1992), the solution of this constrained optimization problem exists, is unique, and can be written as:

$$w_i^* = w_i \mathrm{F}(\eta_i), \quad i = 1, \ldots, n \tag{3}$$

where $\eta_i = \mathrm{x}_i^\top \lambda$ is a linear combination of the calibration variables x_i, $\lambda = (\lambda_1, \ldots, \lambda_q)^\top$ is the q-vector of Lagrangian multipliers associated with the constraints (2), and $F(\cdot)$ is a

calibration function, which is uniquely determined by the distance function $G(w_i^*, w_i)$.

A key feature of the calibration approach is that many traditional reweighting methods, such as post-stratification, raking, and generalized linear regression (GREG), correspond to special cases of the calibration estimator:

$$\hat{t}_y = \sum_{i \in s} w_i^* y_i, \tag{4}$$

for particular choices of the calibration function $F(\cdot)$ (or, equivalently, of the distance function $G(\cdot, \cdot)$). Deville and Särndal (1992) present various functional forms for $G(w_i^*, w_i)$ and $F(\eta_i)$. The chi-square distance function $G(w_i^*, w_i) = (w_i^* - w_i)^2/2w_i$, which leads to the widely used GREG estimator, has the advantage of ensuring a closed-form solution for the calibrated weights w_i^*. However, this distance function is unbounded, and depending on the chosen set of calibration variables, it may also lead to negative weights. Different specifications of the calibration function may avoid these issues, but the underlying optimization problems may not admit a solution, and the Lagrange multipliers must be computed numerically. In SHARE, we rely on the logit specification of the distance function:

$$G(w_i^*, w_i) \propto \left(\frac{w_i^*}{w_i} - l\right) \ln\left(\frac{w_i^*/w_i - l}{1 - l}\right) + \left(u - \frac{w_i^*}{w_i}\right) \ln\left(\frac{u - w_i^*/w_i}{u - 1}\right),$$

which leads to a calibrated function of the form:

$$F(\eta_i; u, l) = \frac{l(u-1) + u(1-l)exp(a\eta_i)}{u - 1 + (1-l)exp(a\eta_i)},$$

where $a = [(1-l)(u-1)]^{-1}(u-l)$. Unlike other distance functions, these functional forms restrict in advance the range of feasible values for the calibrated weights by suitable choices of the lower bound l and the upper bound u. Specifically, if a solution exists, then it must satisfy the restriction $w_i l \leqslant w_i^* \leqslant w_i u$. As discussed in Deville and Särndal (1992), the effectiveness of the calibrated weights depends crucially on the correlation between the study variable y and the calibration variables x. In the extreme case when y can be expressed as a linear combination of x, it is clear that the calibrated estimator gives an exact estimate of t_y for every realized sample s. Under suitable regularity conditions, the class of calibration estimators satisfies other desirable asymptotic properties. For example, the estimators obtained by alternative specifications of the distance function are asymptotically equivalent to the GREG estimator based on a chi-squared distance function. Thus, in large samples, calibrated weights are robust to arbitrary choices of the calibration function $F(\cdot)$.

Unfortunately, this property does not necessarily extend to the more realistic cases where survey data are affected by non-response errors. Previous studies by Lundström and Särndal (1999) and Haziza and Lesage (2016) suggest that in these cases, alternative specifications of the calibration function $F(\cdot)$ correspond in practice to imposing different parameterizations of the relationship between response and calibration variables. Moreover, statistical properties of calibration estimators depend as usual on the validity of the missing-at-random

assumption. Brick (2013), Molenberghs et al. (2015), Vermeulen and Vansteelandt (2015), and Haziza and Lesage (2016), among others, discuss a variety of robust weighting methods based on a propensity-score approach. One key issue in the implementation of these methods for SHARE is that selection probabilities and auxiliary variables are usually known for the subsample of respondents only.

5 Calibrated cross-sectional weights for the CAPI subsample

The calibrated cross-sectional weights of the CAPI subsample of Wave 8 were computed separately by country to match the size of the national 50+ populations of individuals in 2019. In each country, we used a logit specification of the calibration function $F(\cdot)$ and a set of calibration margins for the size of the target population across the eight gender-age groups, i.e., males and females in the age classes ([50 – 59], [60 – 69], [70 – 79], [80+]), as reported in Table 2.3 in the Appendix.

In 11 countries (Austria, Belgium, Bulgaria, Germany, Hungary, Italy, the Netherlands, Poland, Romania, Spain, and Sweden), we also included an additional set of calibration margins for the size of the 50+ population across 2016 NUTS1 regional areas (Israel is excluded from the figure). Notice that this additional set of calibration margins were ineffective in all countries containing only one NUTS1 region12[1]. In France and Greece, NUTS1 calibration margins were excluded because of inconsistency between sample and population data. In Israel, where no NUTS nomenclature is available, we used an additional set of calibration margins for the Jewish Israeli and Arab Israeli population groups and immigrants from the former USSR. Population data about the calibration margins come from the Central Bureau of Statistics for Israel and from the EUROSTAT regional database for all other countries.

As usual, calibrated cross-sectional weights are computed at the individual level for inference to the target population of individuals and at the household level for inference to the target population of households. At the individual level, we assign an individual-specific weight to each 50+ respondent that depends on the household design weight and the respondents' set of calibration variables (namely, gender, age class and NUTS1 code). At the household level, we assign instead a common calibrated weight to all interviewed household members that depends on the household design weight and the set of calibration variables for all 50+ respondents in that household. By construction, calibrated cross-sectional weights are missing for respondents younger than 50 (i.e. age-ineligible partners of an age-eligible respondent), for those with missing information on the calibration variables and for those with missing sampling design weights (i.e. respondents from households for which we do not have sampling frame information).

[1] That is the case in Croatia, Cyprus, the Czech Republic, Denmark, Estonia, Finland, Latvia, Lithuania, Luxembourg, Malta, Slovakia, Slovenia and Switzerland.

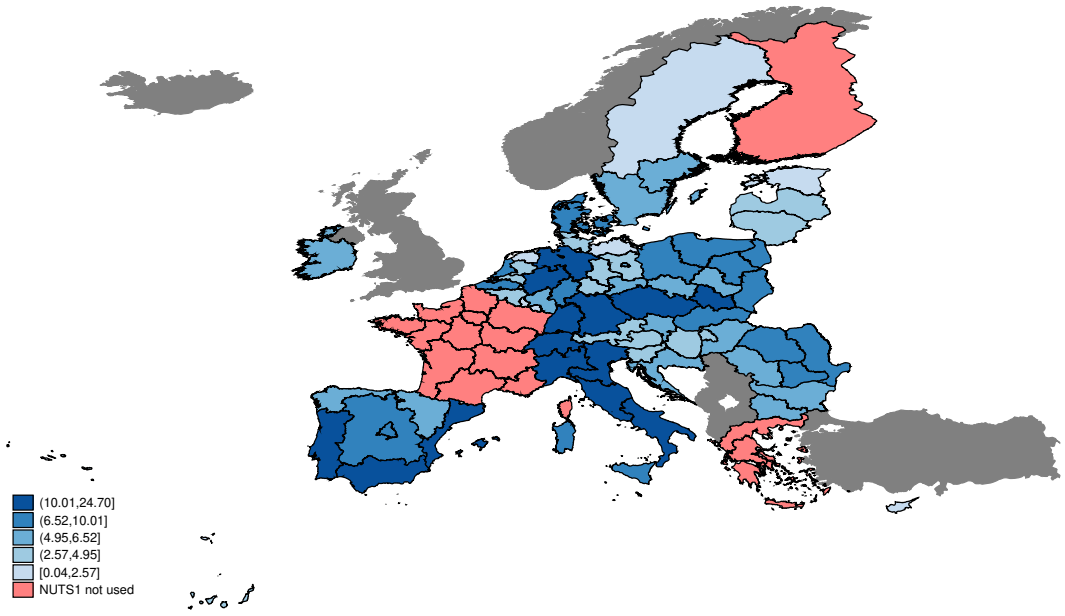

Figure 2.1: NUTS1 Population Margins for the Calibrated Cross-sectional Weights of Wave 8 (Millions of People)

6 Calibrated longitudinal weights for the CAPI subsample

In addition to calibrated cross-sectional weights, SHARE Wave 8 Release 8.0.0 also includes calibrated longitudinal weights for the purposes of panel data analyses. Although calibration relies on the same procedure, calibrated longitudinal weights differ from calibrated cross-sectional weights in two important respects. First, the calibrated longitudinal weights are defined only for the balanced subsample of respondents who have participated in at least two waves of the study. Second, since mortality is a source of attrition that affects both the sample and the population, calibrated longitudinal weights account for the mortality of the target population across waves. In other words, the target population for panel data analysis is defined as the target population at the beginning of a reference time period that survives up to the end of the period considered (see, for example, Lynn, 2009).

To simplify the structure of the public release of the data, we provide calibrated longitudinal weights only for selected wave combinations of the SHARE panel. Those available in Release 8.0.0 are the seven possible couples of any two adjacent waves (namely, the wave combinations 1 – 2, 2 – 3, 3 – 4, 4 – 5, 5 – 6, 6 – 7 and 7 – 8) and the fully balanced panel (i.e. the wave combination 1 – 2 – 3 – 4 – 5 – 6 – 7 – 8). The weights of the generic wave combination t – ... – s are computed separately by country to represent the national 50+ population of Wave t that survives up to the interview year of Wave s. For example, the wave combination 1 – 2 allows the population of people aged 50+ in 2004 that survived up

to 2006 to be represented, while the fully balanced panel allows the population of people aged 50+ in 2004 that survived up to 2019 to be represented.

For the calibrated longitudinal weights of two adjacent waves, we use a logit specification of the calibration function $F(\cdot)$ and a set of calibration margins for the size of the target population across eight gender-age groups (i.e. males and females whose ages at the time of the starting wave were in the four classes [50 – 59],[60 – 69], [70 – 79] and [80+]). Compared to calibrated cross-sectional weights, we do not control for NUTS1 calibration margins due to the smaller number of observations available in the national longitudinal subsamples. Moreover, we account for the mortality of the target population by subtracting from each calibration margin the corresponding number of deaths that occurred between the interview years of Wave t and Wave s. Table 2.4 in the Annex provides the population margins used to compute the calibrated longitudinal weights of the wave combination 7 – 8. Population margins for the calibrated longitudinal weights of the other wave combinations can be found in De Luca and Rossetti (2019a, Tables A.3 – A.8).

For the calibrated longitudinal weights of the fully balanced panel, we further restricted the set of calibration margins to six gender-age groups (i.e. males and females whose ages in 2004 were in the three classes [50 – 59],[60 – 69] and [70+]). Table 2.5 in the Appandix shows the population margins used to construct the longitudinal weights of the fully balanced panel.

As with calibrated cross-sectional weights, calibrated longitudinal weights are available both at the individual level and at the household level. For the individual weights, the balanced sample consists of respondents interviewed in each wave of the selected wave combination. For the household weights, the balanced sample consists instead of households with at least one eligible member interviewed in each wave of the selected wave combination. Note that, according to these definitions, the balanced sample of households is larger than the balanced sample of individuals. For example, couples with one partner participating in Wave 7 and the other partner participating in Wave 8 belong to the balanced sample of households for the wave combination 7-8, even if neither of the two partners belongs to the corresponding balanced panel of individuals.

7 Supplementary material and user guide on calibrated weights

Since the SHARE panel now consists of eight waves, one can compute many different types of calibrated longitudinal weights depending on the selected combination of waves and the selected unit of analysis (either individuals or households). In addition, one can compute many different types of calibrated cross-sectional weights for specific subsamples of the data collected in each wave (e.g. the respondents to the vignette questionnaires of Waves 1 and 2 or the drop-off questionnaires of Waves 1 to 8). These considerations make it clear why the strategy of providing all possible calibrated cross-sectional and longitudinal weights is not feasible, especially in the future when additional waves will be available. For cross-

sectional studies based on specific subsamples and longitudinal studies based on other wave combinations, users are required to control for the potential selection effects of unit non-response and attrition by computing their own calibrated weights or by implementing some alternative correction method.

To support users in this non-trivial methodological task, we provide a set of Stata do-files and ado-files that illustrate step by step how to compute calibrated cross-sectional and longitudinal weights. In addition, we provide one data set with updated information on population size and number of deaths by year, gender, age and NUTS1 regions. Registered users can download this supplementary material on calibrated weights from the SHARE Research Data Center dissemination website (https://releases.sharedataportal.eu/releases), under the link "Generate Calibrated Weights Using Stata (2018)". A discussion of the step-by-step operations can also be found in the SHARE Technical Report "Computing Calibrated Weights in Stata" (De Luca and Rossetti, 2019b).

8 Imputations of missing values in the CAPI data

Imputations of the missing values due to item non-response errors in the regular face-to-face interview of Wave 8 were constructed using the same procedure adopted in the previous regular waves of SHARE (see, for example, De Luca *et al.* 2015). Of course, we adapted the imputation model to the specific features of the regular Wave 8 interview in terms of branching, skip patterns, proxy interviews, country-specific deviations from the generic version of the questionnaire and availability of partial information from the sequence of unfolding bracket questions. However, we also attempted to preserve as much as possible the comparability of the imputations across different waves of the SHARE panel. The imputation procedure is essentially based on either the hot-deck method or the fully conditional specification (FCS) method depending on the prevalence of missing values for the variables collected in the regular interview of Wave 8.

9 Hot-deck imputations

In SHARE, we always used the hot-deck method for variables affected by negligible fractions of missing values (usually, much less than 5 per cent of the respondents eligible to answer a specific item on the CAPI questionnaire). The hot-deck method consists of replacing the missing values in one or more variables for a non-respondent (called the recipient) with the observed values in the same variables obtained from a respondent (called the donor) who is "similar" to the recipient according to some metric (see, for example, Andridge and Little, 2010).

In Wave 8, we computed hot-deck imputations in an early stage, separately by country, and according to a convenient order that accounts for branching and skip patterns included in the various modules of the CAPI questionnaire. For each variable imputed through this method, we select the donors randomly from imputation classes determined by auxiliary

variables that are observed for both donors and recipients. We imputed first basic socio-demographic characteristics such as age and education, which contained very small fractions of missing values. These characteristics were then used as auxiliary variables to impute the missing values in the other variables. Our baseline set of auxiliary variables consisted of country, gender, five age classes ([– 49], [50 – 59], [60 – 69],[70 – 79], [80+]), five groups for years of education and two groups for self-reported good/bad health. For some variables, we exploited a larger set of auxiliary variables. For example, we also used the number of children to impute the number of grandchildren and an indicator for being hospitalised overnight during the last year to impute other health-related variables. Variables that are known to be logically related, such as respondent's weight, height and body mass index, were imputed jointly.

10 FCS imputations

In the second stage of the imputation procedure, we dealt with the more worrisome issue of item non-response in monetary variables, such as income from various sources, assets and consumption expenditures, which were typically collected by retrospective and open-ended questions that are sensitive and difficult to answer precisely (see Figure 2.2). Figure 2.2

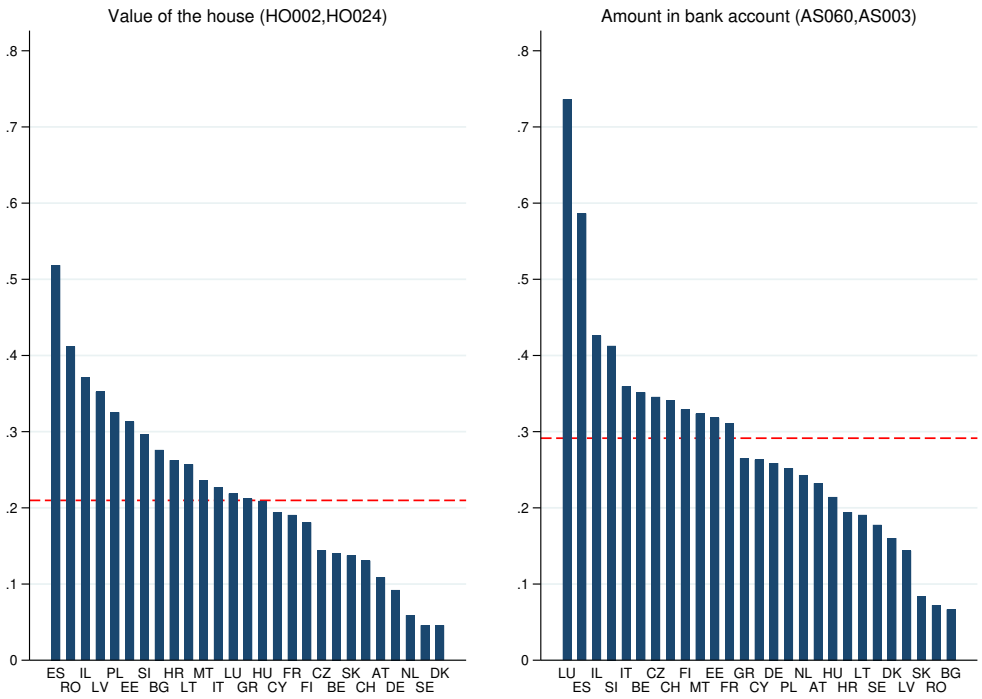

Figure 2.2: Item non-response rates for value of the house and amount in bank account by country

shows the item non-response rates of two monetary variables: value of the house (HO002, HO024) and amount in bank account (AS060, AS003). For the first variable, the percentage of missing values among the eligible respondents ranges from a minimum of 5 per cent in Denmark and Sweden to a maximum of 52 per cent in Spain (21 per cent on average). The percentage of missing values becomes even more dramatic for questions that are likely to be very sensitive for the respondents. For example, the financial respondent was asked "Do you (or your husband/wife/partner) currently have a bank account, or transaction account, or saving account or postal account?" (AS060) and then "About how much do you (and your husband/wife/ partner) currently have in bank accounts, transaction accounts, saving accounts or postal accounts?" (AS030). In 12 out of 27 countries participating in Wave 8, more than 30 per cent of the eligible respondents either refused or did not know how to answer these two questions. The unweighted cross-country average of the item non-response rate is equal to 29 per cent.

In the current body of research, two main methods for addressing missing data in multivariate imputation with arbitrary missing-data patterns are the joint modeling (JM) approach and the fully conditional specification (FCS) approach.

The JM approach assumes a genuine multivariate distribution for all imputation variables, making it particularly suitable for cases where specific arbitrary missing-data patterns can be identified. Complex mathematical methods, such as various Markov Chain Monte Carlo (MCMC) techniques, are often employed to impute values. For instance, the MVN method (Schafer 1997) is a notable JM approach that assumes a multivariate normal distribution for the data. While JM is feasible for simpler cases, especially when data can be reasonably modeled using a multivariate normal distribution, its practical application may be limited for more complex data structures. Nevertheless, it does offer a stronger theoretical basis, as imputed values are derived from a genuine multivariate distribution.

Since Wave 1, we have handled these large fractions of missing values with the fully conditional specification (FCS) method of van Buuren et al. (1999). The FCS method uses a Gibbs sampling algorithm, which imputes multiple variables jointly and iteratively through a sequence of regression models. Assume we want to impute arbitrary patterns of missing values on a set of J variables. The basic idea of the FCS method is that, at each step of the iterative process, we impute the missing values on the j-th variable (j=1,. . . ,J) by drawing from the predictive distribution of a regression model that includes as predictors the most updated imputations of the other J – 1 variables (as well as other fully observed predictors). The process is applied sequentially to the whole set of J variables and is repeated in a cyclical manner by overwriting at each iteration the imputed values computed in the previous iteration. Despite a lack of rigorous theoretical justification (see, for example, Arnold *et al.* 1999, 2001; van Buuren, 2007), the FCS method has become one of the most popular multivariate imputation procedures due to its flexibility in handling complicated data structures and its ability to preserve the correlations of the imputed variables (Raghunathan *et al.* 2001; van Buuren *et al.* 2006). Comparisons of the FCS method with other multivariate imputation techniques can be found in Lee and Carlin (2010).

In Wave 8, we computed FCS imputations separately by country and household type.

The household types considered were singles and third respondents (sample 1), couples with both partners interviewed (sample 2), and all couples with and without a non-responding partner (sample 3). The distinction between the first two samples was primarily motivated by the fact of using socio-demographic characteristics of the partner of the designed respondent as additional predictors to impute the missing monetary amounts within couples. The overlapping partition of the last two samples was instead motivated by the need to impute properly total household income in the couples with a non-responding partner.

The set of monetary variables imputed jointly in the Gibbs sampling algorithm was country- and sample-specific as we required a minimum number of donor observations for estimating the regression model associated with each variable[2]. Variables that did not satisfy this requirement were imputed first (either by hot-deck or by regression imputations) and then used as fully observed predictors for computing the FCS imputations of missing values in the other monetary variables.

The imputation of each monetary variable was typically based on a two-part model that involved a probit model for ownership and a linear regression model for the amount conditional on ownership[3]. Depending on eligibility and ownership, we converted (if needed) non-zero values of monetary variables in annual euro amounts to avoid modelling differences in the time reference periods of the various variables and the national currencies of non-euro countries. In an early stage of the imputation process, we also symmetrically trimmed 2 per cent of the complete cases from the country-specific distribution of annual euro amounts to exclude (and then impute) outliers that may have a large influence on survey statistics. Moreover, we applied logarithm or inverse hyperbolic sine transformations to reduce skewness in the right tails of the conditional distribution of each monetary variable[4].

The set of fully observed predictors was also sample-specific. For singles and third respondents (sample 1), it included gender, age, years of education, self-perceived health, number of children, number of chronic diseases, score of the numeracy test, employment status and willingness to answer (as perceived by the interviewer in the IV module of the CAPI instrument). For couples with both partners interviewed (sample 2), we added a similar set of predictors for the partner of the designed respondent. For couples with a non-responding partner (those remaining in sample 3 after excluding the couples in sample 2), we restricted the additional set of predictors referring to the non-responding partner to age and years of education only[5].

Imputations of the monetary amounts were always constrained to fall within individual-level bounds that incorporated the partial information available on the missing observations

[2] The minimum number of observations was equal to 100 in sample 1 and 150 in samples 2 and 3.

[3] For the few variables without an ownership question, such as food at home expenditure (CO002) and total household income (HH017), we used a simple linear regression model.

[4] We apply the log transformation to variables with a positive support and the inverse hyperbolic sine transformation to variables that may take negative values (e.g. income from self-employment, bank account and value of own business).

[5] In the few cases where the number of donor observations available in the estimation step was lower than 30, we employed a smaller subset of predictors, namely gender, age, years of education and self-reported health.

(e.g., country-specific thresholds used to trim outliers in the tails of the observed distribution of each monetary variable, bounds obtained from the sequence of unfolding bracket questions asked by design to non-respondents of open-ended monetary variables and lower bounds based on the observed components of aggregated monetary variables).

As usual, the imputation of total household income received particular attention because the CAPI questionnaire provides two alternative measures of this variable. The first measure (thinc) can be obtained by a suitable aggregation at the household level of all individual income components, while the second (thinc2) can be obtained via the one-shot question on monthly household income (HH017). As discussed in De Luca *et al.* (2015), it is not easy to find strong arguments to prefer one measure over the other. Moreover, the availability of two alternative measures may greatly improve the imputation process because each measure could contribute relevant information on the missing values of the other measure. Specifically, to avoid understating the first measure of total household income in couples with a non-responding partner, we adopted the following three-stage algorithm:

Stage 1. For singles and third respondents (sample 1), we imputed all monetary variables by the FCS method discussed before. At the end of each iteration of the Gibbs sampling algorithm, we also computed total household income (thinc), household net worth (hnetw) and total household expenditure (thexp) by suitable aggregations of the imputed income, wealth and expenditure items. Next, we imputed the second measure of total household income (thinc2) using the first measure of total household income (thinc), household net worth (hnetw), total household expenditure (thexp) and socio-demographic characteristics of the household respondent as predictors. The imputed values of thinc2 were constrained to fall in the bounds derived from the sequence of unfolding bracket questions for the variable HH017.

Stage 2. For couples with both partners interviewed (sample 2), the imputation strategy is similar to that adopted in stage 1 for the sample of singles and third respondents (sample 1). The main difference is that in each iteration of the Gibbs sampling algorithm we employed a larger set of predictors that also included socio-demographic characteristics and the most updated imputations of the monetary variables of the partner of the designed respondent.

Stage 3. Imputed values of all monetary variables for the subsample of couples with both partners interviewed were obtained in stage 2. In stage 3, these couples were included in the imputation sample only as donor observations to impute the missing values in monetary variables for the remaining subsample of couples with a non-responding partner. In this case we imputed first all monetary variables for the responding partners using the FCS method. Unlike stage 2, the predictors referring to the non-responding partner now consisted, however, of age and years of education only. At the end of each iteration of the Gibbs sampling algorithm, we also imputed the second measure of total household income (thinc2) using household net worth (hnetw), total household expenditure (thexp) and socio-demographic characteristics of the responding partner as predictors and bound information obtained from the sequence of unfolding bracket questions for the variable HH017. Finally, we imputed the first measure of total household income (thinc) using the second measure of total household income (thinc2), household net worth (hnetw), total household expenditure (thexp) and

socio-demographic characteristics of the responding partner as predictors, couples with two partners interviewed as donor observations and the imputed sum of individual income sources of the responding partner as a lower bound.

To account for the additional variability generated by the imputation process, we always provide five different imputations of the missing values. Multiple imputations were constructed through five independent replicates of the hotdeck/FCS imputation method. Notice that neglecting this additional source of uncertainty by selecting only one of the five available replicates in the generated imputations module (gv_imputations) may result in misleadingly precise estimates. Convergence of the Gibbs sampling algorithm for FCS imputations was assessed by the Gelman– Rubin criterion (Gelman and Rubin, 1992; Gelman *et al.* 2004) applied to the mean, the median and the 90th percentile of the five imputed distributions of each monetary variable.

11 Appendix

11.1 Appendix A: Tables

Table 2.3: Gender-Age National Calibration Margins for the Calibrated Cross-sectional Weights of Wave 8

Country	Men				Women				Total
	[50-59]	[60-69]	[70-79]	[80+]	[50-59]	[60-69]	[70-79]	[80+]	
AT	692,191	474,962	350,075	161,732	690,277	514,324	429,922	280,785	3,594,268
BE	802,163	648,581	415,223	238,790	792,093	677,454	487,635	408,179	4,470,118
BG	470,525	433,913	273,241	117,795	476,004	520,171	407,681	220,815	2,920,145
CH	646,594	457,734	328,274	168,216	635,284	473,791	377,379	275,436	3,362,708
CY	53,572	45,812	29,742	13,348	54,845	47,747	34,164	18,665	297,895
CZ	666,784	643,683	415,820	145,542	653,390	715,219	557,948	287,365	4,085,751
DE	6,767,896	4,987,359	3,503,497	2,025,017	6,706,270	5,315,052	4,182,432	3,364,089	36,851,612
DK	400,964	325,943	262,429	103,820	396,811	336,986	289,629	159,926	2,276,508
EE	82,421	69,650	39,223	19,117	89,403	92,800	70,564	55,600	518,778
ES	3,427,300	2,509,740	1,734,281	1,068,505	3,486,303	2,703,891	2,084,157	1,812,379	18,826,556
FI	366,458	351,333	246,439	108,055	366,324	373,436	293,480	194,655	2,300,180
FR	4,292,095	3,793,351	2,481,629	1,464,385	4,496,175	4,205,880	2,964,334	2,642,280	26,340,129
GR	714,255	602,074	450,866	312,655	786,535	676,684	545,121	447,779	4,535,969
HR	284,323	264,078	147,496	71,412	297,231	296,359	210,691	146,221	1,717,811
HU	590,879	579,975	324,657	127,425	627,392	732,233	514,932	305,608	3,803,101
IL	406,588	347,274	210,114	109,198	426,100	388,107	252,035	163,941	2,303,357
IT	4,578,610	3,511,037	2,727,000	1,605,281	4,773,621	3,826,173	3,235,533	2,724,793	26,982,048
LT	197,840	143,592	80,935	43,367	227,204	199,804	154,548	118,172	1,165,462
LU	45,729	30,197	17,583	9,021	42,096	29,946	19,751	15,261	209,584
LV	125,437	100,102	57,730	26,854	145,230	140,106	113,267	80,659	789,385
MT	30,110	29,526	19,955	7,912	28,852	29,778	22,730	12,934	181,797
NL	1,258,588	1,038,005	730,336	307,968	1,249,800	1,051,908	791,774	490,852	6,919,231
PL	2,314,260	2,362,249	1,070,816	514,817	2,406,732	2,790,943	1,574,555	1,145,559	14,179,931
PT	696,521	595,393	415,892	236,885	782,400	691,534	548,704	424,571	4,391,900
RO	1,242,261	1,157,669	602,331	316,528	1,233,703	1,389,591	880,666	589,870	7,412,619
SE	651,921	554,220	466,408	207,684	634,895	560,157	497,859	314,449	3,887,593
SI	153,747	136,366	75,445	36,422	149,902	140,115	95,536	74,611	787,533
SK	349,422	315,265	147,993	54,719	358,875	369,921	227,013	124,794	1,948,002

.

Table 2.4: Gender-Age National Calibration Margins for the Calibrated Cross-sectional Weights of Wave 8

Country	Men				Women				Total
	[50-59]	[60-69]	[70-79]	[80+]	[50-59]	[60-69]	[70-79]	[80+]	
AT	641,916	425,171	293,227	101,159	649,968	474,602	377,426	197,514	3,160,983
BE	780,772	600,343	338,619	150,758	781,348	641,318	426,507	287,540	4,007,205
BG	462,860	404,824	210,042	69,239	486,490	518,959	338,629	136,263	2,627,306
CH	609,106	427,735	273,061	107,327	597,391	452,618	330,130	194,427	3,123,558
CY	51,799	42,136	24,034	7,671	53,876	44,947	28,440	11,674	264,577
CZ	655,340	623,733	296,595	87,898	660,149	726,857	435,224	191,820	3,677,616
DE	6,398,988	4,387,725	3,330,805	1,132,381	6,402,957	4,796,985	4,204,401	2,149,719	32,803,961
DK	377,632	320,726	205,745	60,700	376,853	336,264	238,206	105,935	2,022,061
EE	80,083	61,275	32,113	10,799	91,955	88,242	65,943	35,666	466,076
ES	3,186,720	2,273,951	1,459,554	717,612	3,269,944	2,508,630	1,861,721	1,316,017	16,594,149
FI	364,057	348,678	184,520	65,441	368,773	379,871	237,100	134,048	2,082,488
FR	4,179,699	3,649,049	1,936,799	965,970	4,442,788	4,105,991	2,467,897	1,913,064	23,661,257
GR	676,197	567,672	396,558	198,600	755,403	647,373	506,928	292,844	4,041,575
HR	288,285	234,875	122,872	39,536	307,264	277,605	192,420	90,979	1,590,340
HU	569,112	523,511	251,723	76,040	635,270	696,793	445,832	196,556	3,394,837
IL	380,842	323,108	161,092	69,022	406,329	367,274	202,439	109,958	2,020,064
IT	4,271,615	3,392,051	2,313,368	1,000,711	4,497,087	3,747,290	2,915,342	1,902,851	24,040,315
LT	193,340	118,862	71,173	24,963	232,518	180,405	149,832	75,514	1,046,607
LU	41,157	26,281	14,261	5,696	38,514	26,498	17,380	10,754	180,541
LV	125,019	85,017	49,187	14,585	150,639	130,578	108,960	49,985	713,970
MT	29,833	28,068	14,697	4,486	29,610	29,283	17,890	8,377	162,244
PL	2,442,865	2,053,353	784,862	311,989	2,620,141	2,556,668	1,302,070	766,896	12,838,844
PT	678,140	550,935	358,266	143,870	761,523	656,475	500,876	285,679	3,935,764
RO	1,128,197	1,014,797	494,872	181,635	1,196,540	1,285,956	798,850	354,295	6,455,142
SE	612,168	550,185	375,389	128,547	600,384	565,018	417,055	217,611	3,466,357
SI	150,661	121,342	61,937	21,233	148,776	129,297	85,917	50,960	770,123
SK	351,918	271,024	108,079	32,221	372,424	337,740	186,921	79,537	1,739,864

.

Table 2.5: Gender-Age National Calibration Margins for the Calibrated Cross-sectional Weights of Wave 8

Country	Men			Women			Total
	[50-59]	[60-69]	[70+]	[50-59]	[60-69]	[70+]	
AT	395,994	265,326	67,998	444,475	351,754	148,713	1,674,260
BE	557,019	304,036	97,811	602,188	399,757	212,144	2,172,955
CH	423,805	245,054	73,879	446,730	303,050	150,123	1,642,641
DE	4,133,958	3,273,280	748,171	4,516,076	4,205,946	1,525,452	18,402,883
DK	316,423	165,945	36,175	334,147	201,451	74,751	1,128,892
ES	2,046,682	1,238,257	437,780	2,333,482	1,701,382	914,894	8,672,477
FR	3,320,951	1,756,371	694,005	3,775,566	2,342,234	1,519,876	13,409,003
IT	3,112,610	2,106,999	608,107	3,455,659	2,784,107	1,329,301	13,396,783
SE	547,372	306,847	83,886	563,213	356,514	162,161	2,019,993

References

Andridge, R. R., and Little, R. J. A. (2010). A review of hot deck imputation for survey non-response. *International Statistical Review*, 78, 40—64.

Arnold, B. C., Castillo, E., and Sarabia, J. M. (1999). Conditional specification of statistical models. *Springer*.

Arnold, B. C., Castillo, E., and Sarabia, J. M. (2001). Conditionally specified distributions: An introduction. *Statistical Science*, 16, 249–274.

Börsch-Supan, A. (2020a). Survey of Health, Ageing and Retirement in Europe (SHARE) Wave 8. COVID-19 Survey 1. Release version: 0.0.1. beta. SHARE-ERIC. Data set. https://doi.org/10.6103/SHARE.w8cabeta.001

Börsch-Supan, A. (2020b). Survey of Health, Ageing and Retirement in Europe (SHARE) Wave 8. Release version: 0. SHARE-ERIC. Preliminary data set.

Brick, J. M. (2013). Unit nonresponse and weighting adjustments: A critical review. *Journal of Official Statistics*, 29, 329–353.

De Luca, G., Celidoni, M., and Trevisan, E. (2015). Item nonresponse and imputation strategies in SHARE Wave 5. In F. Malter and A. Börsch-Supan (Eds.), SHARE Wave 5: Innovations and methodology (pp. 85-100). MEA, Max Planck Institute for Social Law and Social Policy.

De Luca, G., and Rossetti, C. (2019a). Weights and imputations. In M. Bergmann, A. Scherpenzeel and A. Börsch-Supan (Eds.), SHARE Wave 7 methodology: Panel inno-

vations and life histories (pp. 167–189). MEA, Max Planck Institute for Social Law and Social Policy.

De Luca, G., and Rossetti, C. (2019b). Computing calibrated weights in Stata. SHARE Working Paper Series 43–2019. Munich Center for the Economics of Aging (MEA).

Deville, J.-C., and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376–382.

Gelman, A., and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7, 457–511.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). Bayesian data analysis (2nd ed). Chapman and Hall

Groves, R. M., and Peytcheva, E. (2008). The impact of nonresponse rates on self-selection bias: A meta-analysis. *Public Opinion Quarterly*, 72, 167–189.

Haziza, D., and Lesage, E. (2016). A discussion of weighting procedures for unit nonresponse. *Journal of Official Statistics* 32, 129–145.

Horvitz, D. G., Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* 47, 663-–685

Lee, K. J., and Carlin, J. B. (2010). Multiple imputation for missing data: Fully conditional specification versus multivariate normal imputation. *American Journal of Epidemiology*, 171, 624–632.

Lundström, S., and Särndal, C. E. (1999). Calibration as a standard method for treatment of nonresponse. *Journal of Official Statistics*, 15, 305–327.

Lynn, P. (2009). Methods for longitudinal surveys. In P. Lynn (Ed.), *Methodology of longitudinal surveys*, 1–19. Wiley.

Molenberghs, G., Fitzmaurice, G., Kenward, M. G., Tsiatis, A., and Verbeke, G. (2015). *Handbook of missing data methodology*. CRC Press.

Raghunathan, T. E., Lepkowski, J. M., Van Hoewyk, J., and Solenberger P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology*, 27, 85–95.

Rubin, D. B. (1987). Multiple imputation for nonresponse in surveys. Wiley.

Schafer, J. L. (1997). Analysis of Incomplete Multivariate Data. Boca Raton, FL: Chapman & Hall/CRC

Van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research*, 16, 219–242.

Van Buuren, S., Boshuizen, H.C., and Knook, D. L. (1999). Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine*, 18, 681–694.

Van Buuren, S., Brands, J. P. L., Groothuis-Oudshoorn, C. G. M., and Rubin, D. B. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation*, 76, 1049–1064.

Vermeulen, K., and Vansteelandt, S. (2015). Biased-reduced doubly robust estimation. *Journal of the American Statistical Association*, 110, 1024–1036.

Chapter 3

Effects of interviewers on response to income and wealth items

Abstract

Nonresponse to items is a prevalent issue frequently encountered in survey data, particularly with regard to items related to income and wealth. In face-to-face surveys, interviewers influence item nonresponse. This study examines interviewer effects on nonresponse to financial items in the sixth wave of the Survey of Health, Ageing and Retirement in Europe (SHARE). The study investigates how interviewer expectations on the response rate to income questions affect actual response rates to income and asset questions achieved in the field.

To deal with missing covariate values, we use three different approaches: the complete-case analysis (CCA), fill-in (FI), and generalized missing indicator (GMI). The comparison of these approaches shows that the interviewer's expectations matter in the context of income and wealth questions, and positive expectations lead to obtaining more meaningful data for financial questions. Although interviewer expectations may change during the field period, training to build the interviewer's confidence can reduce the occurrence of item nonresponse.

Keywords: SHARE; item nonresponse; missing data; logit; complete-case analysis; multiple imputation; model averaging

JEL classification: To follow

1 Introduction

Sample surveys frequently suffer from various sources of nonsampling errors, such as coverage errors, unit and item nonresponse errors, attrition, and measurement errors, which may affect sample representativeness and quality of the data. These errors may depend on a number of features of the interview process (e.g., interview mode, length of the survey period, interviewers, interview instruments, questions wording, and so on...) (Groves and Couper 1998; West and Blom 2017; Banks *et al.* 2011; Olson 2014).

In interviewer-administered surveys, the interviewer has an important role in obtaining unit and item nonresponse outcomes (Korbmacher *et al.* 2013; Friedel *et al.* 2019; Durrant *et al.* 2010; Tourangeau and Yan 2007; Pickery and Loosveldt 2001, and Essig and Winter 2009). While some studies investigate that socio-demographical interviewer's characteristics such as age, gender, race, ethnicity, education level, and experience influence unit and item nonresponse (Berk and Bernstein 1988; Vercruyssen *et al.* 2017; Riphahn and Serfling 2005 and Bergmann *et al.* 2022), there is several empirical evidence of how interviewer non-demographic characteristics (e.g., interviewer personality traits, interviewer attitudes and so on...) affect unit and item nonresponse (see Lynn *et al.* 2013; Blom and Korbmacher 2013; Lipps and Pollien 2011; Silber *et al.* 2021; Wuyts and Loosveldt 2017; Schrapler 2006).

The importance of the interviewer information is first to map out the likelihood of reducing nonsampling errors by special interviewer training activities; special training activities may alter them to decrease in advance the occurrence of such nonsampling errors (Groves and Couper 1998; Schaeffer *et al.* 2010). Second, since these are important determinants of the response probability, such information could be used in ex-post adjustment methods (e.g., weights and imputations). The interviewer's characteristics are also important for ex-post adjustment methods based on missing at random assumption (e.g., weights and imputations). As emphasized by Fitzgerald *et al.* (1998), Nicoletti and Peracchi (2005), and De Luca and Peracchi (2012), interviewers' characteristics may also provide a valid set of exclusion restrictions to identify more general missing data mechanisms.

In this paper, we use data from the sixth wave of the Survey of Health, Aging, and Retirement in Europe (SHARE) and associated interviewer survey (SHARE_IWS) to study how interviewers' expectations on response to income questions affects the actual response rates achieved in the field. As studied by Sudman *et al.* (1977) and Singer and Kohnke-

Aguirre (1979), interviewers who expect difficulties obtain lower response rates on sensitive questions like gambling, income, excessive alcohol consumption, mental health, and sexual behaviors. Friedel (2020) shows that the interviewer's expectations affect item income and asset nonresponse rates, and Cunha *et al.* (2022) exhibit that optimistic and self-confident interviewers perform better on income response rates.

Although our paper and Friedel's (2020) share a common research question about the impact of interviewer expectations on nonresponses regarding income and assets, significant differences set them apart.

Key differences with respect to Friedel (2020) are that we employ three distinct approaches to address our research question while encountering missing covariate values in our model. We estimate and compare these estimators — the complete-case analysis (CCA), fill-in (FI), and generalized missing indicator (GMI) — to obtain the best estimates. In contrast, Friedle (2020) used a multilevel approach, recognizing the nested structure of respondents within interviewers and the underlying hierarchical setup.

Additionally, our analysis expands to the sixth wave of SHARE, covering 12 countries in the extended interviewer survey (IWS). We also delve into the issue of missing values in the covariates attributed to nonparticipation and unanswered items in both the interviewer survey and the regular SHARE interview. Specifically, we enhance the SHARE multiple imputation database for respondents' characteristics to accommodate a hundred imputations and supplement it with hot-deck multiple imputations for the interviewers' attributes.

Finally, we examine and demonstrate the relevance of country heterogeneity in understanding how the expected response rate of interviewers to income inquiries impacts the actual response probability. This is crucial as there are different determinants that influence the response variables at the country level, including the variability of survey agencies, cultural factors, and so on.

To study the effects of interviewers on the nonresponse errors in income and wealth questions, we combine the interviewer survey database of the 2015 wave with the auxiliary source of data about the interviewers, namely the interviewer roster database.[1] that contains additional data for the interviewer's socio-demographic characteristics (e.g., interviewer age, gender, years of experience). While the primary objective of this paper is not to capture

[1] The interviewer roster data are not included in the public release of the SHARE data. The SHARE central administration kindly provided this additional source of data.

the causal effects of our covariate of interest, interviewers are not randomly assigned to respondents, unlike in experiments, in order to reduce survey costs. However, the collection of interviewer survey (IWS) data prior to the commencement of the main survey (CAPI) ensures that estimated effects are not affected by issues of reverse causality.

The remainder of the paper is organized as follows. Section 2 briefly describes the SHARE data of wave 6 and summary statistics of response variables. The "Choice of predictors and missing data patterns" section 3 is devoted to description and summary statistics of regressors. The statistical methods adopted in the analysis are reported in the "Methodology" section 4 . Analysis and results of the model estimation are presented in the "Results" section 5. Finally, the "Conclusions" section 6 offers a discussion of the relevant results and conclusions.

2 SHARE data

This paper is based on release 7.1.0 of the Survey of Health, Aging and Retirement in Europe (SHARE), a multidisciplinary and cross-national panel database of micro data on health, socio-economic status and social networks of the elderly European population. The panel currently comprises seven regular waves (2004-05, 2006-07, 2011, 2013, 2015, 2017, and 2019) on current living circumstances and two retrospective waves (2008-09 and 2017) on life histories. We focus on the 2015 regular wave (i.e. wave 6) where survey data about respondents have been supplemented by auxiliary survey data about interviewers (the so-called interviewer survey - IWS) in two-thirds of the participating countries (Austria, Belgium, Estonia, Germany, Greece, Italy, Luxembourg, Poland, Portugal, Slovenia, Spain, and Sweden). In this study, the focus is directed exclusively towards wave 6 of the SHARE Interviewer Survey, despite the availability of data for waves 5, 6, and 7. The decision to exclude waves 5 and 7 is rooted in specific considerations. Firstly, the limited participation of countries in the interviewer survey during wave 5 has led to a constraint in the sample size, thereby restricting the analytical scope. Additionally, the retrospective nature of the survey in wave 7 for the majority of respondents introduces comparability challenges, as the data is obtained through varied questioning methods, potentially influencing the coherence and consistency of the analysis. To reduce missing values on basic socio-demographic characteristics of the

interviewers, we also combine the IWS data with the other paradata obtained from the national survey agencies (the so-called interviewer roster - IWR). In the following subsections, we describe these two sources of data and the criteria used to select our sample.

2.1 Main survey data of SHARE wave 6

The target population of wave 6 consists of people born in 1964 or earlier, who speak (one of) the country's official languages (regardless of nationality and citizenship), and who do not live either abroad or in institutions such as prisons and hospitals during the entire fieldwork period.

National samples are selected through probability-based sampling designs. However, sampling procedures are not completely standardized across countries because of the lack of suitable sampling frames for the target population of interest (see e.g., Bergmann *et al.* 2017). To limit the impact of sample representativeness issues and coverage errors for certain population groups, we restrict our sample to respondents born between 1934 and 1964 who live in residential households. Younger cohorts of respondents are included in the sample only because they are spouses/partners of age-eligible respondents, but are not representative of the underlying population. Similarly, we exclude older cohorts of respondents and respondents living in nursing homes or other healthcare institutions because of likely coverage errors in the national sampling procedures for the institutionalized population. In total, our sample includes 41,934 respondents from 12 countries that have also participated in the SHARE interviewer survey of wave 6.

Like all other regular waves of SHARE, the interview mode adopted in wave 6 is face-to-face computer-assisted personal interviewing (CAPI), supplemented by show cards and a self-administered paper-and-pencil questionnaire. The CAPI questionnaire, which represents the largest part of the interview, is organized into 23 modules that cover a wide range of topics such as demographics and family composition, physical and mental health, behavioral risks, cognitive abilities, well-being, labor force participation, incomes, health and consumption expenditures, assets, financial transfers, social relations, and expectations. To reduce the burden of the interview, 7 modules are asked only to one person per household/couple: questions about assets and financial transfers are asked only to the financial respondent, questions about children and social support are asked only to the family respondent, and

questions about the income of non-eligible household members, housing, and consumption expenditure are asked only to the household respondent.[2] Another exception is the module on interviewer observations, which collects information on the interview process and is completed by the interviewer at the end of each interview without involving the respondent.

Most of the variables available in SHARE present negligible fractions of missing values (usually, much less than 5 percent of the eligible respondents to each item), but this is not the case for financial variables about incomes, assets, and consumption expenditures which are collected by open-ended questions that are sensitive and difficult to answer. Table 3.1 shows the response rates in the eligible respondents' sample of 16 financial variables such as income from various sources (first panel), real and financial assets (second panel), and health and food expenditures (third panel). The number of eligible respondents after removing outliers varies across items because of branching and skip-patterns included in the CAPI questionnaire, and it is extracted from the imputations module (gv_imputations). The response rate on some financial variables reaches a particularly low worrisome level. For example, around one-third of the eligible respondents do not answer questions about money hold in bank accounts and the value of the main residence. A single one-shot question on total household income suffers from about 24 percent of missing values. These large fractions of missing values lead to serious concerns on the potential selection bias and efficiency loss generated by item nonresponse errors. Notice that we focus on four specific financial variables: "total household income", "old age, early retirement, survivor pensions", "value of the main residence", and "bank account", only. We have excluded the other 12 financial variables from our analysis for two main reasons. Firstly, this is due to the limited number of eligible observations for certain variables, such as "bonds, stocks, and mutual funds". Secondly, we have observed a low nonresponse rate in some financial variables.

[2] The "financial respondent" is either a single or the partner of each couple who is most knowledgeable about financial matters, the "family respondent" is either a single or the partner of each couple who is interviewed first, while the "household respondent" is the household member who is knowledgeable about housing matters.

Table 3.1: Response rate of answers on the financial variables of SHARE wave 6 in the eligible respondent's sample

	Respondents		
Variable	Type	Elig.	RR
Total household income	**HR**	**28204**	**0.761**
Earnings from employment	AR	11429	0.794
Earnings from self-employment	AR	2736	0.632
Old age, early ret., survivor pensions	**AR**	**23615**	**0.848**
Interests from financial assets	FinR	15130	0.384
Value of main residence	**HR**	**22341**	**0.660**
Value of real estate	HR	7238	0.601
Value of cars	HR	20551	0.816
Bank accounts	**HR**	**24919**	**0.615**
Bond, stock and mutual funds	FinR	5750	0.550
Mortgage on main residence	HR	4196	0.495
Financial liabilities	FinR	4863	0.771
Out-of-pocket exp.: outpatient care	AR	24925	0.855
Out-of-pocket exp.: prescribed drugs	AR	30362	0.904
Food at home exp.	FamR	28204	0.865
Food outside home exp.	FamR	16982	0.886

Notes: AR means "all respondents", HR means "household respondents", FinR means "financial respondents", and FamR means "Family respondents". RR is the response rate on each financial variable, while UB is the percentage of missing values with some informative UB answer

In this paper, our focus is solely on examining the determinants of the response process for the 4 financial variables, as indicated among others in Table 3.1. In addition to the observable characteristics of the eligible respondents to each item, we shall exploit the auxiliary data collected in the SHARE interviewer survey to evaluate the impact of observable interviewers' characteristics on the response probability to financial variables.

2.2 Interviewer survey data of SHARE wave 6

Table 3.2: Number of interviewers, number of participants, and participation rate to the interviewer survey (IWS) of SHARE wave 6 by country

		IWS	
Country	Total	Obs.	PR
Austria	70	51	0.729
Belgium	132	106	0.803
Estonia	82	35	0.427
Germany	147	128	0.871
Greece	170	88	0.518
Italy	140	132	0.943
Luxembourg	44	24	0.545
Poland	60	27	0.450
Portugal	51	39	0.765
Slovenia	59	48	0.814
Spain	116	57	0.491
Sweden	101	73	0.723
Total	1172	808	0.689

Notes. PR denotes the participation rate to the IWS.

Since its pilot study in wave 4, SHARE conducted an interviewer survey (IWS) to supplement the survey data about respondents with detailed information about their interviewers. The IWS of wave 6 was conducted as an online survey after the national interviewer training sessions, but prior to the fieldwork of the main survey in each country. Although interviewers are not randomly assigned to respondents, this feature of the IWS ensures that the variables used to study the effects of interviewers on the survey outcomes do not suffer from reverse causality problems.

In addition to basic socio-demographic characteristics such as gender, age, educational attainments, and occupational status, the IWS questionnaire is based on the conceptual framework developed by Blom and Korbmacher (2013) which identifies four key dimensions of interviewer characteristics that are important to study the impact of the interviewer effects

on various forms of nonsampling errors such as unit and item nonresponse, lack of consent to record linkage, and lack of cooperation with other survey requests. Here, we focus on the interviewer characteristics that are likely to play an important role in explaining item response errors on financial questions. In particular, the first dimension of the IWS questionnaire refers to interviewers' attitudes towards the survey process and their job, which are measured by a set of questions on the reasons for being an interviewer, circumstances under which deviating from the interview protocol to best approach difficult respondents, trust in other people, and data protection concerns. The second dimension refers to interviewers' own behavior regarding data collection requests and how interviewers would behave in similar situations as their respondents. For example, the IWS includes a set of questions on whether an interviewer should respect the privacy of respondents, whether a refusal from a reluctant respondent should be accepted, and whether putting great effort into persuading the respondents affect the reliability of their answers. Further, it asks for the total household income of the interviewer to assess possible relationships between the response behavior of respondents and interviewers to sensitive financial questions. The third dimension refers to interviewers' experience with social surveys in general and with the previous waves of SHARE. Finally, the fourth dimension refers to interviewers' expectations about survey outcomes, such as their expectations of the response rate to income questions.

The IWS provides valuable information for understanding the complex process through which interviewers may influence the nonsampling errors of a survey. However, the fact that this survey is also subject to problems of the unit and item nonresponse may lead to biased and inefficient estimates of the interviewers' effects of interest. In SHARE, the countries' participation in the IWS is voluntary, as well as the participation of the interviewers in the participating countries. Table 3.2 shows the number of interviewers who worked for the main survey of wave 6, the number of interviewers who have participated in the IWS, and the resulting participation rate separately by country. The IWS covers 808 out of the 1172 interviewers who performed at least one interview in the main survey of wave 6. The cross-country average participation rate is 69 percent, with a minimum of 43 percent in Estonia and a maximum of 94 percent in Italy.

To limit the impact of nonresponse errors in the IWS, we also exploit the interviewer roster (IWR) data collected by the national survey agencies. This administrative data contain only information on a few interviewers' characteristics (namely gender, age, years of

education, years of experience, and participation in the previous waves of SHARE). Still, they have the advantage of covering a relatively larger number of interviewers and therefore provide additional information on the missing values of the available interviewers' characteristics. Unfortunately, IWR data are not available for all Swedish interviewers, 17 Greek interviewers, and 18 Portuguese interviewers. Further, the available interviewers' characteristics are not observed for all countries (e.g., years of education is missing in Germany and years of experience is missing in Poland).

3 Choice of predictors and missing data patterns

In this study, the analysis aims at modeling the probability of response to financial questions in surveys, with respect to the interviewer's expectations of response rates to income questions, plus a set of control variables.

3.1 Regressor of interest

Our study's main explanatory variable is the interviewer's expectations of the probability that interviewees answer financial questions meaningfully. Before starting the field period, the interviewers are asked the following question: "what does the interviewer expect? How many of his/her respondents (in percentage) in SHARE will provide information about their income?" It is asked by interviewers to provide a numerical answer scaled from 0 to 100 percent with a one percent increment, but its empirical distribution presents several focal values shown in Figure 3.1. Therefore to limit the impact of measurement errors on the regressor of interest, we use a binary indicator that takes the value 1 for all interviewers with an expected response rate greater than the median of its country-specific distribution. Further, as the paper discusses (pag. 26), interviewers are not randomly assigned to respondents. This means that interviewers may have different expectations about whether the people they interview will answer questions about their income. These expectations can be influenced by the interviewers' previous experiences with these people in earlier surveys, which might also be connected to how they will respond in future waves. Interviewers' expectations are, in this sense, endogenous to item nonresponse due to omitted variable bias. Note that within

SHARE, it is not possible to track the same interviewer across different follow-up waves, as their IDs tend to change with each subsequent wave.

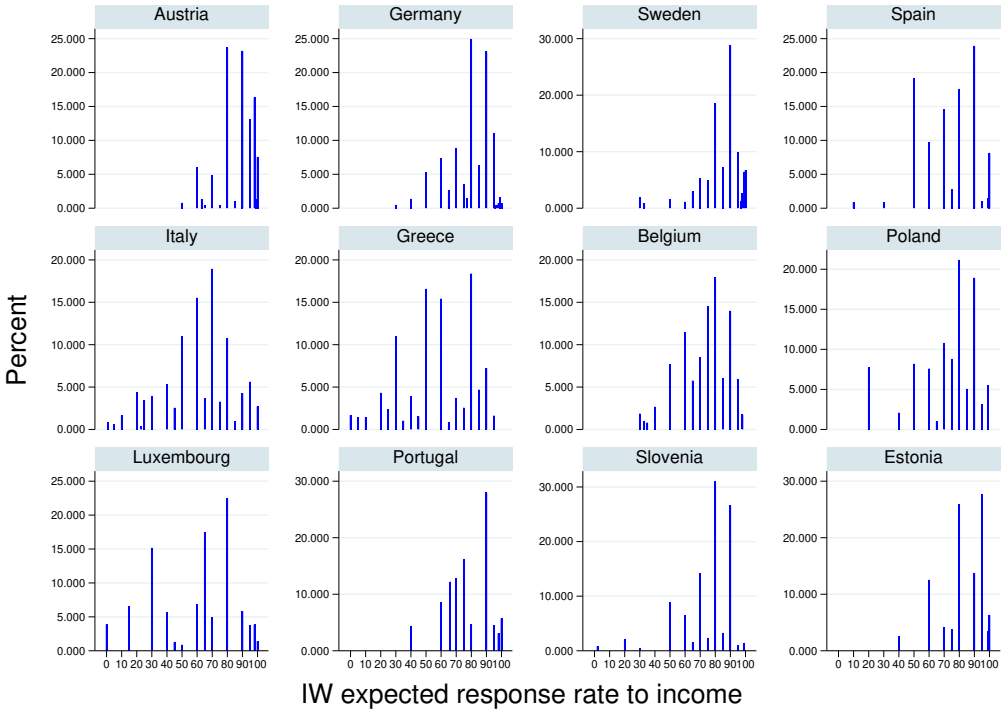

Figure 3.1: Distribution of expectations of the interviewer on response rate to income by country

The paper provides a descriptive analysis of the cross-country heterogeneity in response to the outcomes of interest (i.e., whether and how the item nonresponse varies across countries). The scatterplot (see Figure 3.2) demonstrates the unadjusted correlation between the response rates at the country level and the average confidence levels of the interviewers. This graphical representation elucidates the relationship between these two key variables, contributing to a deeper understanding of the data collection process within different countries.

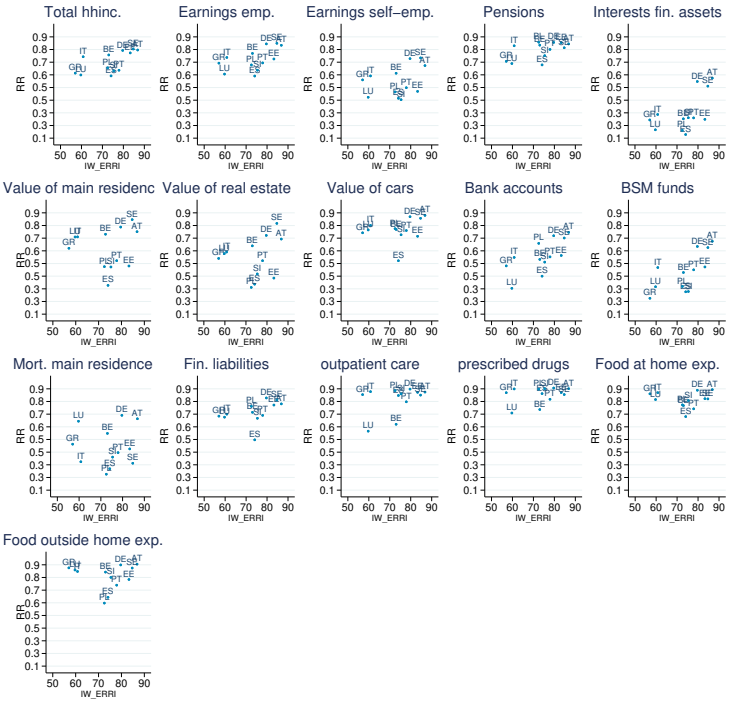

Figure 3.2: the country-level response rate and the country-level average interviewers' confidence level

3.2 Control variables

Besides basic socio-demographic characteristics (e.g., gender, age, years of experience as an interviewer, and education level), control variables at the interviewer level include the workload status during the fieldwork period, self-reported health status, the interviewer's response to total household income: whether the interviewer responds to the question about the average monthly income of his/her household after taxes in the last year or not, and interviewer strategies during CAPI interview, the interviewer speak fast: if the interviewer finds out that the respondent is in a hurry during the interview process, he or she speaks fast, and the interviewer clarified questions: if the respondent does not understand a question, he/she explains what the question really means.

At the respondent level, the control variables comprise socio-demographic characteristics such as gender, age, marital status, education level, and participation status in the past wave. Also, additional covariates that explain outcomes at the respondent level are supplemented to the model: the respondent's score on numeracy, fluency tests and self-rated memory, which assess cognitive abilities, and self-assessed health status, body mass index and depression status that measure the physical and mental health status of respondents. The descriptions and non-missing data summary statistics of the major explanatory variable and the control variables are given in Table 3.3.

Table 3.3: Definitions and summary statistics of the control variables in the eligible respondent's sample

Description	Obs.	Mean	Std.
IW ERR income:high	26577	0.439	0.496
IW female	40537	0.721	0.449
IW high education	29213	0.432	0.495
IW workload: high	41934	0.749	0.434
IW good health	29386	0.585	0.493
IW response to THI	29498	0.702	0.458
IW speak fast	29397	0.425	0.494
IW clarifies questions	29449	0.607	0.489
IW age	40537	51.388	11.767
IW yrs of experience	38296	10.314	8.895
R female	41934	0.552	0.497
R lives in couple	41934	0.758	0.428
R high education	41228	0.592	0.491
R numeracy score	38128	0.657	0.475
R good health	41870	0.613	0.487
R part. past waves	41934	0.807	0.395
R good memory	39006	0.731	0.444
R limit. with activities	41868	0.443	0.497
R depression status	40230	0.730	0.444
R age	41929	65.806	8.120
R fluency score	40216	20.008	7.860
R BMI	41112	27.146	4.594

Note: Obs. is number in the eligible respondents sample, Mean denotes average, and Std. is standard errors.

3.3 Missing data patterns.

In our study, the issue of missing data poses a significant challenge as almost all covariates suffer from some level of data missing. The missing data can be attributed to two distinct sources, each contributing to the complexity of the missing data patterns we observed. These sources include nonresponse errors from the interviewers in the Interviewer Survey (IWS) and nonresponse errors from the respondents in the Computer-Assisted Personal Interview (CAPI) phase.

1. **Nonresponse errors from the interviewers in the IWS.** Data belonging to the interviewer survey are missed because of either nonparticipation or item nonresponse. Hence, this source of missingness at the respondent level leads to approximately a high missing data percentage. Let I_1 be a binary indicator that takes the value 1 for the interviewers who participate and answer all questions in the interviewer survey and value 0 otherwise.

2. **Nonresponse errors from the respondents in the CAPI.** This source concerns all variables related to the respondents' socio-demographic characteristics, health measures, and cognitive abilities. Let I_2 be a binary indicator for this additional source of nonresponses that takes the value 1 for the respondents who participate and respond to all questions in CAPI and value 0 otherwise.

The simultaneous existence of these two sources of missing data results in the emergence of three distinct missing data patterns, which are meticulously summarized in Table 3.4. This table presents a comprehensive overview of the complete case subsample and the various missing data patterns observed across different countries.

In particular, the "CC" column in the table signifies the number of complete cases without any missing data, indicating the robustness of the data collection process for these cases. On the other hand, the columns labeled "NR_IW", "NR_R" and "NR_IW_R" represent the number of cases with missing data due to nonresponse errors from the interviewers in the IWS, nonresponse errors from the respondents in the CAPI, and a combination of nonresponse errors from both the interviewers and the respondents, respectively.

The data reveals variations in the patterns across different countries, with some exhibiting higher proportions of missing data due to nonresponse errors from either interviewers, re-

spondents, or both. Understanding these patterns is crucial for developing robust strategies to handle missing covariate data.

Table 3.4: Complete-case subsample and missing data patterns by country

	Patterns				
Country	CC	NR_IW	NR_R	NR_IW_R	Total
Austria	2006	587	295	65	2953
Germany	3132	683	149	27	3991
Sweden	2018	1010	217	106	3351
Spain	1563	1616	608	777	4564
Italy	3315	822	432	125	4694
Greece	1047	1156	780	1313	4296
Belgium	3380	1248	261	94	4983
Poland	536	751	135	197	1619
Luxembourg	832	473	77	25	1407
Portugal	632	543	187	127	1489
Slovenia	2346	1121	162	109	3738
Estonia	1802	2441	294	312	4849
Total	22,609	12,451	3,597	3,277	41,934

Note: CC denotes complete-case subsample, NR_IW indicates nonresponse in interviewer survey, NR_R is nonresponse in CAPI, NR_IW_R denotes nonresponse in interviewer survey and CAPI.

4 Methodology

In this section, we explore three distinct approaches to determine whether the response to income and assets items is influenced by interviewers' confidence level. Moreover, these approaches aid us in addressing missing covariate values, the core concern in our study.

To handle missing data, understanding missing data mechanisms is crucial in implementing appropriate strategies. While the detailed discussion of missing completely at random, missing at random, and not missing at random (hereafter, MCAR, MAR, NMAR) is beyond the scope of this chapter, their importance in the context of data analysis cannot be overlooked. Therefore, we briefly explain it.

Rubin (1976) categorized missing data problems into three types: MCAR, MAR, and NMAR, each representing different patterns in the missing data mechanism. MCAR refers to data that are missing completely at random, implying that the reasons for the missing data are unrelated to the data itself. On the other hand, MAR indicates that the missing data are dependent on observed data, and the reasons for missing data can be attributed to known properties. NMAR, represents data that are missing not at random, indicating that the probability of missing data varies due to reasons unknown to the analyst.

In the case of MCAR, the missing data mechanism is unrelated to the data, and thus, it is considered the most straightforward but often unrealistic assumption. MAR is a broader class than MCAR, accounting for dependencies on observed data, while NMAR is the most complex case, indicating missing data patterns due to unknown reasons. Rubin's distinctions highlight the importance of understanding the missing data mechanism in selecting appropriate methods for analysis, as many simple fixes only work under the restrictive and often unrealistic MCAR assumption. Handling NMAR may involve acquiring more information about the causes of missingness or conducting sensitivity analyses under various scenarios.

4.1 Complete-case analysis

One common approach to deal with the missing data is the complete-case analysis (CCA), which consists of deleting all observations with missing covariate values and estimating the model of interest using only the complete-case (CC) subsample. Our application corresponds to estimating a set of logit models for the four highlighted binary response indicators of the

financial outcomes listed in Table 3.1. If $\mathrm{Y}_{j,0}$ $(\mathrm{N}_{j,0} \times 1)$ is the indicator vector of observations of outcome interest in the complete-case (CC) subsample, which takes value 1 if the ith eligible respondent answers the jth financial variable, and value 0 otherwise, the logistic regression model of the complete-case (CC) subsample can be considered as following linear predictor

$$\eta_{j,0} = \mathbf{X}_{j,0}^{\top} \beta_j, \tag{1}$$

where $\mathbf{X}_{j,0}$ $(\mathrm{N}_{j,0} \times \mathrm{K}_j)$ is the matrix of observations on the regressors as $\mathrm{N}_{j,0} < \mathrm{N}_j$ is the size of the complete-case (CC) subsample, and β_j $(\mathrm{K}_j \times 1)$ is the unknown parameter vectors.

In addition to standard regularity conditions for maximum likelihood (ML) estimation of binary logit models, properties of the complete-case analysis (CCA) estimator of β_j in model (1) depend crucially on the validity of two assumptions (Dardanoni *et al.* 2015): (i) Fisher information matrix for the subsample with complete data is positive definite with probability approaching one as the sample size goes to infinity; (ii) the conditional on covariates, the response probability is the same in the subsamples with and without missing covariates. The first assumption requires that one can identify the β_j from the complete-case (CC) subsample of each outcome. The second assumption requires that the response variable and missing data mechanism for the covariates are conditionally independent.

Notice that even if two assumptions above are held, and the complete-case analysis (CCA) gives us asymptotically consistent estimates of β_j, we lose much data. For instance, in our study, the complete-case (CC) subsample includes $22,609$ observations shown in Table 3.4, while the complete data sample contains $41,934$, so using the complete-case (CC) subsample can result in losing $19,325$ data in our analysis. Hence, this amount of lost data decreases precision in the CCA approach. Using the CCA method comes at the price of losing data that may be valuable, so a better strategy can be to impute the values of the missing covariate.

4.2 Fill-in approach

Fill-in (FI) is a popular approach to deal with the problem of missing data in estimated values substituting missing values. The fill-in (FI) approach encompasses various methods; we focus on the most prevalent one here, i.e., multiple imputations (MI). Rubin (1987) developed multiple imputation (MI) within the Bayesian framework, where data augmen-

tation is heavily reliant on Bayesian methodology. A multiple imputation analysis consists of three main phases: the imputation phase, the analysis phase, and the pooling phase. In the imputation phase, multiple datasets (e.g., m = 10) are generated, each containing different estimations of the missing values. Subsequently, in the analysis phase, these completed datasets are subjected to standard statistical procedures, with the analysis repeated for each imputed dataset. This process yields m sets of parameter estimates and standard errors. Finally, in the pooling phase, the results are consolidated into a single set of outcomes, often using Rubin's formulas for pooling parameter estimates and standard errors. Multiple imputation encompasses various techniques, and while the three-step process remains consistent, different algorithms exist for the imputation phase, proposed by various methodologists (Schafer, 1997, 2001; van Buuren, 1999, 2006, 2007). Van Buuren et al. (1999, 2006, 2007) used the FCS method which uses a Gibbs sampling algorithm, which imputes multiple variables jointly and iteratively through a sequence of regression models. Assume we want to impute arbitrary patterns of missing values on a set of J variables. The basic idea of the FCS method is that, at each step of the iterative process, we impute the missing values on the j-th variable (j=1,. . . , J) by drawing from the predictive distribution of a regression model that includes as predictors the most updated imputations of the other J – 1 variables (as well as other fully observed predictors). The process is applied sequentially to the whole set of J variables and is repeated in a cyclical manner by overwriting at each iteration the imputed values computed in the previous iteration.

In the following, supposing that Y_j ($\mathrm{N}_j \times 1$) is the indicator vector of observations of outcome interest in complete data, the logit model of the fill-in (FI) sample can be written by following the linear predictor

$$\eta_j = \mathbf{W}_j^{\top} \beta_j, \tag{2}$$

where $\mathbf{W}_j$ ($\mathrm{N}_j \times \mathrm{K}_j$) is a matrix of observed and imputed data on the regressors, and β_j ($\mathrm{K}_j \times 1$) is the unknown parameter vectors.

In the multiple imputations (MI) method, in addition to the assumption of missing at random (MAR) data to get asymptotically equivalent fill-in maximum likelihood (ML) estimator with ML estimator from the complete data sample, we require an additional condition that the model used to construct the imputations is more general than the model used to analyze the imputed values (i.e., congeniality, Meng 1994). Notice that the validity of the

imputation model is not taken for given. Accordingly, the fill-in parameter estimates β_{FI_MI} might not be asymptotically consistent.

As mentioned in the previous paragraph, we fill in missing covariate values on respondents' and interviewers' characteristics by the fully conditional specification (FCS) method of van Buuren *et al.* (2006) is based on an iterative sequence of univariate imputation methods. To ensure that the imputed values of an interviewer do not change across her/his respondents, we exploit two sequential Gibbs samplings: one for the respondent's variables and one for the interviewer's variables. Further, standard errors can decrease either due to the amount of association between auxiliary variables used in the imputation model and the variables being imputed or the number of imputations (von Hippel, 2020), so in this study, in order to the minimally sufficient number of imputations that decrease standard errors, we exploit a rule of thumb: $\mathrm{M} \geqslant 100 \times \mathrm{FMI}$, where FMI is a fraction of missing information (e.i. the ratio of information lost due to the missing) (see e.g., Stata 17 help manual), and since the FMI on interviewer variables is around 0.85, we employ $\mathrm{M} = 100$ multiple imputations.

4.3 Generalized missing-indicator (GMI) and model averaging (MA)

The generalized missing indicator or "grand model"[3] approach introduced by Dardanoni *et al.* (2011, 2012, 2015) allows us to augment the model space by considering not only unrestricted and fully-restricted specification forms of the grand model that correspond to the complete-case (CC) (1)[4] and fill-in (FI) (2) approaches respectively but also all intermediate sub-models with a subset of auxiliary parameters δ_j restricted to zero. The general message of the grand model is that expansion of model space brings up the model uncertainty problem. One approach that treats the model uncertainty is model averaging (MA), based on the idea that each model contributes information on the parameters of interest.

Suppose all possible missing patterns of data are indicated by $\mathrm{h}_j = \{1, \cdots, \mathrm{H}_j\}$. The model space $\mathcal{M}_j$ includes R_j possible LOGITs (i.e. $\mathrm{R}_j = 2^{\mathrm{H}_j}$), that is $\mathcal{M}_j = \{\mathrm{M}_{j,1} \cdots \mathrm{M}_{j,\mathrm{R}_j}\}$.

[3] We consider the "grand model" of the form

$$\eta_j = \mathbf{W}_j^{\top} \beta_j + \mathbf{Z}_j^{\top} \delta_j, \tag{3}$$

where $\mathbf{W}_j$ $(\mathrm{N}_j \times \mathrm{K}_j)$ and $\mathbf{Z}_j$ $(\mathrm{N}_j \times \mathrm{H}_j\mathrm{K}_j)$ are the matrices of the "fill-in" and "auxiliary" regressors, respectively.

[4] Dardanoni *et al.* (2015) show the complete-case analysis (CCA) estimates of β_j in model (1) is numerically equivalent to the ML estimates of β_j in the grand-model

The r_jth logit model, M_{j,r_j}, can be estimated by the following linear predictor

$$\eta_{j,r_j} = \mathbf{W}_j^\top \beta_j + \mathbf{Z}_{j,r_j}^\top \delta_{j,r_j}, \tag{4}$$

where $\mathbf{W}_j$ $(\mathrm{N}_j \times \mathrm{K}_j)$ and $\mathbf{Z}_{j,r_j}$ are the matrices of the "focus" regressors and the subset of $\mathsf{P}_{j,r_j} \in [0 \;\; \mathrm{H}_j\mathrm{K}_j]$ "auxiliary" regressors, respectively. The δ_{j,r_j} is the corresponding vector of auxiliary coefficients of model r_jth.

Consequently, our model averaging estimate of coefficient of interest $\beta_{j,\mathrm{MA}}$ is considered the form

$$\widehat{\beta}_{j,\mathrm{MA}} = \sum_{r_j=1}^{R_j} \lambda_{j,r_j} \widehat{\beta}_{j,r_j}, \tag{5}$$

where $\lambda_{j,r_j} \geqslant 0 \quad$ and $\quad \sum_{r_j=1}^{R_j} \lambda_{j,r_j} = 1$. The $\widehat{\beta}_{j,r_j}$ is the ML estimates of β_j under the r_jth model.

Bayesian model averaging (BMA) with information criteria weights

In this study, we explore two commonly approaches for approximating the posterior model probabilities, denoted as λ_{j,r_j}=$\Pr(\mathrm{M}_{j,rj}|\mathrm{Y}_j)$. The weights used in our model averaging calculations for estimating $\beta_{j,\mathrm{MA}}$ are as follows:

$$\lambda_{j,r_j} = \Pr(\mathrm{M}_{j,r_j}|\mathrm{Y}_j) = \frac{\Pr(\mathrm{Y}_j|\mathrm{M}_{j,r_j})\Pr(\mathrm{M}_{j,r_j})}{\sum_{r_j=1}^{R_j} \Pr(\mathrm{Y}_j|\mathrm{M}_{j,r_j})\Pr(\mathrm{M}_{j,r_j})}, \tag{6}$$

where $\Pr(\mathrm{M}_{j,r_j})$ is the prior probability of the $\mathrm{r}_j th$ model, and

$$\Pr(\mathrm{Y}_j|\mathrm{M}_{j,r_j}) = \int_{\beta_j \in \mathrm{B}_j} \Pr(\mathrm{Y}_j|\beta_{j,r_j}, \mathrm{M}_{j,r_j})\Pr(\beta_{j,r_j}|\mathrm{M}_{j,r_j})\mathrm{d}\beta_{j,r_j}, \tag{7}$$

is the marginal likelihood of model r_jth. The $\Pr(\mathrm{Y}_j|\beta_{j,r_j}, \mathrm{M}_{j,r_j})$ is the sample likelihood, and $\Pr(\beta_{j,r_j}|\mathrm{M}_{j,r_j})$ is the prior density of β_{j,r_j} under the $\mathrm{r}_j th$ model. and β_{j,r_j} is the vector of parameters of interest of model r_jth.

Notice that in order to obtain the posterior probability of the model, r_jth (6), we ought to calculate the marginal likelihood of the data (7). However, there is usually no closed-

form solution for the marginal likelihood of the data analytically. To address this challenge, Raftery (1996) uses the Laplace method for integrals to approximately compute (7), if the following assumptions hold: diffuse priors and equal prior model probabilities, then the marginal likelihood of model r_jth is obtained by Schwarz's theorem (1978) as that is

$$Pr(\mathrm{Y}_j \mid \mathrm{M}_{j,r_j}) \approx \exp\left(\frac{-\mathrm{BIC}_{j,r_j}}{2}\right),$$

BMA weights can then be approximated by

$$\lambda_{j,r_j} = \Pr(\mathrm{M}_{j,r_j}|\mathrm{Y}_{j,r_j}) = \frac{\exp(-\mathrm{BIC}_{j,r_j}/2)}{\sum_{r_j=1}^{R_j}\exp(-\mathrm{BIC}_{j,r_j}/2)} = \frac{\exp(-\Delta\mathrm{BIC}_{j,r_j}/2)}{\sum_{r_j=1}^{R_j}\exp(-\Delta\mathrm{BIC}_{j,r_j}/2)}, \tag{8}$$

$$\Delta\mathrm{BIC}_{j,r_j} = \mathrm{BIC}_{j,r_j} - \mathrm{BIC}_{j,r_j,\min} \tag{9}$$

where $\mathrm{BIC}_{j,r_j,\min}$ is the minimum of the R_j different BIC values.

In the spirit of likelihood ratio methods, Buckland *et al.* (1997) proposed an alternative method for computing the posterior model probabilities using Akaike's Information Criterion (AIC). as the weight assigned to the j-th model:

$$\lambda_{j,r_j} = Pr(\mathrm{M}_{j,r_j}|\mathrm{Y}_j) = \frac{\exp(-\mathrm{AIC}_{j,r_j}/2)}{\sum_{r_j=1}^{R_j}\exp(-\mathrm{AIC}_{j,r_j}/2)} \tag{10}$$

where a higher λ_{j,r_j} indicates a higher probability of the model being plausible. For ease of computation, Burnham and Anderson (2002) propose a slightly different version:

$$\lambda_{j,r_j} = Pr(\mathrm{M}_{j,r_j}|\mathrm{Y}_j) = \frac{\exp(-\Delta\mathrm{AIC}_{j,r_j}/2)}{\sum_{r_j=1}^{R_j}\exp(-\Delta\mathrm{AIC}_{j,r_j}/2)} \tag{11}$$

where

$$\Delta\mathrm{AIC}_{j,r_j} = \mathrm{AIC}_{j,r_j} - \mathrm{AIC}_{j,r_j,\min} \tag{12}$$

where $\mathrm{AIC}_{j,r_j,\min}$ is the minimum of the R_j different AIC values.

5 Results

Generally speaking, the findings reveal that item nonresponse for all four questions asked (i.e. household income; old age, early retirement, and survivor pensions; bank accounts; value of the main residence) are positively affected by the interviewer's expected response rate in three approaches investigated and shown in Tables (3.5), (3.6), (3.7), and (3.8). Only for a small number of countries, the interviewer's expected response rate negatively influences item nonresponse rates to financial questions, but nearly no significant negative effects. Moreover, significant negative effects exist on the value of the main resident items in Greece for FI_MI and BMA_BIC methods only, as well as on pensions in Spain for CCA and BMA_AIC.

Furthermore, the crucial reason we employ the fill-in approach is to incorporate all the available information into the logistic regression model, which might lead to decreasing standard errors. Notice that in the multiple imputations (FI_MI) approach, the pooling standard error stems from two different components that reflect the within and between sampling variance of the mean difference in the multiple imputed datasets. Although squared standard errors in each imputed dataset, i.e., so-called within-imputation variance, decrease in FI_MI approach compared to CCA, the pooling standard errors obtained by the FI_MI approach are almost larger than those obtained by the CCA because of additional uncertainty due to the imputation of missing values, i.e., so-called between-imputation variance (see Tables (3.5) and (3.6)).

In the end, this study compares the predictive performance of the block Bayesian model averaging based on Bayesian and Akaike information criteria estimators with the CCA and FI_MI estimators. Despite that, in our model space, all models are equally likely a priori; our block BMA procedures lead to a posterior distribution concentrated in a few models. Hence, the block BMA_AIC estimates β_{BMA_AIC}, approximately corresponds to the CCA estimates β_{CCA}, and the block BMA_BIC estimates β_{BMA_BIC}, approximately coincides to the FI_MI estimates β_{FI_MI}.

Limitations. There are certain limitations to the present results. First, as said earlier, the results of average marginal effects (AME) of the block Bayesian model averaging based on BIC and AIC are so close to two extreme forms of specification of the GMI model (grand model), i.e., CCA and FI_MI. Therefore, we tried to use the conjugate priors for LOGITs,

which allow estimating posterior model probabilities using a Markov Chain Monte Carlo algorithm. However, the high number of imputation sets increases the computation time. In the future study, we may find a way to implement a new approach like WALS (Magnus and De Luca, 2016) to lower the computation time significantly. Second, because respondents are nested within interviewers, using a simple standard logit model underestimates the true standard errors, meaning that we tend to over-reject the null hypotheses.

6 Conclusions

Our results underscore that there exist associations support the hypothesis that interviewer expectations regarding respondent probability to report their income are correlated with item nonresponse to financial questions. Respondents are more likely to report their income and assets information when interviewed by an interviewer who expected more than 50 percent of their respondents to report their income than when interviewed by an interviewer who expected 50 percent or fewer of their respondents.

In the current study, optimistic interviewers whose respondents answer questions on their income of more than 50 percent are found to decrease item income nonresponse by up to 14 percent in some countries and item assets nonresponse by up to 26 percent in some countries. These effects are substantial, with an average response rate of 80 percent for the two income questions and 64 percent for the two asset questions. These findings indicate that interviewer expectations are important in older people's surveys. In order to reduce nonresponse rates to financial questions, this study suggests hiring and training interviewer strategies.

7 Appendix

7.1 Appendix A: Tables

Table 3.5: Estimates of the average marginal effects of interviewer's expected response rate over complete-case subsample by country through CCA approach

Country	thinc2	ypen1	bacc	home
ES	0.042	0.092 **	0.098 **	0.069 *
	(0.030)	(0.028)	(0.034)	(0.034)
IT	0.101 **	0.098 **	0.262 **	0.017
	(0.017)	(0.016)	(0.025)	(0.021)
GR	0.075 *	0.107 **	-0.085	-0.043
	(0.035)	(0.039)	(0.049)	(0.041)
PT	-0.036	0.020	-0.033	0.131 *
	(0.042)	(0.034)	(0.049)	(0.055)
PL	0.080	0.146 **	0.087	0.010
	(0.051)	(0.041)	(0.064)	(0.063)
SI	0.139 **	0.039	0.177 **	0.224 **
	(0.024)	(0.021)	(0.028)	(0.028)
AT	0.013	0.046 **	0.028	0.044
	(0.018)	(0.017)	(0.022)	(0.030)
DE	0.040 **	0.032 *	0.096 **	0.035
	(0.016)	(0.014)	(0.018)	(0.019)
BE	0.025	0.046 **	0.031	0.030
	(0.015)	(0.015)	(0.020)	(0.018)
LU	0.082 *	0.057	0.174 **	-0.005
	(0.039)	(0.043)	(0.041)	(0.038)
SE	0.004	-0.015	0.023	0.037 *
	(0.022)	(0.025)	(0.027)	(0.019)
EE	0.042	-0.053 **	0.125 **	0.116 **
	(0.023)	(0.018)	(0.031)	(0.035)

Table 3.6: Estimates of the average marginal effects of interviewer's expected response rate over complete-case subsample by country through FI_MI approach

Country	thinc2	ypen1	bacc	home
ES	0.079	0.096	0.093	0.079
	(0.074)	(0.072)	(0.055)	(0.068)
IT	0.085 **	0.073 **	0.211 **	0.024
	(0.026)	(0.019)	(0.042)	(0.029)
GR	0.038	0.062	0.003	-0.102 *
	(0.054)	(0.047)	(0.062)	(0.051)
PT	0.080	0.076	0.038	0.151
	(0.074)	(0.041)	(0.072)	(0.077)
PL	0.081	0.095	0.120	0.059
	(0.073)	(0.050)	(0.092)	(0.101)
SI	0.161 **	0.117 *	0.191 **	0.218 **
	(0.051)	(0.056)	(0.048)	(0.059)
AT	0.020	0.042 *	0.036	0.035
	(0.025)	(0.017)	(0.028)	(0.032)
DE	0.038 *	0.023	0.085 **	0.036
	(0.017)	(0.015)	(0.022)	(0.022)
BE	0.033	0.059 **	0.031	0.051 *
	(0.018)	(0.017)	(0.025)	(0.021)
LU	0.088 *	0.054	0.116	0.023
	(0.041)	(0.041)	(0.061)	(0.046)
SE	-0.003	-0.011	0.039	0.036
	(0.025)	(0.029)	(0.032)	(0.020)
EE	0.037	-0.019	0.082	0.046
	(0.037)	(0.027)	(0.066)	(0.068)

Table 3.7: Estimates of the average marginal effects of interviewer's expected response rate over complete-case subsample by country through BBMA_BIC approach

Country	thinc2	ypen1	bacc	home
ES	0.079	0.096	0.093	0.079
	(0.074)	(0.072)	(0.055)	(0.068)
IT	0.085 **	0.073 **	0.211 **	0.024
	(0.026)	(0.019)	(0.042)	(0.029)
GR	0.038	0.062	0.003	-0.102 *
	(0.054)	(0.047)	(0.062)	(0.051)
PT	0.080	0.076	0.038	0.151 *
	(0.074)	(0.041)	(0.072)	(0.077)
PL	0.081	0.113 **	0.120	0.059
	(0.073)	(0.042)	(0.091)	(0.101)
SI	0.164 **	0.072 *	0.196 **	0.218 **
	(0.046)	(0.036)	(0.043)	(0.059)
AT	0.020	0.042 *	0.036	0.035
	(0.025)	(0.017)	(0.028)	(0.032)
DE	0.038 *	0.023	0.085 **	0.036
	(0.017)	(0.015)	(0.022)	(0.022)
BE	0.033	0.059 **	0.031	0.051 *
	(0.018)	(0.017)	(0.025)	(0.021)
LU	0.088 *	0.054	0.116	0.023
	(0.041)	(0.041)	(0.061)	(0.046)
SE	-0.003	-0.011	0.039	0.037
	(0.025)	(0.029)	(0.032)	(0.020)
EE	0.037	-0.019	0.082	0.050
	(0.037)	(0.027)	(0.066)	(0.066)

Table 3.8: Estimates of the average marginal effects of interviewer's expected response rate over complete-case subsample by country through BBMA_AIC approach

Country	thinc2	ypen1	bacc	home
ES	0.042	0.092 **	0.098 **	0.069 *
	(0.030)	(0.028)	(0.034)	(0.035)
IT	0.101 **	0.098 **	0.262 **	0.017
	(0.017)	(0.016)	(0.025)	(0.021)
GR	0.075 *	0.107 **	-0.085	-0.043
	(0.035)	(0.039)	(0.049)	(0.042)
PT	-0.036	0.020	-0.033	0.132 *
	(0.042)	(0.034)	(0.049)	(0.056)
PL	0.065	0.145 **	0.085	-0.006
	(0.050)	(0.041)	(0.061)	(0.065)
SI	0.144 **	0.039	0.194 **	0.222 **
	(0.024)	(0.021)	(0.029)	(0.028)
AT	0.014	0.045 **	0.026	0.038
	(0.019)	(0.017)	(0.022)	(0.031)
DE	0.040 *	0.029 *	0.093 **	0.036
	(0.016)	(0.015)	(0.018)	(0.021)
BE	0.025	0.057 **	0.035	0.047 *
	(0.016)	(0.017)	(0.020)	(0.020)
LU	0.083 *	0.067	0.154 **	0.016
	(0.039)	(0.043)	(0.040)	(0.044)
SE	0.002	-0.017	0.024	0.038 *
	(0.023)	(0.025)	(0.028)	(0.019)
EE	0.040	-0.050 *	0.119 **	0.112 **
	(0.025)	(0.020)	(0.032)	(0.034)

7.2 Appendix B: Technical questions in regular and interviewer SHARE wave 6 questionnaires

Here, we document the some questions posed to respondents and interviewers that inform the formulation of dependent and independent variables in our model within the regular SHARE and interviewer wave 6 questionnaires.

Dependent variables:

HH017_TotAvHHincMonth: Total household income: How much was the overall income, after taxes and contributions, that your entire household had in an average month?

EP671_IncomeSources: Old age, early retirement, survivor pensions (components of 6 different pensions): Have you received income from any of these sources in the year?

Main public sickness benefits: They are contribution-based payments received as an earnings replacement when an employee is off sick.

Main public disability insurance pension: if the sickness turns out to be long-standing, and a return to work is not to be expected, then the claimant will typically be transferred to a disability insurance pension (e.g., invalidity or incapacity benefit).

The term 'pension' in the heading of this category is to be meant as 'regular payment', rather than relating to old age.

Public unemployment benefit or insurance: they are received, for a limited time period, by previous employees, later finding themselves unemployed. Eligibility is based on a history of insurance contribution.

Public long-term care insurance: it includes cash payments meant to provide for long-term care needs; receipt does not necessarily depend on having previously paid contributions.

Social assistance: it includes cash or voucher programs meant to provide a general 'safety net', guaranteeing minimum resources to those otherwise lacking resources from either employment or contributory-based social security benefits/pensions.

AS003_AmBankAcc: Bank accounts: About how much do you [and/ and/ and/ and] [your/ your/ your/ your] [husband/ wife/ partner/ partner] currently have in bank accounts, transaction accounts, saving accounts, or postal accounts?

HO024_ValueH: Value of the main residence: In your opinion, how much would you receive if you sold your property today?

Independent variables:

Expected response rate to income (IW ERR income): Social surveys very often ask about respondents' income. What do you expect, how many of your respondents (in percentage) in SHARE will provide information about their income? Percentage 0..100 -1 I don't know, -2 I refuse to say

R and IW Self-reported health: Would you say your health is: 1 Excellent, 2 Very good, 3 Good, 4 Fair, 5 Poor, -1 I don't know, -2 I refuse to say

Below follow two statements about difficult respondents and contact attempts. We would like to know how you react in the following situations.

IW clarifies questions: If the respondent doesn't understand a question, I explain what the question really means: 1 Perfectly, 2 Somewhat, 3 Not really, 4 Not at all, Don't know, Refuse to say

IW speak fast: If I notice that the respondent is in a hurry, I speak faster: 1 Perfectly, 2 Somewhat, 3 Not really, 4 Not at all, Don't know, Refuse to say

Fluency score: Now, I would like you to name as many different animals as you can think of. You have one minute to do this. Ready, go. IWER: Allow one minute precisely. If the respondent stops before the end of the time, encourage him/her to try to find more words. If he/she is silent for 15 seconds, repeat the basic instruction ('I want you to tell me all the animals you can think of'). No extension on the time limit is made in the event that the instruction has to be repeated.

Numeracy 1: CF012_NumDis: If the chance of getting a disease is 10 percent, how many people out of 1,000 (one thousand) would be expected to get the disease?

CF013_NumHalfPrice: In a sale, a shop is selling all items at half price. Before the sale, a sofa costs 300. How much will it cost in the sale?

CF014_NumCar: A second-hand car dealer is selling a car for 6,000. This is two-thirds of what it costs new. How much did the car cost new? The respondent should not use paper and pencil.

CF015_Savings: Let's say you have 2,000 in a savings account. The account earns ten percent interest each year. How much would you have in the account at the end of two

years?

CF108_Numarcy subtraction: Now, let's try some subtraction of numbers. One hundred minus 7 equals what? The respondent should not use paper and pencil. If R adds 7 instead, you may repeat the question.

CF103_Memory: How would you rate your memory at the present time? Would you say it is excellent, very good, good, fair, or poor? SELF-RATED WRITING SKILLS 1. Excellent 2. Very good 3. Good 4. Fair 5. Poor

PH005-Limitation activities: For the past six months at least, to what extent have you been limited because of a health problem in activities people usually do? LIMITED ACTIVITIES 1. Severely limited 2. Limited, but not severely 3. Not limited

EURO-D: The 12 EURO-D items (depressed mood, pessimism, wishing death, guilt, sleep, interest, irritability, appetite, fatigue, concentration, enjoyment, and tearfulness) were all taken from the Geriatric Mental State [31]; each item is scored 0 (symptom not present) or 1 (symptom present), generating a simple ordinal scale

References

Anderson, D.R. and Burnham, K.P. (2002). Avoiding pitfalls when using information-theoretic methods. *The Journal of wildlife management* 66, 912–918.

Banks, J., Muriel, A., and Smith, J. (2011). Attrition and health in ageing studies: Evidence from ELSA and HRS. *Longitudinal and Life Course Studies*, 2: 1–29.

Bergmann, M., Franzese, F., and Schrank, F. (2022). Determinants of consent in the SHARE accelerometer study. *SHARE Working Paper Series*, N. 78.

Bergmann, M., Kneip, T., De Luca, G. and Scherpenzeel, A., (2017). Survey participation in the survey of health, ageing and retirement in Europe (SHARE), Wave 1-6. *Munich: Munich Center for the Economics of Aging.*

Berk, M. L., and Bernstein, A. B. (1988). Interviewer characteristics and performance on a complex health survey. *Social Science Research*, 17: 239–251.

Blom, A. G., and Korbmacher, J. M. (2013). Measuring interviewer effects in SHARE Germany. *SHARE Working Paper Series*, N. 3.

Buckland, S.T., Burnham, K.P. and Augustin, N.H. (1997). Model selection: an integral part of inference. *Biometrics* 53, 603–618.

Cunha, C., Matos, A.D., Voss, G. and Machado, C., 2022. Interviewer characteristics and nonresponse survey outcomes: A Portuguese case study. *In Challenges and Trends in Organizational Management and Industry. Springer*, Cham. 95–111

Dardanoni, V., Modica, S., and Peracchi, F. (2011). Regression with imputed covariates: A generalized missing-indicator approach. *Journal of Econometrics*, 162: 362–368.

Dardanoni, V., De Luca, G., Modica, S. and Peracchi, F. (2012). A generalized missing-indicator approach to regression with imputed covariates. *The Stata Journal*, 12: 575-604.

Dardanoni, V., De Luca, G., Modica, S., and Peracchi, F. (2015). Model averaging estimation of generalized linear models with imputed covariates. *Journal of Econometrics*, 184: 452–463.

De Luca, G., and Peracchi, F. (2012). Estimating Engel curves under unit and item nonresponse. *Journal of Applied Econometrics*, 27: 1076–1099.

Durrant, G. B., Groves, R. M., Staetsky, L., and Steele, F. (2010). Effects of interviewer attitudes and behaviors on refusal in household surveys. *Public Opinion Quarterly*, 74: 1–36.

Essig, L. and Winter, J.K. (2009). Item non-response to financial questions in household surveys: An experimental study of interviewer and mode effects. *Fiscal Studies*, 30: 367–390.

Fitzgerald, J., Gottschalk, P., and Moffitt, R. (1998). An analysis of sample attrition in panel data: The Michigan panel study of income dynamics. *The Journal of Human Resources*, 33: 251–299

Friedel, S. (2020). What they expect is what you get: The role of interviewer expectations in nonresponse to income and asset questions. *Journal of Survey Statistics and Methodology*, 8: 851–876.

Friedel, S., Bethmann, A., and Kronenberg, M. (2019). The third round of the SHARE interviewer survey. *SHARE Wave 7 Methodology: Panel Innovations and Life Histories*, 101–106.

Groves, R. M., Couper, M. P. (1998). Nonresponse in household interview surveys. *John Wiley and Sons.*

Korbmacher, J. M., Friedel, S., Wagner, M., and Krieger, U. (2013). Interviewing interviewers: The SHARE interviewer survey. *SHARE Wave 5: Innovations & Methodology*, 67–74.

Lipps, O., and Pollien, A. (2011). Effects of interviewer experience on components of nonresponse in the European Social Survey. *Field Methods*, 23: 156–172.

Lynn, P., Sinibaldi, J. and Tipping, S. (2013). The effect of interviewer experience, attitudes, personality and skills on respondent co-operation with face-to-face surveys. *Survey Research Methods*, 7: 1–15.

Magnus, J. R., and De Luca, G. (2016). Weighted-average least squares (WALS): A survey. *Journal of Economic Surveys* 30: 117–148.

Meng, X.L. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statistical Science* 9, 538–558.

Nicoletti, C., and Peracchi, F. (2005). Survey response and survey characteristics: microlevel evidence from the European Community Household Panel. *Journal of the Royal Statistical Society*, 168: 763–781.

Olson, K. (2014). Do non-response follow-ups improve or reduce data quality? A review of the existing literature. *Quality Control and Applied Statistics*, 59: 61–62.

Pickery, J., and Loosveldt, G. (2001). An exploration of question characteristics that mediate interviewer effects on item nonresponse. *Journal of Official Statistics*, 17: 337–350

Raftery, A.E. (1996). Approximate Bayes factors and accounting for model uncertainty in generalised linear models. *Biometrika*, 83, 251–266.

Riphahn, R.T., and Serfling, O. (2005). Item non-response on income and wealth questions. *Empirical Economics*, 30: 521–538.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63, 581–592

Schaeffer, N. C., Dykema, J., and Maynard, D. W. (2010). Interviewers and interviewing. *Handbook of Survey Research*, 2: 437–471.

Schafer, J. L. (1997). Analysis of Incomplete Multivariate Data. Boca Raton, FL: Chapman & Hall/CRC

Schraepler, J. P. (2006). Explaining income nonresponse-a case study by means of the British Household Panel Study (BHPS). *Quality and Quantity*, 40: 1013-1036.

Silber, H., Roßmann, J., Gummer, T., Zins, S. and Weyandt, K.W. (2021). The effects of question, respondent and interviewer characteristics on two types of item nonresponse. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 184: 1052–1069.

Singer, E. and Kohnke-Aguirre, L. (1979). Interviewer expectation effects: A replication and extension. *Public Opinion Quarterly*, 43: 245–260.

Sudman, S., Bradburn, N.M., Blair, E.D., and Stocking, C. (1977). Modest expectations: The effects of interviewers' prior expectations on responses. *Sociological Methods & Research*, 6: 171–182.

Tourangeau, R., and Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin*, 133: 859–883.

Van Buuren, S., Brand, J.P., Groothuis-Oudshoorn, C.G. and Rubin, D.B. (2006). Fully conditional specification in multivariate imputation. *Journal of statistical computation and simulation*, 76, 1049–1064.

van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research*, 16, 219–242.

Vercruyssen, A., Wuyts, C., and Loosveldt, G. (2017). The effect of sociodemographic (mis)match between interviewers and respondents on unit and item nonresponse in Belgium. *Social Science Research*, 67: 229–238.

Von Hippel, P.T., (2020). How many imputations do you need? A two-stage calculation using a quadratic rule. *Sociological Methods & Research*, 49, 699–718.

West, B. T. and Blom, A. G. (2017). Explaining interviewer effects: A research synthesis. *Journal of Survey Statistics and Methodology*, 5: 175–211.

Wuyts, C., and Loosveldt, G. (2017). The Interviewer in the respondent's shoes: What can We learn from the way interviewers answer survey questions?. *Field Methods*, 29: 140–153.

Chapter 4

Analysis response propensity score using WALS in SHARE-HCAP

Abstract

Propensity-score adjustment is a widely used technique for handling nonresponse errors in sample surveys. In this study, we estimate propensity scores for the SHARE-HCAP study that can be used to construct nonresponse weights. Traditionally, single-specification models have been employed to compute propensity scores. However, with the availability of a large set of predictors collected across multiple waves, such as Waves 6, 7, and 8, and the SHARE-HCAP fieldwork, researchers can consider the strategy of using a large set of variables. Nevertheless, it is noted that less parsimonious models might introduce larger variability in predicted probabilities and their corresponding propensity score weights. Shrinkage estimators come into play to tackle this trade-off between bias and precision. Due to its computational efficiency, this study uses the Weighted Average Least Squares (WALS) method, among other model averaging techniques. Our findings underscore the influential role of cognitive abilities in self-selection processes and subsequent nonresponse errors. Incorporating these cognitive variables demonstrates the potential for crafting more accurate nonresponse weights.

Keywords: Nonresponse errors, propensity score methods, shrinkage estimators, WALS method

JEL classification: To follow

1 Introduction

Survey research encounters a persistent challenge in nonresponse bias, where selected individuals decline or fail to respond to survey requests, potentially bias sample estimates and compromising the representativeness of findings (Groves, 2006). These errors can significantly impact the accuracy and validity of survey results, necessitating robust strategies to address nonresponse bias effectively.

In addressing this issue, propensity score methods have emerged as a valuable tool to account for the selection process of respondents and nonrespondents. These methods involve estimating the probability of response based on observed covariates, facilitating adjustments in survey weights and imputing missing values (Little and Rubin, 2019).

While these methods have found wide application across various disciplines, including economics, social sciences, and public health, our study primarily focuses on the implementation of parametric propensity score models, notably logistic regression, to estimate the likelihood of response based on observed characteristics (Brick, 2013). We acknowledge the potential limitations of these models, as they may not fully capture the complexities of the data-generating processes and can lead to biased estimates if the underlying assumptions are violated. We should say that there is an increasing attention towards nonparametric methods, such as kernel-based and machine-learning techniques, which offer greater flexibility in modeling complex relationships and demonstrate reduced sensitivity to distributional assumptions (Da Silva and Opsomer, 2006; Earp et al., 2018, 2014; Ferri-García and del Mar Rueda, 2020; Buskirk and Kolenikov, 2015; Cham and West, 2016; Griffin et al., 2017).

Our study employs the model averaging method to estimate propensity scores, particularly where the model uncertainty issue arises. With the availability of a large set of

predictors collected across multiple waves, such as Waves 6, 7, and 8, researchers can consider the strategy of using a large set of variables in the single specification propensity score models. It is noted that less parsimonious models might introduce larger variability in predicted probabilities and their corresponding propensity score weights. Additionally, suppose we have two models to choose from: unrestricted and restricted. No matter which one we choose, we are making a trade-off. If we go with the unrestricted model, our results might be less biased, but the results might not be very precise. If we choose the restricted model, it might be the other way around– our results might be more precise (lower variance), but the results might be biased.

To effectively handle the challenge of model uncertainty in the nonresponse adjustment process, we adopt the Weighted-Average Least Squares (WALS) method, as proposed by Magnus et al. (2010). This method serves as a hybrid model-averaging approach, incorporating both frequentist and Bayesian perspectives, allowing us to address the challenge of model uncertainty effectively.

By leveraging the WALS estimator, our study innovatively addresses model uncertainty while estimating predicted response probabilities, significantly reducing computation time compared to other model averaging methods, especially Bayesian model averaging (BMA).

The next parts of this paper are carefully structured in the following way: In Section 2, we provide a brief overview of the SHARE-HCAP data used in our study, along with an explanation of the sample design. The "Methodology" (section 3) explains the WALS estimator for the logistic regression model in detail. The "Results" (section 5) presents the findings from our analyses and the model estimations. Lastly, the "Conclusions" (section 6) summarizes this study's important findings and conclusions.

2 SHARE-HCAP Data and sample design

This section entails a concise exploration of the SHAR-HCAP dataset, providing a brief overview of the gross sample under investigation in this study. It also offers a succinct overview of the key explanatory variables used in our models, along with a concise description of the sample design employed in our study.[1]

2.1 SHARE-HCAP data

SHARE-HCAP is a new research protocol that augments the cognitive measures collected in the regular waves of the SHARE panel with a more detailed set of cognitive tests and health assessments.[2] In our study, SHARE-HCAP study covers about 2,651 respondents from five European countries (Czech Republic, Denmark, France, Germany, and Italy) who are aged 65+ years in 2022 and have participated in one of the last three regular waves of SHARE (either Wave 6, Wave 7, or Wave 8), and were eligible for SHARE Wave 9. SHARE-HCAP study aims to assess cognitive status in the sample study and then use that information to extrapolate the cognitive status of the full SHARE population. The study also calculates the prevalence rates of moderate and low cognitive impairment in each country. It compares these rates to those obtained from other studies, including the HRS and HCAP studies. Further, the study examines the impact of biomedical and socioeconomic factors on late-life cognition.

[1] SHARE HCAP technical report - sampling. Bergmann and Bethmann (2022).

[2] The interview protocols and cognitive measures of the SHARE-HCAP are harmonized with those of the Health and Retirement Study (HRS) in the USA. In addition to HRS-HCAP and SHARE-HCAP studies, the HCAP study is conducted in several regions and countries, including Chile, China, the Caribbean, England, Africa, Ireland, India, Mexico, and Northern Ireland, under different names such as Chile-Cog, CHARLS, CADAS, ELSA-HCAP, HAALSI Dementia Study, TILDA, LASI-DAD, Mex-Cog, and NICOLA.

In Wave 9 of the SHARE-HCAP study, the sample frame consists of 12,847 units. At last, out of these 12,847 units, 3,880 respondents, referred to as the gross sample, were contacted by SHARE to inquire about their interest in participating in the SHARE-HCAP study interview.

The sample sizes available across each country exhibit a range, with a minimum of 487 observations in the case of the Czech Republic and a maximum of 565 observations recorded in Denmark (Table 4.1). Notice that participation in one of the last three regular waves was used to define the SHARE-HCAP sampling frame and sampling design, while eligibility for Wave 9 was used to account for data deletion requests and notified deaths before the start of the SHARE-HCAP fieldwork. For households with two or more SHARE panel respondents, eligibility for SHARE-HCAP was further restricted to only one respondent per household selected randomly.

The fact that the SHARE-HCAP sampling frame was constructed from the pool of respondents in the last three regular waves of SHARE has clear advantages and disadvantages. On the one hand, it implies that the underlying gross sample (see Table 4.1) may suffer from selection effects due to unit nonresponse and attrition in the previous waves of the SHARE panel. On the other hand, the underlying design weights can be computed for all units of the gross sample. One can exploit a considerably larger set of auxiliary variables when accounting for the unit nonresponse errors that occurred in the study's first wave.

Table 4.1: Number of respondents, participants, and participation rates in the gross sample SHARE-HCAP by country

		SHARE-HCAP interview	
Country	Total	Obs.	PR
CZ	739	487	0.659
DE	801	544	0.679
DK	868	565	0.651
FR	706	524	0.742
IT	766	531	0.693
Total	3880	2,651	0.683

Note: PR denotes the participation rate.

We also report, separately, the sample averages and standard deviations of the explanatory variables for respondents who participate in the survey and those who do not. Simple t-tests of the differences in the averages between these two groups provide a prima facie evidence of the relevance of non-random non-response shown in Table 4.2. The t-statistic indicates that factors such as age, orientation in time ability, and some other variables played a role in the respondent's decision not to participate, suggesting reasons beyond random chance.

Table 4.2: Description of statistics of explanatory variables in response and nonresponse samples

	Response		Nonresponse		
Variable	Mean	SD	Mean	SD	t-stat
Female	1.54	0.498	1.56	0.496	1.20
Age	76.34	7.480	77.12	7.922	2.96
Education level: high	0.64	0.481	0.61	0.488	-1.40
Living with partner	0.40	0.490	0.40	0.490	-0.09
Recalling words in memory	2.38	0.741	2.34	0.787	-1.30
Orintation in time: good	0.89	0.317	0.85	0.356	-3.10
Numeracy score 1: good	0.83	0.373	0.81	0.391	-1.57
Numeracy subtr. score: good	0.88	0.326	0.86	0.351	-2.02
Memory: good	0.72	0.449	0.68	0.468	-2.70
Fluency	20.19	7.852	19.88	7.964	-1.16
Self-rated health: good	0.63	0.483	0.59	0.492	-2.39
Limitation activities	0.50	0.500	0.45	0.498	-2.82
Number of chronic disease	0.16	0.365	0.17	0.371	0.59
Activities of daily living	0.88	0.330	0.82	0.384	-4.58
Instr. act. of daily living	0.81	0.394	0.74	0.438	-4.74
Depression	0.23	0.419	0.21	0.411	-0.85
Eyesight reading: good	0.85	0.355	0.83	0.377	-1.91
Hearing: good	0.78	0.418	0.76	0.426	-0.88
Body mass index	2.86	0.782	2.80	0.796	-2.05
participation in w6 and w7	0.92	0.271	0.89	0.312	-3.00
Willingness to answer: good	0.92	0.265	0.88	0.325	-4.45
N		2651		1229	

.

The sample design of the SHARE-HCAP study was stratified based on household type (single or couple) and cognitive impairment level (i.e., immediate word recall and delayed word recall tests, indicators of low, moderate, or healthy). In addition, self-reported dementia diagnosis was used to categorize respondents into low group if respondents were diagnosed with Alzheimer's disease, dementia, organic brain syndrome, senility, or any other serious memory impairment by a doctor.

3 Weighted-Average Least Squares (WALS) estimator

Our study employs the model averaging method to estimate propensity scores where the model uncertainty problem exists. This approach contributes all available information to address model uncertainty, a concept first discussed by Leamer (1978). Amid various dimensions of model uncertainty, such as functional forms, we underscore the role of covariate selection as a significant source of uncertainty. In this section, we present the pertinent statistical theory within our context: The WALS, as introduced by Magnus *et al.* (2010) and developed in various papers (see, e.g., Magnus and De Luca 2016; De Luca *et al.* 2018, 2021, 2023).

3.1 Statistical framework

We consider modeling a data matrix $[y : X]$ consisting of n observations on a scalar outcome and k regressors. Thus, y is an n-vector with ith element y_i, and X is an $n \times k$ matrix of full column-rank k with ith row x_i'. As in a standard logit setup, we assume that the elements of y are realizations of n independently distributed random variables with mean, finite nonzero

variance, and Bernoulli distribution as following form

$$f(y_i;\beta,X)=\pi_i^{y_i}(\beta,X)(1-\pi_i(\beta,X))^{n-y_i}\quad(i=1,2,\cdots,n)\quad y\in\{0,1\},\tag{1}$$

by the properties of the Bernoulli distribution, the mean and variance of y_i are equal to $\mu_i=\pi_i(\beta,X)$ and $\sigma_i^2=\pi_i(\beta,X)(1-\pi_i(\beta,X))$ (McCullagh and Nelder, 1989).

We depart from a standard logit setup by allowing for uncertainty in the specification of the linear predictor. Specifically, we partition the k regressors into two subsets, $X=[X_1:X_2]$, where X_p is an $n\times k_p$ matrix with the ith row equal to x_i' for $p=1,2$ and $k_1+k_2=k$. The k_1 columns of X_1 contain the regressors that we want in the model for theoretical or other grounds (focus regressors in the terminology of Danilov and Magnus, 2004), while the k_2 columns of X_2 contain the additional regressors of which we are less certain (auxiliary regressors). Stacking the linear predictors for the n observations on top of each other gives the n-vector $\eta(\beta)=X\beta=X_1\beta_1+X_2\beta_2$, with $\beta=(\beta_1',\beta_2')'$, where β_1 is the vector of focus parameters and β_2 is the vector of auxiliary parameters.

In total, 2^{k_2} possible models contain all focus regressors and arbitrary subsets of auxiliary regressors. We represent the jth model as a logistic regression with the added restriction $R_j'\beta_2=0$, where R_j denotes a $k_2\times r_j$ matrix of rank $0\leqslant r_j\leqslant k_2$ such that $R_j'=[I_{r_j}:0]$ (or a column-permutation thereof) and I_{r_j} denotes the identity matrix of order r_j. The matrix R_j thus specifies which auxiliary regressors are excluded from the jth model, and the scalar r_j denotes the number of excluded auxiliary variables.

We assume that $\{(y_i,x_i')'\}_{i=1}^n$ are independent observations, and the log-likelihood of the

logistic regression model is of the form

$$\ell(\beta) = \sum_{i=1}^{n} [y_i \log(\pi_i(\beta, X)) + (1 - y_i) \log(1 - \pi_i(\beta, X))] \tag{2}$$

Since $x_i = (x'_{i1}, x'_{i2})'$ and $\beta = (\beta'_1, \beta'_2)'$, the gradient of the log-likelihood (the score) is the k-vector $s(\beta)$ consisting of the following subvectors:

$$s_p(\beta) = \frac{\partial \ell(\beta)}{\partial \beta_p} = \sum_{i=1}^{n} [y_i - \pi_i(\beta)] \, x_{ip} \quad (p = 1, 2), \tag{3}$$

We also define a $k \times k$ matrix $H(\beta)$, which is equal to minus the Hessian of the log-likelihood and consists of the following submatrices:

$$H_{pq}(\beta) = -\frac{\partial^2 \ell(\beta)}{\partial \beta_p \partial \beta_q} = \sum_{i=1}^{n} [\pi_i(\beta)(1 - \pi_i(\beta))] x_{ip} x'_{iq} \quad (p, q = 1, 2). \tag{4}$$

3.2 One-step ML estimators

The Maximum Likelihood (ML) estimator of β for the j-th model maximizes the log-likelihood $\ell(\beta)$ subject to the constraint $R'_j \beta_2 = 0$, or equivalently, it solves the system of $k_1 + k_2 + r_j$ equations:

$$\begin{aligned} 0 &= s_1(\beta), \\ 0 &= s_2(\beta) - R_j \nu_j, \\ 0 &= R'_j \beta_2, \end{aligned} \tag{5}$$

where ν_j denotes the r_j-vector of Lagrange multipliers associated with the constraint $R'_j \beta_2 = 0$. One challenge in extending the WALS approach to logistic regression models is that unless the elements of y are normally distributed, the system of likelihood equations (5) are

nonlinear and must be solved through an iterative method such as Newton-Raphson or the scoring method. To address this issue, De Luca *et al.* (2018) exploited the class of one-step ML estimators that admit closed-form expressions and are asymptotically equivalent to (fully-iterated) ML estimators.

Given a starting value $\bar{\beta} = (\bar{\beta}_1', \bar{\beta}_2')'$, with properties to be discussed below, expanding the likelihood equations (5) around $\bar{\beta}$ yields the approximation:

$$\begin{aligned} 0 &= \bar{s}_1 - \bar{H}_{11}(\beta_1 - \bar{\beta}_1) - \bar{H}_{12}(\beta_2 - \bar{\beta}_2), \\ 0 &= \bar{s}_2 - \bar{H}_{21}(\beta_1 - \bar{\beta}_1) - \bar{H}_{22}(\beta_2 - \bar{\beta}_2) - R_j\nu_j, \\ 0 &= R_j'\beta_2, \end{aligned} \tag{6}$$

where $\bar{s}_p = s_p(\bar{\beta})$ and $\bar{H}_{pq} = H_{pq}(\bar{\beta})$, $p, q = 1, 2$. An estimator $\hat{\beta}_j$ that solves the linearized system of constrained likelihood equations (6) is termed a one-step ML estimator of β under the jth model, as it corresponds to the first step of the Newton-Raphson method. The one-step ML estimators have been used in many different studies; see, e.g., Janssen *et al.* (1985) (M-estimation), Rothenberg (1984) (generalized least squares), Frazier and Renault (2017) (efficient two-step estimation) and Gupta (2023) (spatial autoregressions), among others.

Let us consider for simplicity the unrestricted model where $R_j = 0$ and define the data transformations:

$$\bar{y} = \bar{X}_1\bar{\beta}_1 + \bar{X}_2\bar{\beta}_2 + \bar{u}, \quad \bar{X}_1 = \bar{D}^{1/2}X_1, \quad \bar{X}_2 = \bar{D}^{1/2}X_2, \tag{7}$$

where $\bar{u} = \bar{D}^{-1/2}(y - \bar{\pi})$, $\bar{D} = D(\bar{\beta})$ is an $n \times n$ diagonal matrix with diagonal element equal to $\pi_i(\bar{\beta})(1 - \pi_i(\bar{\beta}))$, and $\bar{\pi} = \pi(\bar{\beta})$ is an n-vector with the ith element equal to $\pi_i(\beta)$. Then,

the solutions $\hat{\beta}_{1u}$ and $\hat{\beta}_{2u}$ to the linearized system of likelihood equations (6) can be written in closed form as:

$$\hat{\beta}_{1u} = (\bar{X}_1'\bar{X}_1)^{-1}\bar{X}_1'\bar{y} - (\bar{X}_1'\bar{X}_1)^{-1}\bar{X}_1'\bar{X}_2\hat{\beta}_{2u},$$

$$\hat{\beta}_{2u} = (\bar{X}_2'\bar{M}_1\bar{X}_2)^{-1}\bar{X}_2'\bar{M}_1\bar{y},$$

where $\bar{M}_1 = I_n - \bar{X}_1(\bar{X}_1'\bar{X}_1)^{-1}\bar{X}_1'$ is a symmetric idempotent matrix of rank $n-k_1$. These expressions show that the unrestricted one-step ML estimators $\hat{\beta}_{1u}$ and $\hat{\beta}_{2u}$ coincide numerically with the least squares coefficients in the linear regression of $\bar{y}$ on $\bar{X}_1$ and $\bar{X}_2$. Proposition 1 of De Luca *et al.* (2018) shows that this result extends to the (constrained) one-step ML estimators of the jth model under the restriction $R_j'\beta_2 = 0$. Hence, the WALS estimator of a logit model can be easily derived from the standard WALS approach to linear regression models after applying the data transformations in (7).

3.3 WALS estimation of linear regression models

Thus motivated, we now consider WALS estimation of a linear regression model for the transformed outcome $\bar{y}$ with linear predictor $\eta = \bar{X}_1\beta_1 + \bar{X}_2\beta_2$. The k_1 columns of $\bar{X}_1 = \bar{D}^{1/2}X_1$ are the transformed focus regressors which we want in the model on theoretical or other grounds, while the k_2 columns of $\bar{X}_2 = \bar{D}^{1/2}X_2$ contain the transformed auxiliary regressors of which we are less certain. As in the original setup of our logit model, there are 2^{k_2} possible models that contain all (transformed) focus regressors and arbitrary subsets of the (transformed) auxiliary regressors.

Unlike other model-averaging approaches, WALS relies on a preliminary semi-orthogonal transformation of auxiliary predictors in $\bar{X}_2$ and the associated vector of auxiliary parameter β_2 that greatly reduces the computational burden from order 2^{k_2} to order k_2. Specifically,

we implement the following one-to-one transformations:

$$\bar{Z}_2 = \bar{X}_2\bar{\Delta}_2\bar{\Psi}^{-1/2}, \qquad \gamma_2 = \bar{\Psi}^{1/2}\bar{\Delta}_2^{-1}\beta_2, \tag{8}$$

where $\bar{\Delta}_2$ is a diagonal $k_2 \times k_2$ matrix that ensures that the diagonal elements of the positive definite matrix $\bar{\Psi} = \bar{\Delta}_2'\bar{X}_2'\bar{M}_1\bar{X}_2\bar{\Delta}_2/n$ is equal to one, where $\bar{\Psi}^{1/2}$ denotes the unique square root of $\bar{\Psi}$. We also rescale the (transformed) focus regressors $\bar{X}_1$ and the associated vector of focus coefficients β_1:

$$\bar{Z}_1 = \bar{X}_1\bar{\Delta}_1, \qquad \gamma_1 = \bar{\Delta}_1^{-1}\beta_1, \tag{9}$$

where $\bar{\Delta}_1$ is a diagonal $k_1 \times k_1$ matrix such that the diagonal elements of $\bar{Z}_1'\bar{Z}_1/n$ are all equal to one.

Since $\bar{Z}_1\gamma_1 = \bar{X}_1\beta_1$ and $\bar{Z}_2\gamma_2 = \bar{X}_2\beta_2$, the linear predictor of the unrestricted model can be rewritten equivalently as $\eta = \bar{Z}_1\gamma_1 + \bar{Z}_2\gamma_2$.

This is convenient because it implies that $\bar{Z}_2'\bar{M}_1\bar{Z}_2/n = I_{k_2}$. The one-step ML estimators for the jth model are given by

$$\hat{\gamma}_{1j} = \hat{\gamma}_{1r} - \bar{Q}W_j\hat{\gamma}_{2u}, \quad \hat{\gamma}_{2j} = W_j\hat{\gamma}_{2u}, \tag{10}$$

where $\hat{\gamma}_{1r} = (\bar{Z}_1'\bar{Z}_1)^{-1}\bar{Z}_1'y$ is the least squares (LS) estimator of γ_1 in the fully restricted model (when $\gamma_2 = 0$), $\hat{\gamma}_{2u} = \bar{Z}_2'\bar{M}_1\bar{y}/n$ is the LS estimator of γ_2 in the unrestricted model, $Q = (\bar{Z}_1'\bar{Z}_1)^{-1}\bar{Z}_1'\bar{Z}_2$, and $W_j = I_{k_2} - R_jR_j'$ is a diagonal matrix with diagonal elements equal to zero or one.

The WALS estimators of γ_1 and γ_2 are obtained by averaging the LS estimators in (10) over the full set of 2^{k_2} models that contain all focus regressors and a subset of the auxiliary

regressors:

$$\widetilde{\gamma}_1 = \sum_{j=1}^{2^{k_2}} \lambda_j \hat{\gamma}_{1j}, \qquad \widetilde{\gamma}_2 = \sum_{j=1}^{2^{k_2}} \lambda_j \hat{\gamma}_{2j}, \tag{11}$$

where the λ_j are non-negative data-dependent model weights that depend only on $\sqrt{n}\hat{\gamma}_{2u}$ and add up to one.

Since the dependence of $\hat{\gamma}_{1j}$ and $\hat{\gamma}_{2j}$ on the model index j is fully captured by the diagonal matrix W_j, we can also write (11) as follows:

$$\widetilde{\gamma}_1 = \hat{\gamma}_{1r} - \bar{Q}\widetilde{\gamma}_2, \qquad \widetilde{\gamma}_2 = W\hat{\gamma}_{2u}, \tag{12}$$

where $W = \sum_{j=1}^{2^{k_2}} \lambda_j W_j$ is a random diagonal matrix whose k_2 diagonal elements w_h (the 'WALS weights') are partial sums of the model weights λ_j, and random vector $W\hat{\gamma}_{2u}$ is asymptotically independent of $\hat{\gamma}_{1r}$.

In proposition 3, De Luca *et al.* (2018) extend the equivalence theorem that Magnus and Durbin (1999) and Danilov and Magnus (2004) proposed for the finite-sample results. The *Asymptotic Equivalence Theorem* states that the WALS estimator $\widetilde{\gamma}_1$ will be a "good" estimator of γ_1 (in the mean square error sense) if and only if $\widetilde{\gamma}_2$ is a "good" estimator of γ_2. That is, if we can find λ_i's such that $W\hat{\gamma}_{2u}$ is an "optimal" estimator of γ_2, then the same λ_i's will provide an "optimal" estimator of γ_1.

The components of $\widetilde{\gamma}_2 = W\hat{\gamma}_{2u}$ are shrinkage estimators of the components of γ_2 as $0 \leqslant w_h \leqslant 1$, and the components of $\hat{\gamma}_{2u}$ are asymptotically independent. If we assume that w_h depends only on the h-th component of $\hat{\gamma}_{2u}$ (see Magnus and De Luca 2016), the shrinkage estimators in $\widetilde{\gamma}_2$ are also asymptotically independent. As a result, our k_2-dimensional problem reduces to k_2 (identical) one-dimensional problems: given one observation $x \sim \mathcal{N}(\theta, \sigma^2/n)$,

we look for a shrinkage estimator $m(x)$ of θ with good MSE properties. This is the so-called normal location problem. Given that σ is assumed to be known, we can set this parameter equal to 1 without loss of generality (see Danilov 2005).

We follow a Bayesian approach because of concerns related to the admissibility of our shrinkage estimator. In addition to x, this approach requires to specify a prior distribution for the location parameter θ.

Magnus and De Luca (2016) focus on the family of reflected generalized gamma distributions to address the issue of choosing the prior for the Bayesian step. Laplace, Weibull, and Subbotin priors have been employed in many previous studies, and each one has its own advantages and disadvantages. In this study, we exploit the Laplace and the horseshoe priors. We use the horseshoe prior (see Carvalho *et al.* 2010) because the Laplace prior is not robust, while the horseshoe prior has a high degree of robustness. As theoretical t-ratios in the normal location model increase with sample size, the asymptotic estimation bias converges to zero only for the horseshoe prior (not for the Laplace). The Bayesian posterior mean is used to obtain the optimal estimator of θ in the normal location model.

3.4 Implications for WALS

For each component x_h of x, the Bayesian approach allows to estimate θ_h by the posterior mean $m_h = \mathrm{m}(x_h)$ $(\mathrm{h} = 1, \cdots, k_2)$. The WALS estimators of γ_1 and γ_2 are then given by:

$$\widetilde{\gamma}_1 = \hat{\gamma}_{1r} - \bar{Q}\widetilde{\gamma}_2, \qquad \widetilde{\gamma}_2 = \sigma m,$$

where $\mathrm{m} = (m_1, \cdots, m_{k_2})$. After estimating the WALS parameters γ_1 and γ_2, we can use the transformations in equations (8) and (9) to obtain the WALS estimators of the original

parameters β_1 and β_2. These estimators are given by $\widetilde{\beta}_1 = \bar{\Delta}_1\widetilde{\gamma}_1$ and $\widetilde{\beta}_2 = \bar{\Delta}_2\bar{\Psi}^{-1/2}\widetilde{\gamma}_2$.

As in De Luca *et al.* (2018), it is important to note that in the context of the logit model, the focus often shifts towards estimating and inferring marginal effects-that is a nonlinear function, $g(\beta; x)$. From a frequentist perspective, ML estimation of each possible model yields a set of 2^{k_2} conditional ML estimates $\hat{\beta}_j$, from which we obtain the conditional ML estimates $\hat{g}_j = g(\hat{\beta}_j; x)$ of $g(\beta; x)$. The key issue is how to combine them best to construct an unconditional estimate of $g(\beta; x)$ that incorporates the uncertainty due to the model selection and estimation steps. The standard frequentist model averaging (FMA) solution is an estimator of the form.

$$\widetilde{g}_{\text{ma}} = \sum_{j=1}^{2^{k_2}} \lambda_j^* \hat{g}_j, \tag{13}$$

where the λ_j^* are model weights chosen based on some optimality criterion (see, e.g., Hjort and Claeskens, 2003). BMA estimators have a similar form, that is, they are a weighted average of the means of the conditional posterior distributions of $g(\beta; x)$ under each possible model with weights equal to the posterior model probabilities (see, e.g., Hoeting *et al.* 1999). Unfortunately, in WALS, we cannot construct the model-averaging estimator in (13) due to lack of information on the $\widetilde{g}_j$ and the λ_j^*. This is a consequence of the semi-orthogonal transformation (8) which leads to important simplifications when estimating β, but also implies some loss of flexibility compared to standard FMA and BMA approaches. Here, loss of flexibility means that we can only compute a model-averaging estimator $\widetilde{\beta}$ of β, that is $\widetilde{\beta} = \sum_{j=1}^{2^{k_2}} \lambda_j \hat{\beta}_j$, on the basis of which we then obtain a plug-in estimator $\widetilde{g}_{\text{pi}} = g(\widetilde{\beta}; x)$ of $g(\beta; x)$. Thus, instead of averaging over nonlinear transformations of the ML estimators, we can only apply a nonlinear transformation of the model-averaging estimator of β.

4 Choice of focus and auxiliary regressors

Estimation is always performed separately by country using as focus regressors a set of socio-demographic variables and a set of cognitive ability measures from the previous regular waves of SHARE. The socio-demographic variables include a quadratic polynomial in age, plus binary indicators for gender, high level of education, whether living with a spouse/partner and NUTS1 (Nomenclature of Territorial Units for Statistics, level 1) regional area. The cognitive ability measures include a quadratic polynomial in the verbal fluency test scores (as a continuous variable), a three-level factor for the combined score of the immediate and delayed word recall tests (low, moderate, and healthy), and a set of binary indicators for good orientation in time, good performance in two numeracy tests, and good self-rated memory. As discussed in Section 2, the binary indicator for living with a spouse/partner and the three-level factor from the word recall tests have also been used as stratum variables in the national sampling designs of SHARE-HCAP. More generally, early studies (Van Beijsterveldt *et al.* 2002; Salthouse, T.A. 2014) suggest that our focus regressors can be highly correlated with the more refined cognitive measures collected in this study. Hence, the fact of controlling for these regressors helps to capture possible selection effects that nonresponse errors may have on the key survey variables of interest.

Our set of auxiliary regressors includes variables for physical health status, conditions of the interview process in the previous waves, a set of interaction terms between socio-demographic characteristics and cognitive abilities measures, and an additional set of interaction terms between the different measures of cognitive abilities. In particular, physical health status is measured by a set of binary indicators for good self-perceived health, no chronic disease, no activity limitations, no limitations in daily living and instrumental ac-

tivities, no depression symptoms, good eyesight, good hearing, and four categories of body mass index ($<$ 18.5=underweight, [18.5,25)=health weight, [25, 30)=overweight, and $\geqslant$ 30=obese). Interview conditions in previous waves are captured by binary indicators for participation patterns in waves 6-8 and good willingness to answer. Here, we deliberately omit the interaction terms involving cognitive and socio-demographic variables for couples and NUTS1 regions.

In our study, because we estimate the predicted response probability and focus on prediction, Thus, both β_1 and β_2 are important. Unlike the restricted model, the unrestricted model allows us to assess whether interactions of different cognitive domains are important predictors of the response probability.

Our modeling framework encompasses different countries, each with its own regional classifications (NUTS1). The number of the focus variables varies across countries due to differences in the NUTS1 regional area (between a minimum of $k_1 = 12$ variables in CZ and DK and a maximum of $k_1 = 27$ variables in DE), while the number of the auxiliary variables is equal to $k_2 = 72$ variables in all countries. Although the model space contains 2^{72} different models, the computational burden of the WALS estimator is order $k_2 = 72$. Compared to other model averaging estimators, this is a great computational advantage.

5 Results

This section presents and compares the findings of four approaches for the propensity score model: the ML estimates of the restricted model, which includes only the subset of focus regressors in X_1, the ML estimates of the unrestricted model which includes all focus and auxiliary regressors in X_1 and X_2, the WALS estimates based on the Laplace prior, and

the WALS based on the horseshoe prior. Since the interpretation of the model coefficients is complicated by the nonlinear nature of logit models, the large number of regressors, and the presence of interaction terms, we shall present the estimated average marginal effects in Section 5.1 and the predicted response probabilities in Section 5.2.

5.1 Estimation of Marginal Effects

Marginal effects represent the change in the predicted response probability when an independent variable undergoes a change while keeping other variables constant. Specifically, we focus on the average marginal effects (AME) of key socio-demographic variables (gender, age, education, and couple) and cognitive ability measures (moderate and healthy recall, good orientation in time, good self-rated memory, good first numeracy, good second numeracy, and verbal fluency), which are defined as the average over the estimation sample of the individual marginal effects computed at the observed values of the regressors.[3] The country-specific estimates of the AME are presented in Tables 4.3-4.7.

Consistent patterns emerge across models and countries regarding the AME of our socio-demographic and cognitive ability variables on response probabilities in the HCAP-SHARE. While there may be variations in the magnitudes of these effects, several noteworthy trends come to the forefront:

- Education positively correlates with increased participation probabilities in multiple countries (CZ, DE, FR, IT).

- Couples tend to be associated with a decreased likelihood of participation (DE, FR).

[3]We do not present the AME of the auxiliary regressors because they may be subject to issue of comparability across different models.

- Positive influences on participation probabilities are associated with cognitive abilities such as a robust orientation in time (CZ, DK, FR, IT).

- Word recall abilities present mixed effects on participation, with moderate word recall generally demonstrating a positive effect (CZ, DK, IT), while healthy word recall exhibits a negative effect (DE, IT).

Further, an approximately consistent pattern is observed in the table results, indicating that standard errors of AME derived from the two WALS methods tend to be between the restricted and unrestricted logit models. In the end, the AME masks the changes in response probabilities at the individual levels. Therefore, it is important to illustrate these changes at the individual level using graphs in Section 5.2.

5.2 Predicted response probabilities

This section presents a set of figures that illustrate the predicted response probabilities. We compute and display the confidence intervals for the WALS based on the horseshoe prior. The WALS confidence intervals for the predictions are derived and are based on the finite-sample properties of frequentist bias estimators of posterior means in the normal location model (De Luca *et al.* 2021), and these properties allow us to study the sampling distribution of the bias-corrected WALS estimator by simulations. We compute point estimates, their moments, and confidence intervals for all coefficients using a total of 1000 replications.

Figure 4.1 illustrates the age profiles of response probabilities estimated from restricted, unrestricted, and WALS models based on the horseshoe prior. Each point on the estimated age profiles corresponds to the response probabilities of one representative elderly aged a year. While the restricted and unrestricted Maximum Likelihood (ML) estimates show consider-

able differences except for DK, the WALS estimate is closely aligned with the unrestricted ML estimates. Additionally, despite a general decline in response probabilities with age, the elderlies in CZ, DE, and DK exhibit lower interview response probabilities than those in FR and IT.

Figure 4.2 presents the response probabilities for three distinct groups categorized by their word recall scores across different age ranges and five countries, estimated from different models. Although small discrepancies exist between restricted and unrestricted ML estimates for healthy subgroup, significant differences emerge in moderate and low groups. Notably, confidence intervals for the WALS estimate are wider in the low group compared to the moderate and healthy groups, indicating increased uncertainty in the presented measurements across countries.

Figures 4.3, 4.4, 4.5, and 4.6 depict the response probabilities for two different groups categorized by cognitive abilities scores (self-rated memory, numeracy 1, numeracy 2, and orientation in time) aged 65-85 estimated from various models (restricted, unrestricted, and WALS based on the horseshoe prior). Notably, substantial discrepancies exist between restricted and unrestricted ML estimates for old people with poor cognitive ability scores. Conversely, the WALS estimates closely resemble unrestricted ML estimates. Moreover, WALS confidence intervals are wider for individuals with poor cognitive functioning than those with better performance.

Figure 4.7 showcases the response probabilities of verbal fluency scores estimated from restricted, unrestricted, and WALS models across countries. While considerable differences are evident between restricted and unrestricted ML estimates, the WALS estimate closely aligns with the unrestricted ML estimates.

Figure 4.8 shows the response probabilities for three distinct groups categorized by their

word recall scores across different fluency scores and estimated from different models. Although small discrepancies exist between restricted and unrestricted ML estimates for the healthy group, significant differences emerge in moderate and low groups. Notably, confidence intervals for the WALS estimate are wider in the low group compared to the moderate and healthy groups.

Figure 4.9, 4.10, 4.11, and 4.12 provide the response probability for two different groups categorized by cognitive abilities scores (self-rated memory, numeracy 1, numeracy 2, and orientation in time) over different fluency scores estimated from various models—restricted, unrestricted, and the WALS. Notably, substantial discrepancies exist between restricted and unrestricted ML estimates for elderly individuals with poor cognitive abilities. Conversely, the WALS estimates closely resemble unrestricted ML estimates. Moreover, for individuals with poor cognitive functioning across countries, WALS confidence intervals are wider than those for individuals with better performance.

6 Conclusion

In summary, our study yields a few insights. First, cognitive impairments negatively influence old people's response probabilities in the interview. Second, estimations of predicted response probabilities by using the WALS approach are closely similar to the unrestricted model. Using the WALS method allows us to construct better propensity score weights because it accounts for model uncertainty.

There is always more to explore. In the future, we could consider using the approach called double-robust estimators. The assumption of a correctly specified model for the response probabilities can be relaxed with the help of doubly robust weighting procedures that

incorporate the specification of two models, one for response probabilities and one for the conditional distribution of the study variables given the auxiliary variables.

7 Appendix

7.1 Appendix A: Tables

Table 4.3: Marginal effects of key socio-demographic variables and cognitive ability variables on the participation probability in CZ

Variable	Model			
	MLR	MLU	WALS-L	WALS-H
Female	0.002	-0.015	-0.010	-0.008
	(0.038)	(0.038)	(0.038)	(0.038)
Age	0.001	0.002	0.001	0.002
	(0.003)	(0.003)	(0.003)	(0.003)
High education	0.095 *	0.100 **	0.101 **	0.101 **
	(0.038)	(0.037)	(0.037)	(0.037)
Couple	0.001	-0.000	0.002	0.003
	(0.038)	(0.038)	(0.038)	(0.038)
Recall: moderate	0.060	0.092	0.084	0.084
	(0.050)	(0.075)	(0.063)	(0.061)
Recall: healthy	-0.010	0.027	0.019	0.019
	(0.055)	(0.077)	(0.067)	(0.064)
Good orientation in time	0.150 **	0.205 ***	0.193 ***	0.191 ***
	(0.052)	(0.051)	(0.052)	(0.052)
Good in first numeracy test	-0.012	-0.018	-0.020	-0.021
	(0.053)	(0.058)	(0.056)	(0.055)
Good in second numeracy test	0.077	-0.019	-0.010	-0.009
	(0.063)	(0.044)	(0.051)	(0.053)
Good in memory test	0.036	0.041	0.040	0.044
	(0.044)	(0.047)	(0.047)	(0.046)
Score in fluency test	0.001	-0.000	-0.000	-0.000
	(0.003)	(0.003)	(0.003)	(0.003)

Note: ML_R: Restricted Maximum Likelihood model; ML_U: Unrestricted Maximum Likelihood model; WALS-L: Weighted-Average Least Squares with Laplace prior; WALS-H: Weighted-Average Least Squares with horseshoe prior.

Table 4.4: Marginal effects of key socio-demographic variables and cognitive ability variables on the participation probability in DE

	Model			
Variable	MLR	MLU	WALS-L	WALS-H
Female	0.011	0.034	0.025	0.022
	(0.034)	(0.035)	(0.035)	(0.035)
Age	-0.004	-0.004	-0.004	-0.004
	(0.003)	(0.003)	(0.003)	(0.003)
High education	0.075	0.047	0.055	0.059
	(0.056)	(0.060)	(0.059)	(0.059)
Couple	-0.060	-0.052	-0.056	-0.058
	(0.035)	(0.035)	(0.036)	(0.036)
Recall: moderate	-0.017	-0.061	-0.051	-0.048
	(0.049)	(0.074)	(0.063)	(0.060)
Recall: healthy	-0.072	-0.092	-0.093	-0.097
	(0.049)	(0.069)	(0.060)	(0.057)
Good orientation in time	-0.059	-0.035	-0.043	-0.050
	(0.058)	(0.048)	(0.056)	(0.059)
Good in first numeracy test	0.079	0.067	0.070	0.067
	(0.055)	(0.086)	(0.073)	(0.069)
Good in second numeracy test	-0.111	-0.101	-0.113	-0.116
	(0.064)	(0.068)	(0.067)	(0.069)
Good in memory test	0.039	0.029	0.033	0.035
	(0.037)	(0.038)	(0.038)	(0.038)
Score in fluency test	0.002	0.001	0.001	0.001
	(0.003)	(0.003)	(0.003)	(0.003)

Note: ML_R: Restricted Maximum Likelihood model; ML_U: Unrestricted Maximum Likelihood model; WALS-L: Weighted-Average Least Squares with Laplace prior; WALS-H: Weighted-Average Least Squares with horseshoe prior.

Table 4.5: Marginal effects of key socio-demographic variables and cognitive ability variables on the participation probability in DK

Variable	Model			
	MLR	MLU	WALS-L	WALS-H
Female	-0.047	-0.045	-0.046	-0.046
	(0.033)	(0.033)	(0.034)	(0.034)
Age	0.001	0.001	0.001	0.001
	(0.003)	(0.003)	(0.003)	(0.003)
High education	-0.096 *	-0.046	-0.063	-0.069
	(0.040)	(0.052)	(0.048)	(0.047)
Couple	0.035	0.032	0.035	0.036
	(0.035)	(0.035)	(0.036)	(0.036)
Recall: moderate	0.068	0.020	0.025	0.025
	(0.063)	(0.086)	(0.079)	(0.077)
Recall: healthy	0.103	0.058	0.065	0.064
	(0.057)	(0.073)	(0.070)	(0.068)
Good orientation in time	0.048	0.055	0.053	0.049
	(0.054)	(0.059)	(0.058)	(0.057)
Good in first numeracy test	0.095	0.046	0.056	0.054
	(0.058)	(0.064)	(0.062)	(0.062)
Good in second numeracy test	0.149 *	0.097	0.104	0.106
	(0.073)	(0.085)	(0.083)	(0.080)
Good in memory test	0.018	-0.003	0.004	0.005
	(0.043)	(0.047)	(0.046)	(0.046)
Score in fluency test	0.002	0.001	0.001	0.002
	(0.002)	(0.003)	(0.003)	(0.003)

Note: ML_R: Restricted Maximum Likelihood model; ML_U: Unrestricted Maximum Likelihood model; WALS-L: Weighted-Average Least Squares with Laplace prior; WALS-H: Weighted-Average Least Squares with horseshoe prior.

Table 4.6: Marginal effects of key socio-demographic variables and cognitive ability variables on the participation probability in FR

Variable	Model			
	MLR	MLU	WALS-L	WALS-H
Female	-0.020	0.007	-0.003	-0.006
	(0.034)	(0.034)	(0.034)	(0.034)
Age	-0.002	-0.002	-0.002	-0.002
	(0.002)	(0.003)	(0.002)	(0.002)
High education	0.061	0.077 *	0.074	0.074
	(0.038)	(0.038)	(0.038)	(0.038)
Couple	-0.078 *	-0.099 **	-0.095 **	-0.094 **
	(0.036)	(0.036)	(0.036)	(0.036)
Recall: moderate	0.002	-0.019	-0.018	-0.019
	(0.057)	(0.089)	(0.075)	(0.072)
Recall: healthy	0.059	0.019	0.023	0.021
	(0.059)	(0.087)	(0.075)	(0.072)
Good orientation in time	0.195 *	0.181 **	0.180 *	0.183 *
	(0.081)	(0.066)	(0.070)	(0.072)
Good in first numeracy test	-0.127 ***	-0.115 **	-0.115 **	-0.112 **
	(0.036)	(0.039)	(0.038)	(0.038)
Good in second numeracy test	0.102 *	0.106	0.100	0.098
	(0.052)	(0.065)	(0.058)	(0.057)
Good in memory test	0.039	0.022	0.028	0.029
	(0.035)	(0.035)	(0.035)	(0.035)
Score in fluency test	-0.003	0.001	-0.000	-0.001
	(0.002)	(0.003)	(0.003)	(0.003)

Note: ML_R: Restricted Maximum Likelihood model; ML_U: Unrestricted Maximum Likelihood model; WALS-L: Weighted-Average Least Squares with Laplace prior; WALS-H: Weighted-Average Least Squares with horseshoe prior.

Table 4.7: Marginal effects of key socio-demographic variables and cognitive ability variables on the participation probability in IT

Variable	Model			
	MLR	MLU	WALS-L	WALS-H
Female	-0.002	0.024	0.018	0.017
	(0.035)	(0.037)	(0.036)	(0.035)
Age	-0.006 *	-0.006 *	-0.006 *	-0.006 *
	(0.002)	(0.003)	(0.002)	(0.002)
High education	0.035	0.036	0.037	0.038
	(0.040)	(0.049)	(0.044)	(0.043)
Couple	0.023	0.018	0.020	0.020
	(0.034)	(0.034)	(0.033)	(0.033)
Recall: moderate	0.082	0.077	0.076	0.074
	(0.043)	(0.059)	(0.051)	(0.050)
Recall: healthy	-0.181 **	-0.182 **	-0.184 **	-0.186 **
	(0.057)	(0.070)	(0.063)	(0.063)
Good orientation in time	0.040	0.064	0.059	0.059
	(0.039)	(0.042)	(0.041)	(0.040)
Good in first numeracy test	0.015	0.060	0.046	0.042
	(0.036)	(0.038)	(0.037)	(0.037)
Good in second numeracy test	0.018	0.000	0.006	0.007
	(0.040)	(0.042)	(0.041)	(0.041)
Good in memory test	0.044	0.032	0.036	0.037
	(0.036)	(0.037)	(0.036)	(0.036)
Score in fluency test	-0.000	-0.002	-0.002	-0.002
	(0.003)	(0.004)	(0.003)	(0.003)

Note: ML_R: Restricted Maximum Likelihood model; ML_U: Unrestricted Maximum Likelihood model; WALS-L: Weighted-Average Least Squares with Laplace prior; WALS-H: Weighted-Average Least Squares with horseshoe prior.

7.2 Appendix B: Figures

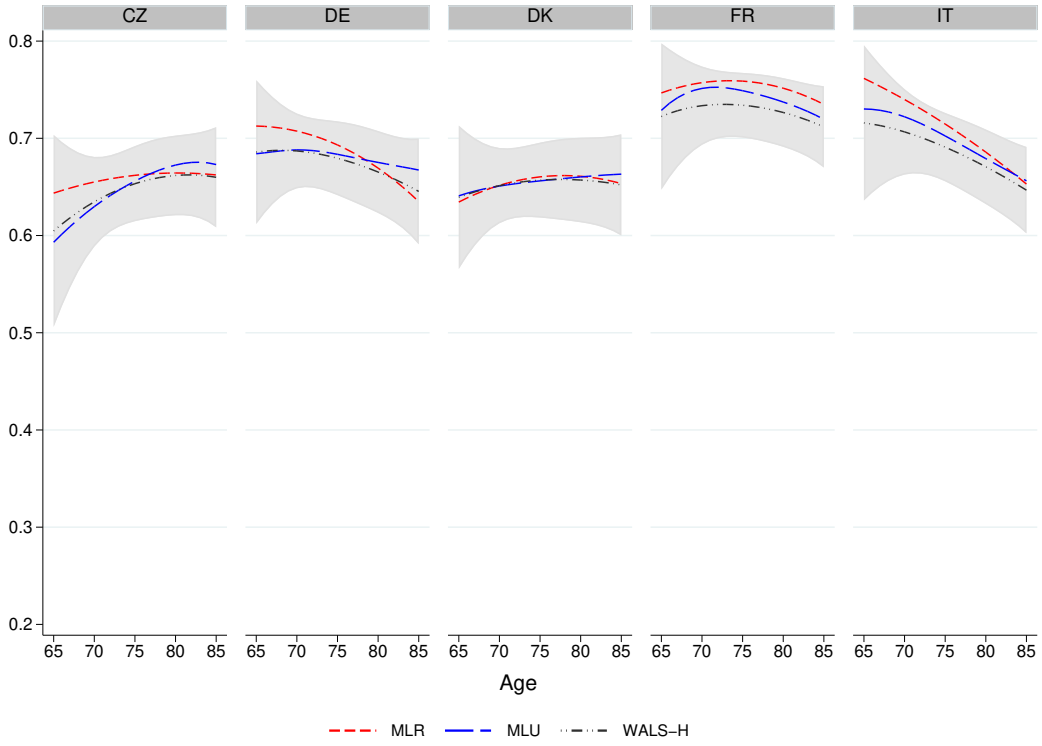

Figure 4.1: Predicted response probability across different ages and selected countries

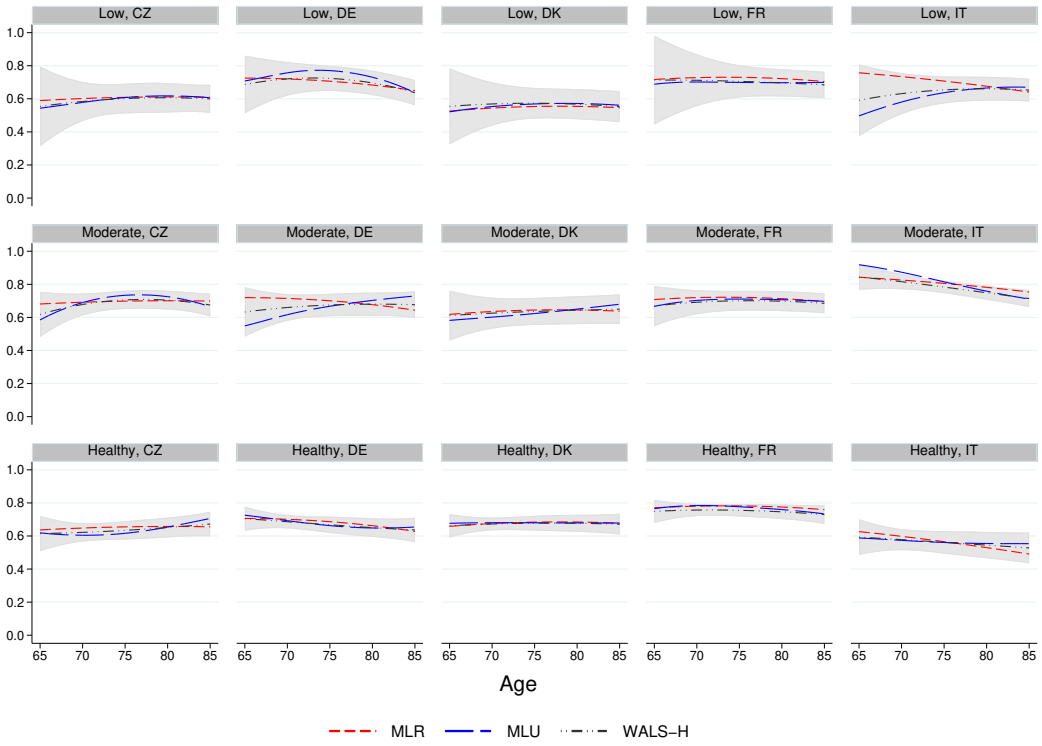

Figure 4.2: Predicted response probability across different ages and selected countries in three different word recall scores groups

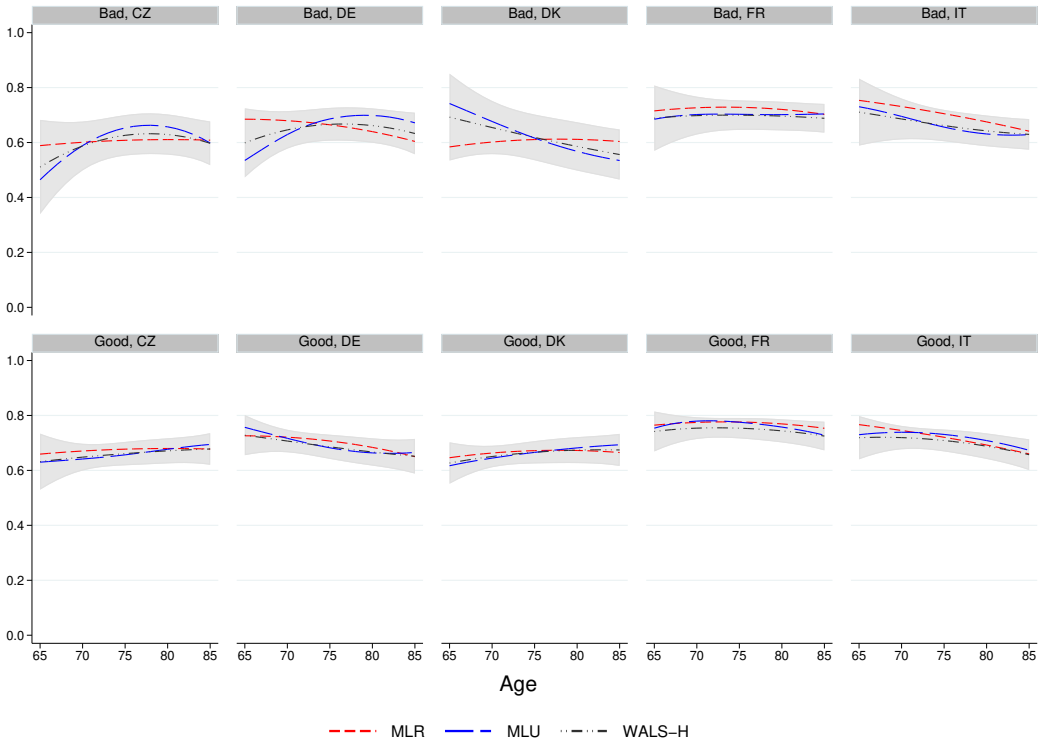

Figure 4.3: Predicted response probability across different ages and selected countries in two different memory test scores groups

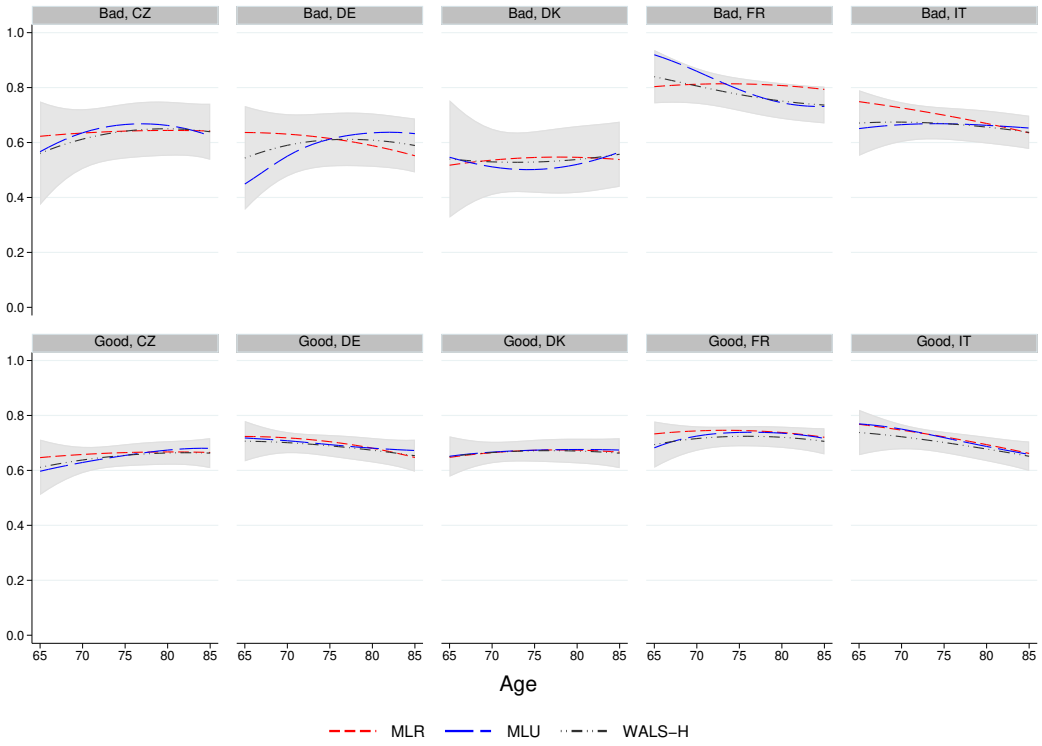

Figure 4.4: Predicted response probability across different ages and selected countries in two different numeracy 1 test scores groups

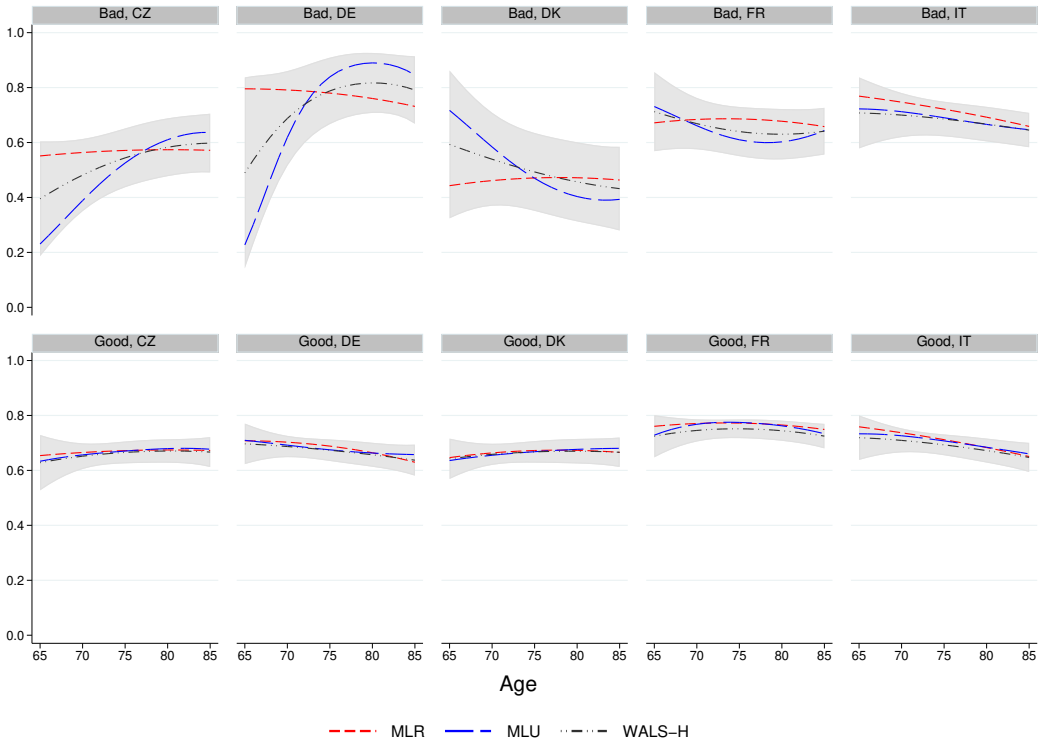

Figure 4.5: Predicted response probability across different ages and selected countries in two different numeracy 2 test scores groups

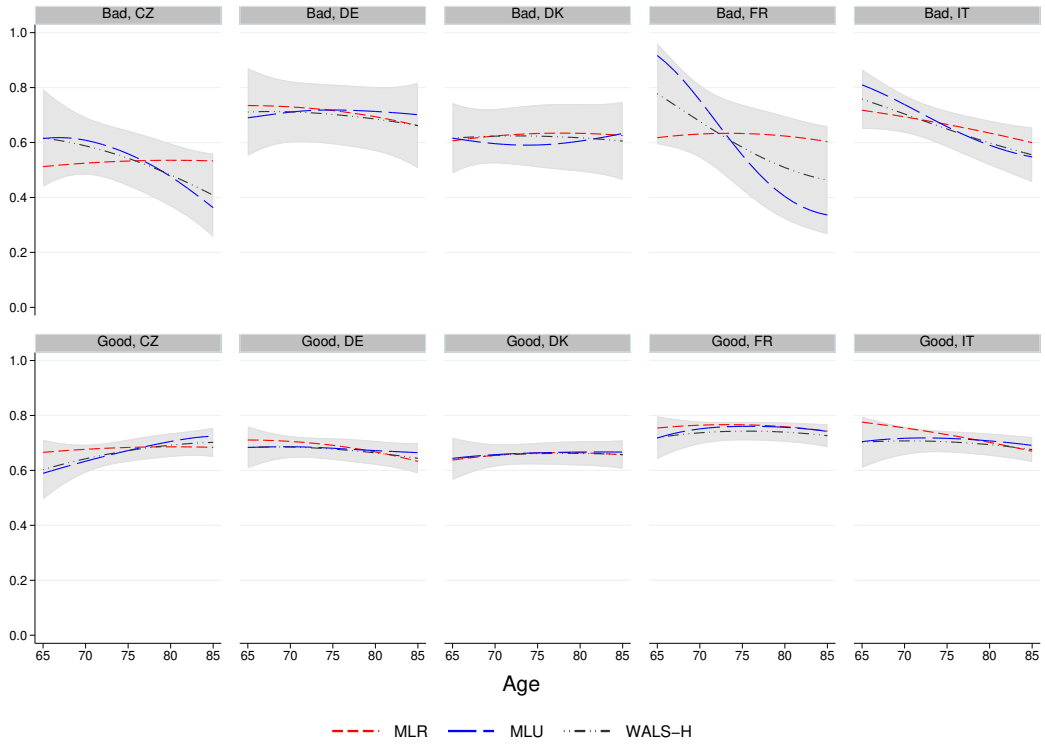

Figure 4.6: Predicted response probability across different ages and selected countries in two different orientations in time scores groups

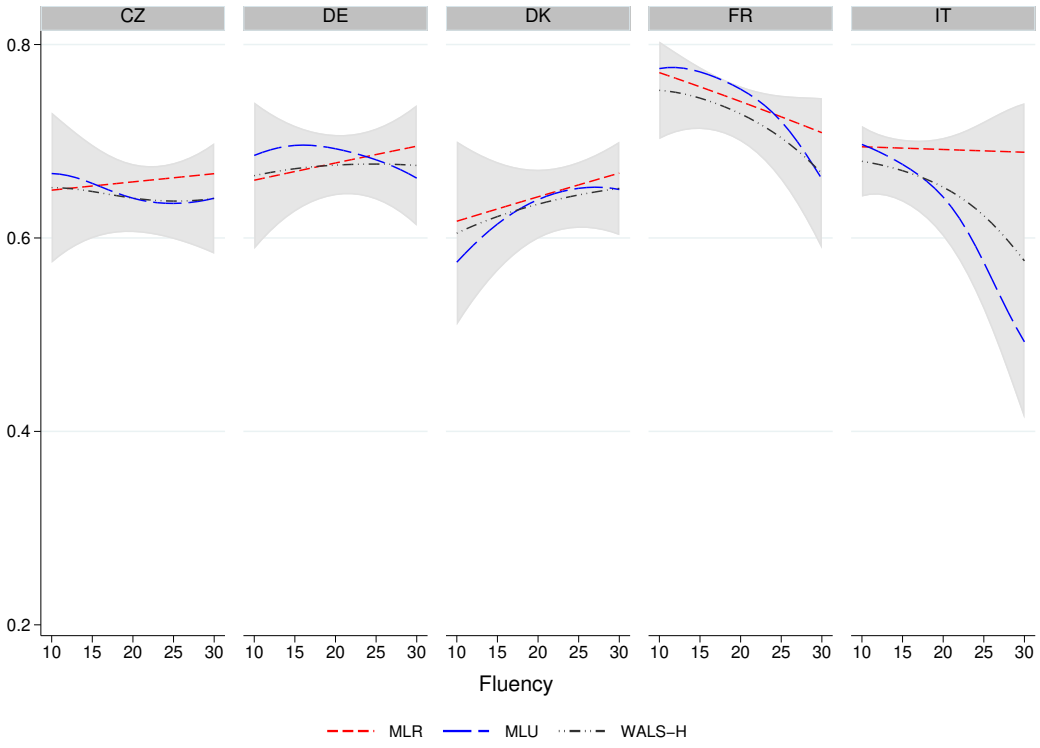

Figure 4.7: Predicted response probability across different fluency scores and selected countries

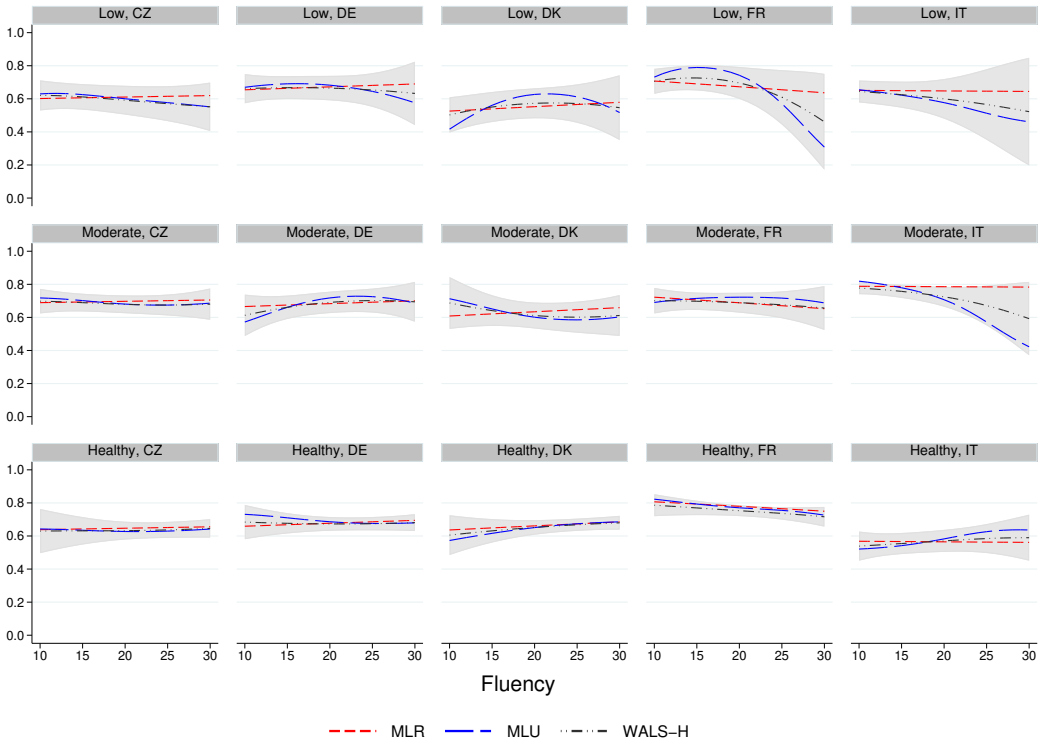

Figure 4.8: Predicted response probability across different fluency scores and selected countries in three different word recall scores groups

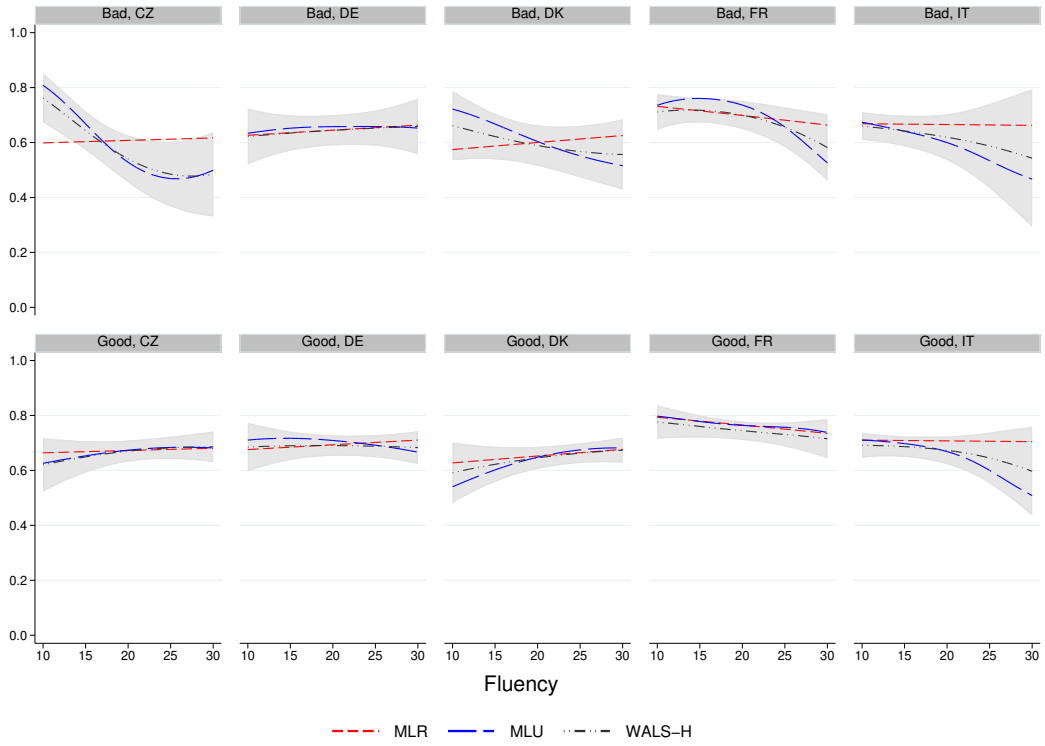

Figure 4.9: Predicted response probability across different fluency scores and selected countries in two different memory scores groups

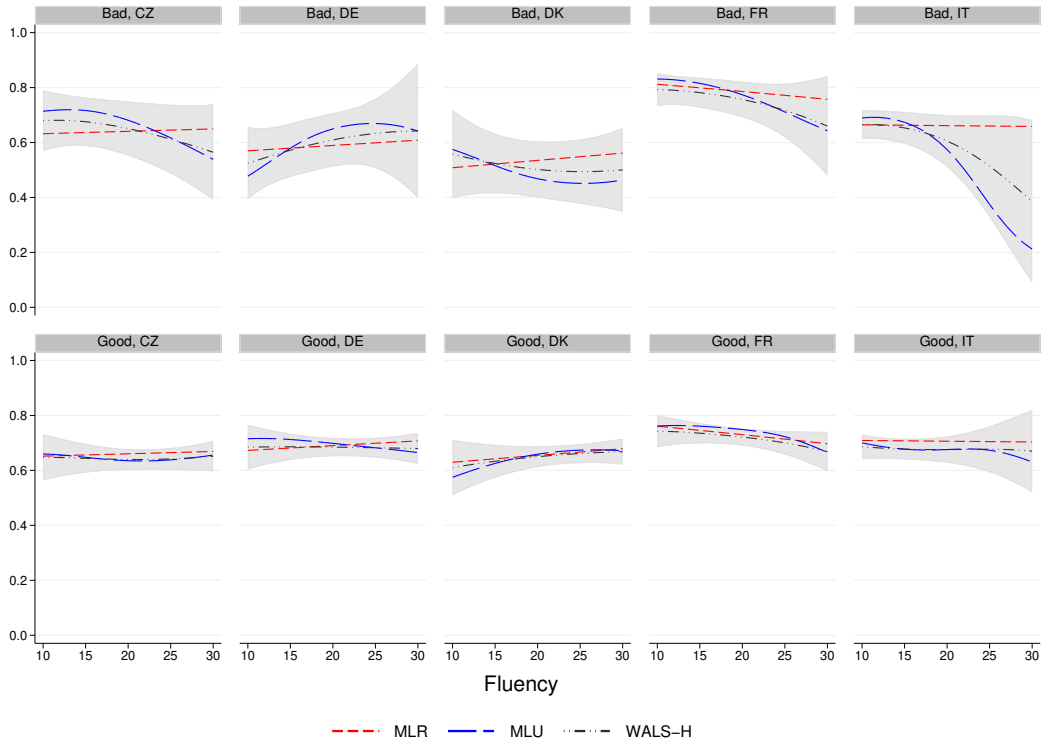

Figure 4.10: Predicted response probability across different fluency scores and selected countries in two different numeracy 1 scores groups

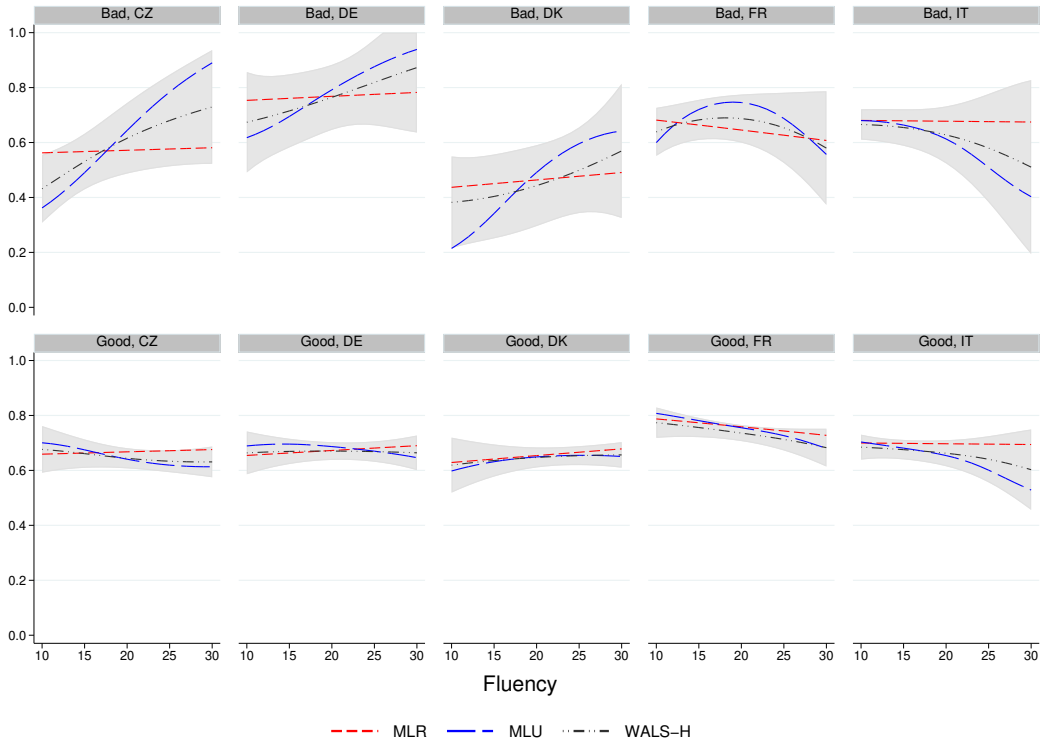

Figure 4.11: Predicted response probability across different fluency scores and selected countries in two different numeracy 2 scores groups

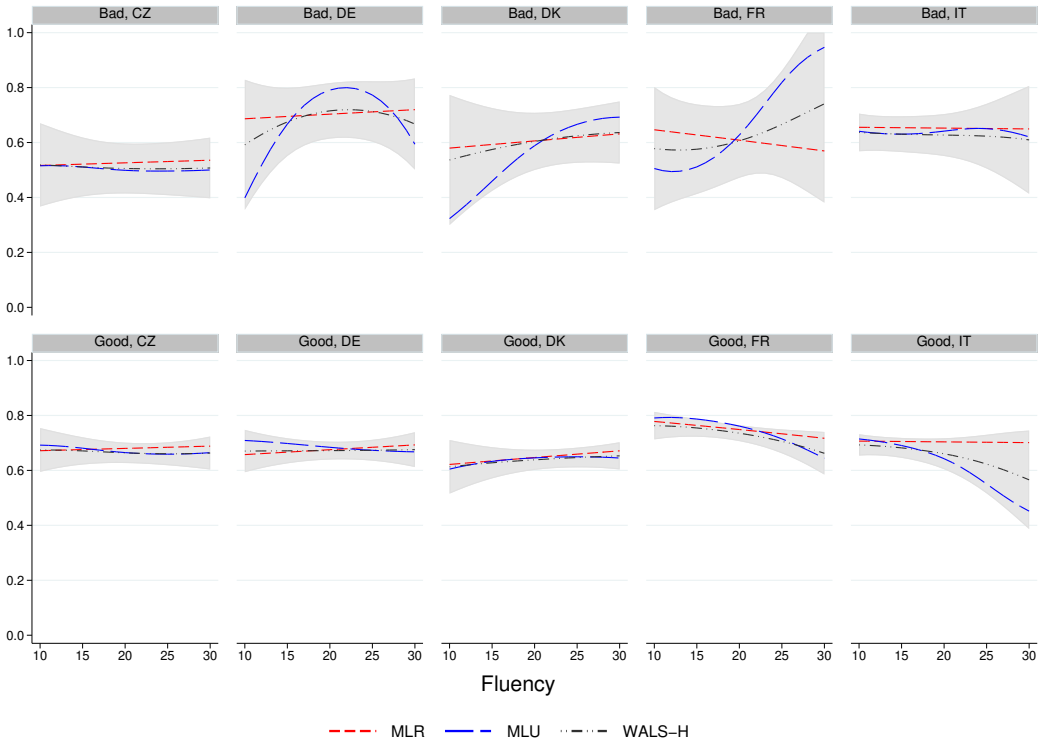

Figure 4.12: Predicted response probability across different fluency scores and selected countries in two different orientations in time scores groups

References

Bergmann and Bethmann (2022).SHARE HCAP technical report - sampling.

Brick, J.M. (2013). Unit nonresponse and weighting adjustments: A critical review. *Journal of Official Statistics*, 29, 329–353.

Buskirk, T.D. and Kolenikov, S. (2015). Finding respondents in the forest: A comparison of logistic regression and random forest models for response propensity weighting and stratification. *Survey Methods: Insights from the Field*, 1–17.

Carvalho, C.M., Polson, N.G. and Scott, J.G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, 97, 465–480.

Cham, H. and West, S.G. (2016). Propensity score analysis with missing data. *Psychological methods*, 21, 427–445.

Danilov, D. and Magnus, J.R. (2004). On the harm that ignoring pretesting can cause. *Journal of Econometrics*, 122, 27–46.

Danilov, D., 2005. Estimation of the mean of a univariate normal distribution when the variance is not known. *The Econometrics Journal*, 8, 277–291.

De Luca, G., Magnus, J.R. and Peracchi, F. (2018). Weighted-average least squares estimation of generalized linear models. *Journal of econometrics*, 204, 1–17.

De Luca, G., Magnus, J.R. and Peracchi, F. (2021). Weighted-average least squares (WALS): Confidence and prediction intervals. *Computational Economics*, 61, 1637–1664.

De Luca, G., Magnus, J.R., and Peracchi, F. (2023). Asymptotic properties of the weighted-average least squares estimator. (Forthcoming)

Earp, M., Mitchell, M., McCarthy, J. and Kreuter, F. (2014). Modeling nonresponse in establishment surveys: Using an ensemble tree model to create nonresponse propensity scores and detect potential bias in an agricultural survey. *Journal of Official Statistics*, 30, 701–719.

Earp, M., Toth, D., Phipps, P. and Oslund, C. (2018). Assessing nonresponse in a longitudinal establishment survey using regression trees. *Journal of Official Statistics*, 34, 463–481.

Ferri-García, R. and Rueda, M.D.M. (2020). Propensity score adjustment using machine learning classification algorithms to control selection bias in online surveys. *PloS one*, 15, p.e0231500.

Frazier, D.T. and Renault, E. (2017). Efficient two-step estimation via targeting. *Journal of Econometrics*, 201, 212–227.

Griffin, B.A., McCaffrey, D.F., Almirall, D., Burgette, L.F. and Setodji, C.M. (2017). Chasing balance and other recommendations for improving nonparametric propensity score models. *Journal of causal inference*, 5.

Groves, R.M. (2006). Nonresponse rates and nonresponse bias in household surveys. *International Journal of Public Opinion Quarterly*, 70, 646–675.

Gupta, A. (2023). Efficient closed-form estimation of large spatial autoregressions. *Journal of Econometrics*, 232, 148–167.

Hjort, N.L. and Claeskens, G. (2003). Frequentist model average estimators. *Journal of the American Statistical Association*, 98, 879–899.

Hoeting, J.A., Madigan, D., Raftery, A.E. and Volinsky, C.T. (1999). Bayesian model averaging: a tutorial *Statistical science*, 14, 382–417.

Janssen, P., Jureckova, J. and Veraverbeke, N. (1985). Rate of convergence of one-and two-step M-estimators with applications to maximum likelihood and Pitman estimators. *The Annals of Statistics*, 1222–1229.

Leamer, E. E. (1978). Specification searches: Ad hoc inference with nonexperimental data. *New York: Wiley.*

Little, R.J. and Rubin, D.B. (2019). Statistical analysis with missing data. John Wiley & Sons.

Magnus, J.R. and Durbin, J. (1999). Estimation of regression coefficients of interest when other regression coefficients are of no interest. *Econometrica*, 67, 639–643.

Magnus, J.R. and De Luca, G. (2016). Weighted-average least squares (WALS): a survey. *Journal of Economic Surveys*, 30, 117–148.

Magnus, J.R., Powell, O. and Prüfer, P. (2010). A comparison of two model averaging techniques with an application to growth empirics. *Journal of econometrics*, 154, 139–153.

McCullagh, P. and Nelder, J.A. (1989). Binary data. In Generalized linear models. *Springer US*, 98–148.

Moral-Benito, E. (2015). Model averaging in economics: An overview. *Journal of Economic Surveys*, 29, 46–75.

Rosenbaum, P.R. and Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 41–55.

Rothenberg, T.J. (1984). Approximate normality of generalized least squares estimates. *Econometrica: Journal of the Econometric Society*, 811–825.

Salthouse, T.A., 2014. Selectivity of attrition in longitudinal studies of cognitive functioning. *Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, 69, 567–574.

Silva, D.D. and Opsomer, J.D. (2006). A kernel smoothing method of adjusting for unit nonresponse in sample surveys. *Canadian Journal of Statistics*, 34, 563–579.

Steel, M.F. (2020). Model averaging and its use in economics. *Journal of Economic Literature*, 58, 644–719.

Van Beijsterveldt, C.E.M., Van Boxtel, M.P.J., Bosma, H., Houx, P.J., Buntinx, F.J.V.M. and Jolles, J. (2002). Predictors of attrition in a longitudinal cognitive aging study: The Maastricht Aging Study (MAAS). *Journal of clinical epidemiology*, 55, 216–223.