PREVALENCE, IMPACT, AND ADJUSTMENTS OF MEASUREMENT ERROR IN RETROSPECTIVE REPORTS OF UNEMPLOYMENT: AN ANALYSIS USING SWEDISH ADMINISTRATIVE DATA

A thesis submitted to the University of Manchester for the degree of PhD in Social Statistics in the Faculty of Humanities

2014

JOSE PINA-SÁNCHEZ

SCHOOL OF SOCIAL SCIENCES
# List of Contents

Abstract .......................................................................................................................... 12

Declaration .................................................................................................................... 13

Copyright ...................................................................................................................... 14

Acknowledgement ....................................................................................................... 15

Abbreviations ................................................................................................................ 16

CHAPTER 1. INTRODUCTION AND THEORETICAL FRAMEWORK ................. 17

1.1. Measurement Error .............................................................................................. 21

1.1.1. Classical Test Theory ....................................................................................... 25

1.1.2. The Implications of Classical Measurement Error ............................................ 29

1.2. Event History Analysis ....................................................................................... 36

1.2.1. Parametric Models ........................................................................................ 41

1.2.2. Semi-Parametric Models ............................................................................... 42

1.2.3. Non-Parametric Models ................................................................................. 43

1.2.4. Single and Repeated Events ......................................................................... 44

1.2.5. Event History Analysis Data .......................................................................... 47

CHAPTER 2. DATA ................................................................................................. 49

2.1. Survey Data from the LSA ................................................................................. 49

2.2. Register Data from PRESO ............................................................................... 51

2.3. Matched Samples ............................................................................................... 54

CHAPTER 3. PREVALENCE AND MECHANISMS OF MEASUREMENT ERROR IN RETROSPECTIVE REPORTS OF UNEMPLOYMENT ............... 62

3.1. Types of Measurement Error in Retrospective Reports of Unemployment ...... 63

3.2. Measurement Error Mechanisms in Retrospective Reports of Unemployment . 69

3.3. Literature Review .............................................................................................. 72

3.4. Measurement Error at the Spell Level ................................................................ 77
3.5. Measurement Error at the Work History Level .......................................................... 82
  3.5.1. Miscounds of the Number of Spells of Unemployment ...................................... 82
  3.5.2. Misreports of Total Durations in Unemployment ............................................. 85
  3.5.3. Mismatches of Work Status in Person-Day Observations .................................. 88
3.6. Discussion of the Measurement Error Mechanisms ............................................. 95

CHAPTER 4. IMPLICATIONS OF MEASUREMENT ERROR IN EVENT HISTORY ANALYSIS ...................................................................................................................... 99

  4.1. Literature Review ................................................................................................. 100
  4.2. Impact in Single Spell Models ........................................................................... 104
    4.2.1. Tools for the Assessment of the Impact of Measurement Error in Event History Analysis Models ................................................................. 108
    4.2.2. Impact on the Accelerated Life-Time Weibull and Exponential Models ....... 110
    4.2.3. Impact on the Proportional Hazards Cox Models ........................................ 113
    4.2.4. Impact on the Proportional Odds Logit Model ............................................. 115
    4.2.5. Summary of the Impact on Models for Single Spells ................................... 117
  4.3. Impact in Multiple Spells Models ......................................................................... 119
    4.3.1. Impact on the Accelerated Life-Time Weibull and Exponential Models ....... 123
    4.3.2. Impact on the Proportional Hazards Cox Model ........................................... 125
    4.3.3. Impact on the Proportional Odds Logit Model ............................................. 126
    4.3.4. Impact on the Random Intercepts Proportional Odds Logit Model ............... 127
    4.3.5. Summary of the Impact on Models for Repeated Spells ............................... 129
  4.4. Discussion of the Implications of Measurement Error in Event History Analysis ................................................................................................................................. 130

CHAPTER 5. ADJUSTMENT OF EVENT-OCCURRENCE MEASUREMENT ERROR .............................................................................................................................. 135

  5.1. Methods for the Adjustments of Measurement Error ......................................... 136
  5.2. SIMEX .................................................................................................................. 138
5.2.1. SIMEX for Measurement Error in the Explanatory Variable of a Simple Regression Model ................................................................. 139
5.2.2. Implementation of SIMEX ............................................................. 144
5.2.3. Conclusions from SIMEX .............................................................. 154
5.3. Regression Calibration and Multiple Imputation ........................................ 155
  5.3.1. Implementation of Regression Calibration and Multiple Imputation ...... 159
  5.3.2. Conclusion from Regression Calibration and Multiple Imputation ........ 174
5.4. Bayesian Adjustments ........................................................................ 176
  5.4.1. Implementation of the Bayesian Adjustments ....................................... 179
  5.4.2. Conclusions from the Bayesian Adjustments ........................................ 198

CHAPTER 6. CONCLUSION ..................................................................... 200
6.1. Findings ......................................................................................... 202
  6.1.1. Large Prevalence and Impact ......................................................... 202
  6.1.2. Different Questions Need to Be Used ............................................ 203
  6.1.3. The Data Needs to Be Treated Differently ....................................... 205
6.2. Generalisations and Caveats .............................................................. 207
  6.2.1. Caveats ...................................................................................... 208
6.3. Further Analyses ............................................................................. 210
  6.3.1. The Study of Measurement Error in Register Data ......................... 210
  6.3.2. Additional Adjustments ................................................................. 211

References ............................................................................................. 213
Appendix A: Descriptive Statistics for the Samples Used ......................... 224
Appendix B: Normality of the Transformed Durations of Unemployment ...... 226
Appendix C: Person-Day Mismatches ....................................................... 227
Appendix D: Bootstrap ........................................................................... 228
Appendix E: R Code for the SIMEX Process ........................................... 229
Appendix F: Extrapolation Functions ................................................................. 231
Appendix G: Visual Test of Convergence ........................................................... 233
Appendix H: WinBUGS Code for the Bayesian Adjustments .............................. 235
Appendix I: DAGs for Future Adjustments ....................................................... 238
List of Tables

Table 1. Status of the first spell reported in LSA and PRESO* ........................................57
Table 2. Status of person-day cases in LSA and PRESO* ..................................................59
Table 3. Estimates (standard errors) for the logit models for misdating and misclassification of the first spells reported*,** .................................................................81
Table 4. Estimates (standard errors) for the logit model for omission* ..........................84
Table 5. Estimates (standard errors) for misreported total durations in unemployment 87
........................................................................................................................................
Table 6. Crosstabulation of person-day cases of PRESO and LSA-1993 for the PRESO categories without job and replacement scheme. ...........................................89
Table 7. Crosstabulation of person-day cases of PRESO and LSA-1993 cases for the PRESO categories permanent job and part-time employed.................................90
Table 8. Classification of spells of unemployment in January 1992* .........................90
Table 9. Classification of spells of unemployment for the day of the interview* .......91
Table 10. Estimates (standard errors) from the random effects logit model for mismatch* .........................................................................................................................94
Table 11. Single event AL Weibull model using register and survey data* ............110
Table 12. Bias in the single event AL Weibull model.......................................................111
Table 13. Single event AL Exponential model using register and survey data ......113
Table 14. Bias in the single event AL Exponential model.............................................113
Table 15. Single event PH Cox model using register and survey data* .................114
Table 16. Bias in the single event PH Cox model .........................................................114
Table 17. Single event PO logit model using register and survey data .................116
Table 18. Bias in the single event PO logit model model............................................116
Table 19. EHA models’ performance in the presence of event-occurrence ME........117
Table 20. EHA models’ performance compared to the PH Cox for register data....119
Table 21. Repeated events AL Weibull model using register and survey data .......124
Table 22. Bias in the repeated events AL Weibull model .............................................124
Table 23. Repeated events AL Exponential model using register and survey data .125
Table 24. Bias in the repeated events AL Exponential model ..............................125
Table 25. Repeated events PH Cox model using register and survey data ...........126
Table 26. Bias in the repeated events PH Cox model ......................................126
Table 27. Repeated events PO logit model using register and survey data ..........127
Table 28. Bias in the repeated events PO logit model ......................................127
Table 29. Random intercepts PO logit model using register and survey data ......128
Table 30. Bias in the random intercepts PO logit model ..................................129
Table 31. EHA models’ performance in the presence of event-occurrence ME ...130
Table 32. EHA models’ performance compared to the PH Cox ..........................130
Table 33. Regression estimates using the standard SIMEX adjustment* ..........150
Table 34. Regression estimates and standard errors using the classical multiplicative SIMEX adjustment ................................................................. 154
Table 35. Effectiveness in R.R.BIAS and R.RMSE with 25% validation subsample ............................................................... 167
Table 36. Effectiveness in R.R.BIAS and R.RMSE with 25% validation subsample ............................................................... 169
Table 37. Effectiveness in terms of R.R.BIAS for different validation subsamples and functional forms ............................................................ 171
Table 38. Bayesian true and naïve models ..........................................................182
Table 39. Results for the adjustment assuming classical ME ..............................184
Table 40. Means (and medians for durations) of variables included in the two samples ...................................................................................... 185
Table 41. Results for the adjustment assuming MAR .........................................188
Table 42. Means (and medians for durations) of variables included in the two samples ...................................................................................... 189
Table 43. Adjustment using a validation subsample and a mixture model ............192
Table 44. Means (and medians for durations) of variables included in the second subsample

Table 45. Adjustment using a mixture model and a different validation subsample

Table 46. Adjustment using mixture models with strong priors
List of Figures

Figure 1. Scatter-plots of a pair of variables with (below) and without ME (above) 30
Figure 2. Demonstration of the effect of classical measurement error on a explanatory variable using scatter plots and lines of best fit. ..............................32
Figure 3. Demonstration of the effect of classical measurement error on a response variable using scatter plots and lines of best fit. .................................35
Figure 4. Example of three work histories durations ..................................45
Figure 5. Work histories affected by ME in a first spells setting .......................66
Figure 6. Work histories affected by ME in a multiple spells setting ..................68
Figure 7. ME-generating mechanisms in retrospective reports of unemployment ....71
Figure 8. Comparing two retrospective designs ........................................74
Figure 9. Problems of ME identification at the spell level .............................78
Figure 10. Frequencies of the starts of spells of unemployment by day of the month: LSA 1993 ..............................................................79
Figure 11. Number of spells of unemployment in LSA and PRESO-1993 (top) and in LSA and PRESO-2001 (bottom) ....................................................83
Figure 12. Scatterplot of the aggregated time spent in unemployment: LSA-1993...86
Figure 13. Scatter plot of the proportion of misclassified time-units and time-span from the interview in LSA-1993 (above) and LSA-2001 (below)* ..................92
Figure 14. Survivor function for the register and survey data ..........................105
Figure 15. Scatterplot (above) and empirical probability density functions (below) of the observed and true durations .........................................................107
Figure 16. Weibull baseline hazard function for the register and survey data ......112
Figure 17. Cox baseline hazard function for the register and survey data ............115
Figure 18. PO baseline hazard function for the register and survey data*,** ..........117
Figure 19. Probability mass functions of the number of spells in the register and the survey ..........................................................120
Figure 20. Probability density function of repeated spells ........................................ 121
Figure 21. Number of subjects unemployed across the window of observation .... 122
Figure 22. Survivor functions restricted to second spells ........................................ 122
Figure 23. Extrapolation function ...................................................................... 142
Figure 24. Comparison of extrapolation functions ............................................... 143
Figure 25. Classical additive simulations for $\lambda_1 = 1$ compared to LSA durations. 147
Figure 26. Observed and true durations compared to the first two sets of classical additive simulations* ............................................................ 148
Figure 27. Extrapolation functions for simulated classical additive ME ............ 150
Figure 28. Bootstrapped sampling distributions for the regression estimates* ...... 151
Figure 29. Classical multiplicative simulations for $\lambda_1 = 1$ compared to LSA durations.................................................................................. 152
Figure 30. LSA and PRESO compared to the first two sets of classical multiplicative simulations* .......................................................... 153
Figure 31. Extrapolation functions for simulated classical multiplicative ME ...... 154
Figure 32. LSA and calibrated durations: logit, linear, exponential, and PMM..... 165
Figure 33. Probability density functions for the different adjusted durations ........ 167
Figure 34. PMM adjusted estimates across validation subsamples of size 25% ..... 170
Figure 35. Adjustment curves for different calibration models ............................. 172
Figure 36. The true AL exponential model............................................................. 181
Figure 37. Adjustment assuming classical ME...................................................... 184
Figure 38. Adjustment assuming MAR ............................................................... 187
Figure 39. Adjustment using a validation subsample and a mixture model........ 191
Figure 40. Effectiveness of the adjustment using a mixture model...................... 193
Figure 41. Empirical probability density functions of the true, observed, and predicted durations................................................................. 194
Figure 42. Posterior distributions of a selection of cases of predicted PRESO...... 194
Figure 43. Effectiveness of the adjustment using a mixture model and an alternative validation subsample ............................................................................................................. 196

Figure 44. Effectiveness of the adjustment using a mixture model with fixed $\pi$ .... 197

Figure A1. Density functions of the difference in the registered and reported time spent in unemployment (left) and the squared root of that difference (right) .......... 226

Figure A2. Extrapolation functions for simulated classical additive ME ............. 231

Figure A3. Extrapolation functions for simulated classical multiplicative ME ...... 232

Figure A4. The naïve AL exponential model ..................................................... 233

Figure A5. The true AL exponential model ..................................................... 233

Figure A6. Adjustment assuming classical ME .............................................. 233

Figure A7. Adjustment assuming MAR and using a validation subsample ......... 234

Figure A8. Adjustment using a mixture model and a validation subsample ....... 234

Figure A9. Adjustment using a mixture model and strong priors .................... 234

Figure A10. Adjustment using a mixture model and paradata ......................... 238

Figure A11. Adjustment using person-day cases and known SN and SP ........... 239
Abstract

University of Manchester
Jose Pina-Sánchez
PhD in Social Statistics
Prevalence, Impact, and Adjustments of Measurement Error in Retrospective Reports of Unemployment: An Analysis Using Swedish Administrative Data
29th July 2014

In this thesis I carry out an encompassing analysis of the problem of measurement error in retrospectively collected work histories using data from the “Longitudinal Study of the Unemployed”. This dataset has the unique feature of linking survey responses to a retrospective question on work status to administrative data from the Swedish Register of Unemployment.

Under the assumption that the register data is a gold standard I explore three research questions: i) what is the prevalence of and the reasons for measurement error in retrospective reports of unemployment; ii) what are the consequences of using such survey data subject to measurement error in event history analysis; and iii) what are the most effective statistical methods to adjust for such measurement error.

Regarding the first question I find substantial measurement error in retrospective reports of unemployment, e.g. only 54% of the subjects studied managed to report the correct number of spells of unemployment experienced in the year prior to the interview. Some reasons behind this problem are clear, e.g. the longer the recall period the higher the prevalence of measurement error. However, some others depend on how measurement error is defined, e.g. women were associated with a higher probability of misclassifying spells of unemployment but not with misdating them.

To answer the second question I compare different event history models using duration data from the survey and the register as their response variable. Here I find that the impact of measurement error is very large, attenuating regression estimates by about 90% of their true value, and this impact is fairly consistent regardless of the type of event history model used.

In the third part of the analysis I implement different adjustment methods and compare their effectiveness. Here I note how standard methods based on strong assumptions such as SIMEX or Regression Calibration are incapable of dealing with the complexity of the measurement process under analysis. More positive results are obtained through the implementation of ad hoc Bayesian adjustments capable of accounting for the different patterns of measurement error using a mixture model.
Declaration

No portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.
Copyright

The author of this thesis (including any appendices and/or schedules to this thesis) owns any copyright in it (the “Copyright”) and he has given The University of Manchester the right to use such Copyright for any administrative, promotional, educational and/or teaching purposes.

Copies of this thesis, either in full or in extracts, may be made only in accordance with the regulations of the John Rylands University Library of Manchester. Details of these regulations may be obtained from the Librarian. This page must form part of any such copies made.

The ownership of any patents, designs, trade marks and any and all other intellectual property rights except for the Copyright (the “Intellectual Property Rights”) and any reproductions of copyright works, for example graphs and tables (“Reproductions”), which may be described in this thesis, may not be owned by the author and may be owned by third parties. Such Intellectual Property Rights and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property Rights and/or Reproductions.

Further information on the conditions under which disclosure, publication and exploitation of this thesis, the Copyright and any Intellectual Property Rights and/or Reproductions described in it may take place is available from the Head of School of (Social Sciences).
Acknowledgement

I am eternally grateful to my supervisors Pr Ian Plewis and Dr Johan Koskinen. Their expertise, availability, and encouragement have offered a perfect guidance in the process of writing this thesis, but above all I am extremely grateful for their endless patience.

I would also like to thank Pr. Stenberg for granting me access to the “Longitudinal Study of the Unemployed”, a truly unique dataset without which this research project would have been un conceivable. Similarly I would like to express my gratitude to the Economic and Social Research Council for funding this PhD.

A special thanks goes to my good friends David Mackie and Will Cook for proof-reading this manuscript. Last but not least, I must mention my life-partner, Maria, for her unconditional love and support through the good and bad times, and my parents, Patricia and Jose, for their constant support and encouragement. Sometimes the easiest relationships are also the easiest to take for granted.
Abbreviations

AL: Accelerated life;
DAG: Directed Acyclic Graph;
EHA: Event history analysis;
FN: False negative;
FP: False positive;
LSA: Survey data from the Longitudinal Study of the Unemployed;
MC: Misclassification;
ME: Measurement error;
MI: Multiple imputation
OLS: Ordinary least squares;
PH: Proportional hazards;
PMM: Predictive mean matching;
PO: Proportional odds;
PRESO: Administrative data from the Swedish Register of Unemployment;
RC: Regression calibration;
SIMEX: Simulation-extrapolation.
CHAPTER 1. INTRODUCTION AND THEORETICAL FRAMEWORK

Retrospective questions are commonly used in survey research to obtain information about life-cycle events. Unlike prospective designs where the respondent is contacted repeatedly across time (waves) and asked about her current state, retrospective questions collect information over a span of time from a single contact with the respondent. This makes them – for the period of enquiry: a) immune to problems of both attrition (subjects dropping out of the study) and inconsistency in the wording of the question or the design of the survey; b) cheaper to administer; and c) more capable of detecting transitions occurring in short periods of time.

However, retrospective questions suffer from one major weakness compared with prospective designs\(^1\), they are especially prone to measurement error (ME). Individuals answering retrospective questions are faced with a higher cognitive challenge than those reporting current status since not only do they need to interpret the question correctly but they also need to recall the events being asked. In addition, the memory failures responsible for the ME found with retrospective questions are often interrelated with the nature of the topic being reported and with the relative difficulty of reporting it (degree of saliency, social desirability, etc.), resulting in complex error-generating mechanisms. For example, the complexity of the MEs found in reports of work status experienced in the past should be expected to be higher than for reports of simpler and more memorable life-course events such as marital status. To make things worse, the complexity of the ME generating processes can be aggravated when a sequential component is included in the question. That is, when subjects are not only requested to report the onset or the duration of an event, or the number of times it occurred, but when they are also asked to order those events chronologically.

These two factors - the magnitude and the complexity of the ME - present in retrospectively reported life-course histories, result in two major problems. First, the use of highly prone to ME life-course histories in regression models will have an impact on the estimated estimates and standard errors. Second, most of the methods

\(^1\) See Solga (2001) for a comparison of data quality derived from prospective and retrospective questions.
that have been used in the literature to adjust for such consequences are based on rather simplistic assumptions about the ME process, which might not be appropriate to tackle the highly complex ME that should be expected from retrospectively collected life-course histories. Unfortunately, however, for studies aiming to analyse transitions between different statuses, retrospective data remains the only source of data available to a majority of researchers.

In this thesis, I analyse the ME found in retrospectively collected work histories - or the sequence of work statuses experienced by subjects across a certain window of observation. To do that, I use a dataset where administrative data from the Swedish register of unemployment has been linked to survey responses to a retrospective question where work history events were asked to be identified and dated. Under the assumption that the register data is a gold standard, I am able to obtain new insights on: a) the nature and extent of ME in retrospective questions; b) the impact that using such prone to ME data could have on various event history analysis (EHA) models; and c) the effectiveness of different statistical adjustments on estimates of interest. Answers to these research questions make a contribution in the fields of survey research, labour market studies, and statistical methodology, which makes this thesis the outcome of an inter-disciplinary research project. Specifically, I intend to achieve three goals: a) to provide evidence to enable survey designers to make more informed decisions regarding the adequacy of retrospective questions to capture work histories accurately; b) to make users of retrospectively collected unemployment data more aware of its shortcomings; and c) to promote, on the basis of evidence, the implementation of methods for the adjustment of the type of ME seen in these kinds of questions.

The thesis is structured as follows. Chapter 1 presents the building blocks that will be used to construct the analyses in the thesis. In Section 1.1 the process of measurement is theoretically discussed. This is followed by a presentation of classical test theory, which will be used to frame the commonly used classical ME model and to describe some key concepts such as those of validity and reliability. The section finishes with a brief description of the implications of using data affected by ME in regression analyses. The second building block, introduced in Section 1.2, is that of EHA - a group of models that are particularly useful for modelling transitions from one state to another across time, which will be used to
study durations of spells of unemployment and the factors that have an effect on them. The three different families of EHA models (parametric, semi-parametric, and non-parametric) will be presented in different subsections, followed by a comparison of models for single and repeated spells. Section 1.2 ends with a review of the different observational schemes used to collect EHA data and how ME can arise in each one of them.

Chapter 2 introduces the last of the building blocks, the Longitudinal Study of the Unemployed, which is the dataset that is used throughout the thesis. The unique feature of this dataset is that it is composed of reports of spells of unemployment matched to administrative records. To understand the characteristics of each of these two distinctive sources of data they will be reviewed in separate subsections.

The core of the thesis is presented in Chapters 3, 4, and 5, in which the prevalence, impact, and adjustments for the ME found in retrospective reports of unemployment are studied. Chapter 3 starts with two theoretical sections. The first describes the forms of ME that we could expect to observe in retrospective questions where entire life-course histories are reported – as opposed to those enquiring about the duration of a specific event; see Section 1.2.5. The second relates to the types of ME mechanisms that have been hypothesised to explain the presence of ME in retrospective reports of unemployment. Section 3.3 reviews the evidence found in the literature. This literature review highlights the importance of the analysis presented here, since it appears that only two previous studies have been able to assess the extent of ME in retrospective reports of unemployment using register data. Section 3.4 explores both the prevalence of ME found in spells of unemployment retrospectively collected and the factors that are associated with this problem. Section 3.5 carries out this same analysis but shifting the interest from specific spells of unemployment to the different forms of ME that can be identified at the work history level. These are: miscounts of the number of spells of unemployment, misreports of total durations of unemployment, and mismatches of work status in person-day observations.

In Chapter 4, I analyse and compare the consequences for the estimated regression estimates and their standard errors of different EHA models when retrospectively reported spells of unemployment are used as the response variable. Section 4.1 reviews the findings from the literature. The analysis in this chapter is carried out in
Sections 4.2 and 4.3, where EHA models for (a) single and (b) multiple spells are considered. Each of these two analyses is presented in different subsections examining the effect of ME in: (i) accelerated life-time (AL) Weibull and exponential models; (ii) proportional hazards (PH) Cox models; and (iii) proportional odds (PO) logit models.

Chapter 5 explores the effectiveness of different statistical methods to adjust for the consequence of using retrospective reports of unemployment in EHA models. The chapter starts with a discussion of the taxonomy of ME adjustment methods. Put simply, the main difference between methods is the extent to which they rely on access to additional data that could inform about the ME process, or on assumptions about the distribution of the ME. Using that classification I examine methods that have been implemented in the literature. Section 5.2 explores the implementation of simulation-extrapolation (SIMEX), an adjustment that requires knowledge about the distribution of the error term and an estimate of its variance. In Section 5.3, I study the implementation of regression calibration (RC) and multiple imputation (MI); two methods that rely on having access to a subsample of observations where the true values of the variable affected by ME are known. Finally, Section 5.4 explores the most flexible of the adjustments, those using a Bayesian approach. Bayesian methods can make use of additional sources of data informing about the ME process when they are available and, given the possibility of using informative priors, different adjustments based entirely on distributional assumptions of the ME term will also be studied.

Chapter 6 concludes with a summary of the results presented in Chapters 3, 4, and 5 (in Section 6.1), and with a discussion of the external and internal validity of these findings (in Section 6.2). In addition, I include a final section, 6.3, where I outline some lines of research that could be explored in the future to improve our understanding of both the presence of ME in retrospectively collected spells of unemployment and of the methods that can be used to adjust for it.
1.1. Measurement Error

The measurement of the object of study is the first stage in the scientific method. To use a metaphor, the measurement process can be understood as the foundations of the building that is the scientific enquiry. Because of the crucial importance of measurement researchers should be concerned about their understanding of the processes that generate their data and be critical about its overall quality. Here I will briefly review the theoretical underpinnings of the measurement process as a means of defining what should be understood by ME. The section then continues with a more formal explanation of ME using some of the principles discussed in classical test theory, and ends with a discussion of the typical consequences that should be expected when data is affected by ME.

A first approach to the theory of measurement shows a subject studied from alternative paradigms. This variety suggests a rich debate that needs to be tackled since, as Hand (1996) argues, different paradigms of measurement can lead to different consequences for inferences.

Hand (1996) distinguishes three main measurement paradigms: representational, operational and classical. The representational is the most dominant, to the extent that the expression “measurement theory” is often used as a shortened description for it. This paradigm stems from the seminal work of Stevens (1946). Stevens understood measurement as the process applied to an entity that has the following three aspects, a characteristic, a scale\(^2\), and a means of assigning a number on the scale to that characteristic. So, in the representational paradigm, objects have attributes, and numbers are assigned to the attributes to describe the relationships between them. In addition, Stevens argued that only those statistics which are invariant to changes between legitimate representations should be used, thus constraining quantitative analysis to those phenomena that can be precisely defined and for which there is a consensus about the way they are measured.

In his conceptualization of the different paradigms, Hand (1996) added the insight that representational measurement theory is about describing real empirical systems. In representational measurement theory we begin with a set of objects, each of which

\(^2\) We also owe to Stevens the now famous nominal, ordinal, interval and ratio scales.
has one or more common attributes, which in turn can be divided into mutually exclusive and exhaustive “equivalence classes”. Then the objects and the relationships between them constitute an “empirical relational system”. In parallel with this we construct a numerical relational system comprising numbers and the relationships between them. Then representational measurement theory is concerned with establishing a mapping from the objects, via the equivalence classes to which they belong, to the number system in such a way that the relationships between objects are matched by relationships between numbers.

To this dominant view of measurement, Hand (1996) contrasts the operational paradigm, which defines scientific concepts in terms of the operations used to describe them, and avoids assuming any underlying reality. In operationalism, things start with the measurement procedure and an attribute has no real existence beyond that. In operationalism, the attribute and the variable are one and the same. This approach thus defines a measurement as any precisely specified operation that yields a number.

To these two theories of measurement, Hand opposed a third one, developed by Michell (1986, 1990), and named the classical paradigm in reference to the traces of the works of Aristotle and Euclid noted by Hand. Under the classical paradigm, measurement addresses the question of how much of a particular attribute an object has and thus only refers to attributes which are quantitative. A quantitative attribute is an attribute whose values satisfy ordinal and additive relationships. Michell distinguishes this approach from the representational theory by stressing that it is the attribute which has these properties and not the objects.

From these three paradigms, the representational paradigm is the hegemonic one in most research fields, and is the one that I adopt here. Its conceptualization of measurement as the mapping of numbers from a scale on to an unchanging object or event that exists before it is measured is intuitive, and makes the operationalisations of variables straightforward. Furthermore, from here, ME can be plainly understood as failure in the assignment of numbers to the phenomenon being measured.

However, the representational paradigm doesn’t fit all scientific fields equally well, and some refinements should be incorporated. Zeller and Carmines (1980) indicate that Stevens’ definition of measurement is much more appropriate for the physical
than the social sciences. “The problem with this definition, from the point of view of the social scientist, is that, strictly speaking, many of the phenomena to be measured are neither objects nor events. Rather, the phenomena to be measured are typically too abstract to be adequately characterized as either objects or events”, (Zeller and Carmines, 1980, p. 2).

This problem was first spotted by Blalock (1968); however Zeller and Carmines deserve the credit for deriving an alternative conceptualisation of measurement which, grounded in representational theory, provides a better fit to the complexity of the social sciences. Zeller and Carmines (1980) argue that, because concepts can be neither directly observed nor measured, the systematic exploration, testing, and evaluation of social theory requires social scientists to use “empirical indicants”, designed to represent given abstract concepts. Thus, in contrast to concepts, empirical indicants or indicators are designed to be as specific, as exact, and as bounded as theoretical formulations and research settings will allow. For example whereas it is commonly agreed that the phenomenon of unemployment is experienced by those who do not have a job and are actively looking for one, the indicator (from now on I will refer to them as the true values) that defines unemployment will often come from the government accounts which require one to be registered at a job centre in order to be considered unemployed.

I take this refinement suggested by Carmines and Zeller into account; I recognise that due to the complexity of some of the concepts used in the social sciences they will not be observable. I will go even further and treat indicators not only as the best possible empirical approximations to the concept but also as the true value. Thus, in combination with Steven’s representational theory I set my definition of measurement as the assignment of numbers according to a certain scale and set of rules to the indicator; failures of these attempts will define the MEs. Therefore, I will not think of true values as unobserved or hypothetical as they are often depicted in the literature of ME, but as the values that are obtained when the measurement procedure was perfect. This distinction will be of use when comparing register with survey data, and it will be consistent with the key assumption that I invoke across the analyses presented in this thesis, where I take the register data as a gold standard.

That is, I assume that the register data are error-free and the extent of ME affecting survey responses can be defined as the difference between the register and survey
data. Such an approach has direct implications for what can be learnt about the ME process since it will offer a direct estimate of the level of validity in the survey responses. The concepts of validity and reliability represent the two basic measures that can be used to assess the size and nature of ME. Reliability concerns the extent to which measurements are repeatable; or put it differently, the extent to which identical\textsuperscript{3} measures of the same indicator offer the same result. Validity looks at the extent to which the measuring instrument actually captures what it intends to capture. Thus, a highly reliable measure in itself does not ensure that one has obtained a good measure of an indicator if validity is low. At the same time, if the set of rules configuring the instrument do not aim at the right indicator systematically, a single measure cannot represent what it intends. Therefore, to have a valid measure, one must have a reliable one; but simply because one has a reliable measure does not mean that it is valid as well. Reliability, then, becomes a necessary but not a sufficient condition for validity.

Measurement flaws in terms of reliability and validity give rise to ME. In fact, two types of ME, random error and systematic (or non-random) error, can be respectively linked to them. This distinction between random and non-random errors represents a first approximation that I will discuss regarding the modelling of ME. Random errors are normally the result of those chance factors that confound the measurement of any phenomenon, and as such the amount of random error is inversely related to the degree of reliability of the measuring instrument. All measures will contain random error to a greater or lesser degree. That is, the very process of measurement introduces random error at least to a limited extent. Regarding survey data, Carmines and Zeller (1980) point to errors due to coding, ambiguous instructions, differential emphasis on different words during an interview, or interviewer fatigue, as some of the sources of random ME. Non-random errors are the ones that appear when the measuring instruments have a systematic biasing effect. As such they are inversely related to the degree of validity. A typical example of systematic ME found in survey data comes from responses to socially sensitive questions; that is, questions that are systematically prone to under or over-represent the true value, e.g. inquiries on drug consumption.

\textsuperscript{3} Measures following the same configuration of scales and rules.
1.1.1. Classical Test Theory

Classical test theory (Novick, 1966) represents the first systematic approach to define ME formally. Originally it was created as a framework to analyse the reliability of educational tests. However, it was quickly extended to the main body of statistical theory that has been used to estimate the reliability of empirical measurements.

The basic idea of classical theory was summarized by Novick in the first paragraph of his seminal paper: “By classical test theory we shall mean that theory which postulates the existence of a true score, that error scores are uncorrelated with each other and with true scores, and that observed, true, and error scores are linearly related.” (Novick, 1966, p. 1). I will present these principles by putting forward the simplest of the measurement models that can be specified within this framework, the classical additive model.

ME models are used to specify how a true but unobserved variable, \( X \), is affected by underlying ME mechanisms, \( U \), generating an observed but ME prone variable, \( X^* \). In the case of the classical additive model the observed variable is defined as the sum of the true variable and the ME term,

\[
X^*_i = X_i + U_i
\]

with the subscript \( i \) indicating any particular subject captured in the sample \( s \).

The imposition of an additive relationship between the true variable and the error term is a specific one from the classical additive model, although other forms can be used within the classical framework. In fact, what classifies a ME model as classical is the following set of six assumptions regarding the ME term \( U \):

\[
\text{Classical Model} \nonumber \\
\begin{align*}
E(U) &= 0; & \text{null expectancy} \\
\text{Var}(U) &= \text{Var}(U); & \text{homoscedasticity} \\
U &\sim N(0, \text{Var}(U)); & \text{normally distributed} \\
\text{Cov}(X, U) &= 0; & \text{indep. of error and true value} \\
\text{Cov}(U_i, U_j) &= 0; & \text{indep. of errors between subjects} \\
E(Y|X, X^*) &= E(Y|X); & \text{non – differentiability}
\end{align*}
\]

1. Null expectancy refers to the assumption that the error term is non-systematic, or in other words, the expected value of the error term is zero, \( E(U) = 0 \).
2. The assumption of homoscedasticity indicates that the variance of the error term is assumed to remain constant across subjects, $\text{Var}(U_i) = \text{Var}(U) = \sigma_u^2$.

3. Once established that the expectation of the error term is zero and its variance constant the third assumption indicates that the error term is Normally distributed, $U \sim N(0, \text{Var}(U))$.

4. The fourth assumption specifies that the correlation between the true value and the error term is assumed to be zero, $\text{Cov}(X, U) = 0$.

5. Furthermore, the correlation between different values of the error term is also assumed to be zero, $\text{Cov}(U_i, U_j) = 0$, where $U_i, U_j$ represent any two subjects captured in $s$.

6. The last assumption, non-differentiality, only becomes relevant when $X^*$ is used in a regression model. It indicates that, given the true value, the ME is not associated with the remaining variability in the response, $Y$; that is, $E(Y|X, X^*) = E(Y|X)$.

The second and sixth assumptions were not originally established by Novick (1966), but they have been included here because they are often required in the application of the adjustment methods. Assumptions 1 and 4 (null expectancy and independence of the error and true value) can be used to define the expected value and the variance of the true value as follows,

$$E(X^*) = E(X) + E(U) = E(X)$$

and,

$$\text{Var}(X^*) = \text{Var}(X) + \text{Cov}(X, U) + \text{Var}(U) = \text{Var}(X) + \text{Var}(U)$$

which can in turn be used to define the reliability of an observed variable affected by classical additive ME, $\rho_{X^*}$, as the ratio of the true to observed variance,

$$\rho_{X^*} = \frac{\text{Var}(X)}{\text{Var}(X^*)} = \frac{\text{Var}(X)}{\text{Var}(X) + \text{Var}(U)}$$

In order to calculate the reliability ratio either one of the unobserved variances ($\text{Var}(X)$ or $\text{Var}(U)$) needs to have been previously estimated. This can only be achieved in the presence of additional validation or replicated data. The former requires observing the true values for at least a subsample of the cases under study. Replicated data is available when some or all observations have been repeatedly
measured. In this scenario, under the assumption of parallel measurements (i.e. measurements that have identical true scores and equal variances), the correlation between repeated measures could be used as an estimate of the variance of the true values.

The classical additive model is by far the most commonly used model in the study of ME. It is simple, it applies to many settings where the ME is thought to be mainly random, and it is assumed by most of the available adjustment methods. In addition, the classical ME model can also be used as the foundation upon which more sophisticated models can be built. A straightforward extension of the model presented in equation 1.1 is the classical multiplicative model, which can be used in problems of ME where the size of the ME is expected to be proportional to that of the true values,

\[ X_i^* = X_i \cdot U_i \]  

(1.6)

The same assumptions about the error term described in equation 1.2 apply, with the exception of assumptions 1 and 3. Specifically, in the classical multiplicative model \( U \) is an independent random variable having a log-normal distribution bounded by 0 and \( \infty \) and with mean equal to 1.

The multiplicative relationship between the true values and the error term presented in this model has been exploited by different authors to represent the memory failures that are so ubiquitous in retrospective questions (Holt et al., 1991; Pickles et al., 1996, 1998; Skinner and Humphreys, 1999; Augustin, 1999; Chesher and Schluter, 2002; Glewwe, 2007; and Dumangane, 2007). For example, we could think of the classical multiplicative model as a better approximation than the standard classical additive model when attempting to specify ME processes derived from questions that require remembering the number of times an event took place in the past, such as the number of changes of partner a person has had in the last ten years. For such questions, it could be expected that respondents who only experienced the event a few times will give more accurate reports than those who experienced it more frequently.

This multiplicative model represents one way the classical ME model can be extended. However, due to the simplicity of the classical ME model, many other extensions could be envisaged. For example, if systematic errors are expected to be
part of the ME process we could model them using the following specification suggested by Carroll et al. (2006),

\[ X_i^* = \gamma_0 + \gamma_1 X_i + U_i \]  

(1.7)

Notice that in comparison with the classical additive model, two constants, \( \gamma_0 \) and \( \gamma_1 \), have been included to account for systematic errors that affect the true variable additively and multiplicatively. This model would be equivalent to the classical additive ME model if \( \gamma_0 = 0 \), and \( \gamma_1 = 1 \). Otherwise, two immediate problems should be expected: a biased mean and a much harder to estimate reliability ratio.

Even if \( \gamma_1 = 1 \), the expected value of the observed variable will be defined as follows;

\[ E(X^*) = \gamma_0 + \gamma_1 E(X) \]  

(1.8)

and the reliability ratio will not be able to be estimated only from knowledge of the variance of the true variable, since the variance of the observed variable is not anymore the variance of the true variable plus the variance of the error term,

\[ \rho_{X^*} = \frac{\text{Var}(\gamma_0 + \gamma_1 X)}{\text{Var}(X^*)} = \frac{\gamma_1^2 \text{Var}(X)}{[\gamma_1^2 \text{Var}(X) + \text{Var}(U)]} \]  

(1.9)

One important example where the classical model is entirely inadequate is in those cases where the unit of observation is categorical. Specifically, the ME model where an error term is added to the true value would not make sense because categorical variables lack a scale. Here, the problem of ME becomes one of misclassification (MC), and the ME should be specified as the probability of correctly observing each of the categories of the variable. In the case of binary variables these are referred to as sensitivity and specificity,

\[
\begin{align*}
P(x^* = 1 | x = 1) &= \theta_{1|1} ; \quad \text{True positive (Sensitivity)} \\
P(x^* = 0 | x = 0) &= \theta_{0|0} ; \quad \text{True negative (Specificity)}
\end{align*}
\]  

(1.10)

These two probabilities are complemented by their opposites, respectively the probability of observing false negative and false positives,

\[
\begin{align*}
P(x^* = 0 | x = 1) &= \theta_{0|1} ; \quad \text{False negative} \\
P(x^* = 1 | x = 0) &= \theta_{1|0} ; \quad \text{False positive}
\end{align*}
\]  

(1.11)
Together these four probabilities can be used to specify a MC model for a binary variable. This is summarised through a MC matrix represented by \( \theta \) in the equation below,

\[ \theta_{X^*|X} = P(X^* = x^*|X = x) \]  

(1.12)

The MC model reflecting ME in categorical data will be applied in some of the adjustments carried out in Chapter 5, but for the moment I continue this introductory section with a review of the consequences of using data affected by classical ME in regression models.

1.1.2. The Implications of Classical Measurement Error

Although it is commonly accepted that survey data is affected by ME, the application of statistical methods to adjust for the consequences of this ME is not especially common. This is partly due to a lack of understanding about the consequences derived from using poorly measured data. “Sociologists are well aware of the fact that most of our variables are measured with considerable error. [...] However, actual analyses and interpretations of data usually presuppose that measurement errors can safely be ignored”, (Blalock, 1965, p. 37). To summarize the effects derived from the inclusion of explanatory variables measured with error in a regression analysis Carroll et al. (2006) use the expression “The triple Whammy of Measurement Error”:

First, ME is responsible for masking the features of the data, making graphical analysis difficult. For example, true quadratic or cubic functions might appear blurred and give the wrong impression of being more suitable for a linear fit. This is shown in Figure 1 below, where a cubic relationship between two variables presented in a scatterplot is masked when one of the variables is affected by classical ME. Second, ME leads to a loss of power for detecting relationships among variables. Moreover, if the standard errors for the regression estimates are not adjusted they will be underestimated. Finally, ME causes bias in estimates of regression estimates, with the direction and magnitude of the bias being determined by the type of model where the data is used (i.e. the outcome model), the type of ME process and its prevalence.
Figure 1. Scatter-plots of a pair of variables with (below) and without ME (above)

![Scatter-plots](source)

Source: Carroll et al. (2006, p. 2)

Probably, the last of these implications – the generation of bias in the regression estimates - is the most serious. However, there seems to be a range of misconceptions with respect to the actual consequences of using variables affected by ME. In particular, many researchers tend to assume more predictable and less severe biases than what might be expected.

These misconceptions could be grouped in three categories:

1. Random ME does not bias the regression estimates. This fallacy is probably grounded on the assumption of missing completely at random (Rubin, 1976), which is often invoked to defend the validity of analyses based on datasets where missing cases are present but not systematically different from the observed cases. However, this implication cannot be extrapolated to the context of ME: even in the simple setting where a explanatory variable included in a multiple linear regression model is slightly affected by classical ME the regression estimates might be biased in unpredictable directions.

2. The second fallacy refers to the idea that ME in explanatory variables will only produce an attenuation bias (a bias towards the null) in the regression estimates. This argument is often used as an excuse to overlook the effect of ME. In particular, an attenuation bias can be thought as the least harmful for
scientific inference since it has an impact on the power to find an specific effect (type I errors) but it does not generate false positives (type II errors). This type of attenuation bias can be predicted for certain regression estimates in very specific ME settings, but it would be wrong to expect to generalise it beyond those special cases.

3. Finally, the last common misconception refers to the belief that ME in the response variable reduces precision but does not produce any bias in the estimates. This is true in standard linear models where the response is affected by classical ME, but again, it cannot be safely generalised beyond those cases.

I now present a set of examples that can be used to illustrate the effect of ME in different settings formally and to refute the above mentioned misconceptions.

### 1.1.3 Classical Measurement Error Affecting the Explanatory Variable in Simple Regression

The case of classical ME affecting the only explanatory variable in a simple linear regression model offers an elegant and intuitive way of appreciating the impact of ME.

Let us take a simple regression where the explanatory variable is measured with error,

\[ Y_i = \beta_0 + \beta_1 X_i^* + \epsilon_i \]  \hspace{1cm} (1.13)

It is well known that, under the assumption that variables are perfectly measured, the ordinary least squares (OLS) estimator of the slope, \( \beta_1 \), is calculated as the ratio between the covariance of \( X \) and \( Y \), \( s_{XY} \), and the variance of \( X \), \( s_X^2 \), so

\[ \beta_1 = \frac{s_{XY}}{s_X^2} \]  \hspace{1cm} (1.14)

However, when the explanatory variable is affected by classical ME the same estimator will now be \( \beta_1^* \), which is formed as the multiplicative relationship of the true slope and a bias term,

\[ \beta_1^* = \beta_1 \left( \frac{s_X^2}{s_X^2 + s_\epsilon^2} \right) \]  \hspace{1cm} (1.15)
This bias is equal to the reliability ratio described in equation 1.5, which by definition is bounded by zero and one, hence biasing the naïve estimate of the slope towards zero. This is the reason why the bias derived from the explanatory variable measured with error in a simple regression model is also known as “regression dilution” or “attenuation bias”.

Invoking the assumption of independence between the error and the true value (4th assumption of the classical model) and their additive relationship seen in equation 1.1, we can see how this biased estimator of the slope is derived,

$$\frac{s_{X,Y}}{s_{X^*}} = \frac{s_{X,Y}}{s_{X}^2 + s_{U}^2} = \frac{s_{X,Y}}{s_{X}^2} = \frac{s_{X,Y}}{s_{X}^2} \left( \frac{s_{X}^2}{s_{X}^2 + s_{U}^2} \right) = \beta_1 \left( \frac{s_{X}^2}{s_{X}^2 + s_{U}^2} \right)$$

The covariance between the observed and the response variable (the numerator of the naïve estimator of the slope) will be unaffected due to the assumption of independence, while the variance of the observed variable (the denominator of the naïve estimator of the slope) will be composed of the sum of the variances of the true variable and the error term.

*Figure 2. Demonstration of the effect of classical measurement error on a explanatory variable using scatter plots and lines of best fit.*

Source: Carroll et al. (2006, p. 42)
The effect of ME in a bivariate relationship can be observed graphically by the comparison of fitted lines for two scatterplots using simulated data where one of the variables is subject to ME in one of the plots. This is shown in Figure 2 above where we can see how the effect of added classical ME in the X variable of the right hand side plot attenuates the line of best fit between the two variables.

This result represents a very elegant demonstration of the effect of ME on the estimate of a regression model. However, as I mentioned before, it should not be generalised beyond this simple setting, where the explanatory variable of a simple regression model is affected by classical ME (Fuller, 1987; and Carroll et al., 2006).

1.1.4. Non-Classical Measurement Error Affecting the Explanatory Variable in Simple Regression

An example that shows the difficulty of anticipating the effect of ME will be in the presence of non-classical ME. Let us take a simple regression model like the one specified in equation 1.13, but where the ME affecting the only explanatory variable can be systematic, as defined in equation 1.7. Notice that the type of error will not only violate the assumption of null expectancy (first item from equation 1.2) but also that of non-differentiality (sixth item from equation 1.2).

Carroll et al. (2006) demonstrated that for such outcome and measurement models, the bias affecting the slope estimate will not be one of attenuation towards the null in a magnitude determined by the reliability ratio. Instead, the naïve estimator of the slope will be defined as follows,

$$
\beta_1^* = \frac{\beta_1 y_1 s^2_X + \rho_{e\epsilon} \sqrt{s^2_\epsilon s^2_U}}{y_1^2 s^2_X + s^2_U} 
$$

(1.17)

So, if $X^*$ is biased in the sense that $\gamma_1 \neq 1$, or if there is significant correlation between the ME and the error about the true line, it is possible for $|\beta_1^*| > |\beta_1|$, an effect exactly the opposite to attenuation.

In summary, in the presence of an ME process that breaches some of the rather restrictive assumptions of the classical ME model could be sufficient to bias.

---

4 “Despite admonitions of Fuller (1987) and others to the contrary, it is a common perception that the effect of ME is always to attenuate the line. In fact, attenuation depends critically on the classical additive ME model” (Carroll et al., 2006, p.46).
regression estimates in ways that are difficult to anticipate\(^5\). In the next and last scenario I will challenge one last misconception referring to how ME affecting the response variable will not bias the regression estimates of the outcome model.

1.1.5. Classical Measurement Error Affecting the Response Variable in Simple Regression

Fuller (1987) demonstrated that classical ME affecting the response variable in a linear model will not produce a bias in the regression estimates. Specifically, if instead of observing the true response variable \(Y\), a different variable \(Y^*\), subject to classical additive ME is observed,

\[
Y_i^* = Y_i + V_i
\]  

(1.18)

the ME term, \(V\), will be absorbed by the residual term of the model having only an impact on the overall precision,

\[
Y_i^* = \beta_0 + \beta_1 X_i + (\varepsilon_i + V_i)
\]  

(1.19)

This result can be seen graphically in Figure 3 below, where the dots on the right-hand side scatterplot are more dispersed across the \(Y\)-axis, but the line of best fit remains practically unchanged.

A similar result holds for other ME models such as the classical multiplicative, where (as seen in equation 1.6) the additive relation between the true variable and the error term is replaced by a multiplicative one, which for the case of the response variable could be represented as follows,

\[
Y^* = Y \cdot V
\]  

(1.20)

Skinner and Humphreys (1999) and Skinner (2000) illustrate how classical multiplicative errors do not produce a bias in the systematic parts of accelerated-life models (see Section 1.2). For the case of the outcome model being an accelerated-life exponential model (which is none other than the linear specification of the logarithmic transformation of a duration variable where the residual term, \(\varepsilon\), follows

---

\(^5\)Furthermore, even in the presence of classical ME, regression estimates could be biased in different directions when an outcome model more complex than simple regression is used. In particular, for a multiple regression model where one of the variables is affected by classical ME, the regression estimates for the variables that are not prone to ME are biased in a direction given by the sign of the correlation between those variables and the one prone to ME (see Carroll et al. 2006, p. 52).
an extreme value distribution; see equation 1.36 in Section 1.2.1) we could follow
the same rationale as in the case of classical additive errors affecting the response
variable seen above. Since the logarithmic transformation of \( Y^* = Y \cdot V \) would be
expressed as \( \log Y^* = \log Y + \log V \) we can see that the impact of classical
multiplicative ME is again absorbed by the stochastic part of the model,

\[
\log Y^* = \beta_0 + \beta_1 X + (\epsilon + \log V)
\]  

(1.21)

**Figure 3. Demonstration of the effect of classical measurement error on a
response variable using scatter plots and lines of best fit.**

![Figure 3](image)

Source: Carroll et al. (2006, p. 339)

The problem is that contrary to what is commonly assumed this result cannot be
safely generalized beyond the simple settings used to derive equations 1.19 and 1.21.
When the outcome model is non-linear, the type of ME is not classical, or both, there
is a strong chance that modelling a response variable prone to ME is going to bias
the regression estimates.

For example, Magder and Hughes (1997) show the effects on a probit model where
the response is a binary variable subject to MC. Models for binary data can be
estimated using maximum likelihood and a link function to adapt the linear
regression to the categorical properties of the response variable. For the case of
probit models, a Normal latent variable is used to map the values of the linear part of
the model to a conditional probability, which could be expressed as 
\[ P(Y = 1|X, \beta_0, \beta_1) = \]
\[ P(Y = 1|X, \beta_0, \beta_1)\theta_{1|1} + P(Y = 0|X, \beta_0, \beta_1)(1 - \theta_{0|0}) \] (1.22)
with \( \theta_{1|1} \) indicating the level of sensitivity, and \( \theta_{0|0} \) the specificity, defined in equation 1.10.

Neuhaus (1999) shows that, for the case of a logit model with just a single explanatory variable and where \( \theta_{1|1} \) and \( \theta_{0|0} \) are bigger than 0.5, \( \beta_1^* \) will approximately converge to
\[ \beta_1^* = \beta_1 \frac{(1 - \theta_{0|1} - \theta_{1|0})\exp(\beta_0)}{[\theta_{0|1}\exp(\beta_0) + \theta_{1|0}][\theta_{0|1}\exp(\beta_0) + 1 - \theta_{1|0}]} \] (1.23)
Hence, although there are settings where the effect of ME in the response variable of a regression model will not affect the systematic parts of that model, those examples are limited to cases where the ME is classical and the outcome model linear. In the presence of complex types of ME, or non-linear outcome models, ME will have the capacity to bias the regression estimates of the outcome model.

### 1.2. Event History Analysis

I now present the group of techniques known as EHA. This could be thought as the second building block of the thesis as in Chapters 4 and 5 I will study the implications and the adjustments of ME using EHA models as the outcome model. I start by presenting the context where EHA becomes relevant, continue by formally reviewing some of its key concepts and then introduce the different families of EHA models (parametric, semi-parametric and non-parametric) separately. After this I present an extension of EHA where repeated events are taken into account and finish by discussing the different observational schemes that have been used to retrieve
EHA data from retrospective questions and the types of MEs that could be expected from each of them.

The group of techniques used in EHA are also known by the names of “time to event”, “hazard modelling”, “survival” or “duration analysis”, depending on the scientific discipline where they are used. However, regardless of the field of application, all of these names refer to the same class of techniques, those that study the durations of spells of a particular state until a transition from that state is made. For example, the duration of the time spent unemployed until a subject becomes employed or abandons the labour market.

The development of this array of techniques stems from the need to overcome some important limitations that standard regression models show when dealing with this type of event data. In increasing order of importance, these are the three major limitations:

1. Since duration data has a lower limit at 0 (it is bound to be positive), it is often the case that the response variable will exhibit considerable asymmetry (skewness to the right). One common fix for this problem is to transform the response variable, for example, using square root or logarithmic transformations. However, what transformation to use is often a subjective decision. Furthermore, after the response variable is transformed the interpretation of the estimates of the model will also need to change.

2. OLS assumes independent and identically distributed observations. Because of this OLS is unable to account for time varying explanatory variables. Here again there are procedures like Generalized Least Squares that could be used to allow for dependence between observations. However, as we will see in the following subsections, there are EHA models such as those from the semi- and non-parametric families that can also account for time varying explanatory variables without requiring any additional extensions.

3. Perhaps the most important circumstance for which OLS methods do not offer adequate solutions is that of censored data. In particular, right-censored data occurs when some observations in the dataset have not experienced a transition from their original state. To treat these censored times as if they represented the time when the event took place would distort the analysis.
since the durations of those episodes would be biased downwards. An alternative solution, discarding censored cases, might be even worse since not only would the sample size be reduced but a form of selection bias might also be induced. EHA models surmount this and the other two problems using likelihood methods. Likelihood estimation allows censored cases to be incorporated into the estimation process by using the information they provide up to the last time at which they were collected\(^6\). This is a preferable solution, although it will require assuming that censoring is non-informative; that is, the censored cases are not different from uncensored ones, and the reason why the event has not occurred is because the time frame has not been stretched far enough.

Before introducing the different EHA models I review two basic concepts for the study of times-to-event that are fundamental to understand how EHA models are specified: the survivor function and the hazard rate.

Let the durations (the time spent in a particular state) be denoted by \(T\), a positive random variable, which for the time being will be assumed to be continuous. Let realizations of \(T\), denoting the duration of a particular unit be denoted by \(t\). If these values are ordered by size a cumulative distribution function, \(F(t)\), can be established as follows,

\[
F(t) = \int_0^t f(t) dt = \Pr(T \leq t) \tag{1.24}
\]

and for all points that \(F(t)\) is differentiable a density function \(f(t)\) can be defined,

\[
f(t) = \frac{dF(t)}{dt} \tag{1.25}
\]

The density function can also be expressed in terms of probability,

\[
f(t) = \lim_{\Delta t \to 0} \frac{\Pr(t \leq T \leq t + \Delta t)}{\Delta t} \tag{1.26}
\]

Hence, when a variable capturing durations is analysed, the cumulative distribution function indicates the probability of observing a case lower or equal to time \(t\), while the density function gives the unconditional failure rate in an infinitesimally small

\(^6\) “A censored event time provides only partial information: it tells you only that the individual did not experience the target event by the time of censoring. In essence it tells you more about event non-occurrence than about event-occurrence”. (Singer and Willet, 2003, p. 325).
differentiable period of time, that is, the proportion of individuals making a transition from one state to another.

In addition, these two functions can be used to define the survivor function and the hazard rate. In fact the survivor function, \( S(t) \), is the complement of \( F(t) \):

\[
S(t) = 1 - F(t) = P r(T > t)
\]  
(1.27)

Thus, the survivor function denotes the probability of a duration being equal to or greater than \( t \), and cases that have not yet experienced the transition from their original state at that time are said to have survived.

Finally, from the concepts of failure and survival the hazard rate can be defined as the ratio between the two,

\[
h(t) = \frac{f(t)}{S(t)}
\]  
(1.28)

Thus, the hazard rate gives the rate at which the unit’s duration ends by \( t \), given that the unit had survived until \( t \);

\[
h(t) = \lim_{\Delta t \to 0} \frac{Pr(t \leq T < t + \Delta t | T \geq t)}{\Delta t}
\]  
(1.29)

In other words, the hazard rate for an interval \([t, t + \Delta t]\) denotes the rate of failure in that interval conditional on survival at or beyond time \( t \).

Here I have shown how the cumulative distribution function, density function, survivor function, and hazard rate are mathematically linked, so, if any one of these is specified, the others can be determined too. However, the estimation of survivor and hazard values for a specific time in continuous data is problematic since they are infinitesimally small.

An alternative is to group durations into intervals. This is the procedure followed in life-tables. However, this solution is not optimal. The basic problem with grouped estimation methods is that they artificially categorize what is by definition a continuous variable, and different categorizations yield different estimates. An alternative to life-tables when estimating the survivor function is to use the Kaplan-Meier method (Kaplan and Meier, 1958).

With respect to life-table methods, the Kaplan-Meier method differs in one fundamental feature: instead of rounding event-times to construct the intervals, it
generates different intervals so that each contains just one observed event at a time. This way, each Kaplan-Meier interval begins at one observed event-time and ends just before the next. The survivor function can be specified as follows,

$$S(t) = \prod_{t_i < t} \frac{n_i - d_i}{n_i}$$  \hspace{1cm} (1.30)

where $d_i$ represents the number of failures in a particular time $t$, the subscript $i$ is used to identify subjects, $i = 1, ..., N$, while the meaning of $n_i$ differs according to whether censored cases are present in the dataset. When there is no censoring, $n_i$ is just the number of survivors just prior to time $t_i$; with censoring, $n_i$ is the number of survivors less the number of censored cases.

The survivor function and hazard rate are fundamental in the exploration of time-to-event processes, but they are only descriptive statistics, that is they only measure parts of the process without controlling for other variables. To ascertain the conditional association of different variables with the observed durations, EHA models need to be used. Most of them are developed from a specification of the hazard rate, which can be treated as having a dependency on time as well as a dependency on the regressors, denoted by $x$. Then we can re-express the hazard rate from equation 1.29 as

$$h(t|x) = \lim_{\Delta t \to 0} \frac{Pr(t \leq T < t + \Delta t | T \geq t, x)}{\Delta t}$$  \hspace{1cm} (1.31)

where the hazard could be understood as an unobserved variable that controls both the occurrence and the timing of events.

More generally, any model used for the analysis of event data can be characterized by two features, the nature of the response variable and whether a specific statistical distribution is used. The first feature distinguishes two groups of EHA models, accelerated life models (AL) and proportional hazard (PH) models. The former group models the observed event time (failure time) directly, whereas the latter models the hazard rate. In terms of statistical distributions, EHA models can be classified into three families: parametric, semi-parametric and non-parametric models. I now review the models that are studied in Chapter 4 using this last distinction.
1.2.1. Parametric Models

The logic underlying parametric event history models is that the time dependency exhibited in event history data is modelled directly. This is done by specifying a density distribution for the failure times. The exponential, Weibull, Gompertz, or Gamma models are examples of models belonging to the parametric family, although in the forthcoming analyses (Sections 4.2.2 and 4.3.2) I focus solely on the Weibull and exponential models. If the Weibull distribution is used, the baseline function (i.e. the function indicating the type of time-dependency) can only be monotonically increasing, monotonically decreasing, or flat with respect to time. The hazard rate is then expressed as:

$$ h(t) = \lambda \alpha (\lambda t)^{\alpha - 1} $$  \hspace{1cm} (1.32)

where $\lambda$ is a positive scale parameter and $\alpha$ is known as the shape parameter. When $\alpha > 1$, the hazard rate is monotonically increasing with time, when $\alpha < 1$, the hazard rate is monotonically decreasing, and when $\alpha = 1$ the hazard is flat, constant at $\lambda$ (this last being the case of the exponential distribution). When conditioning on explanatory variables the Weibull model takes the following shape:

$$ h(t|x) = \alpha (t)^{\alpha - 1} \exp(\beta x) $$  \hspace{1cm} (1.33)

and now the scale parameter is $\exp(\beta x)$, the shape parameter is $\alpha$, and $\beta$ is a row vector of coefficients to be estimated. The impact of the explanatory variables is to alter the scale parameter, while the shape of the distribution remains the same.

The exponential and Weibull models are members of the family of PH models; that is, unit changes in the explanatory variables will imply constant changes in the risk of event-occurrence in a direction and magnitude indicated by their slope estimates. However, these models could also be expressed as modelling the log of the event duration, which turns them into AL models. The Weibull model would then be expressed as:

$$ \log(T) = \beta x + \sigma \epsilon $$  \hspace{1cm} (1.34)

where $\epsilon$ is a stochastic disturbance term with a type-1 extreme value distribution scaled by $\sigma$. Here, in contrast with the specifications using hazard rates, slope estimates are indicative of the change in log-durations (the logs of the expected life time) for different values of the explanatory variables.
The specification for the exponential PH model is shown in equation 1.35 below, which differs from the PH Weibull model defined in equation 1.33 in its use of a constant hazard rate,

$$h(t|x) = h(t)\exp(\beta x)$$

(1.35)

For the transformation into an AL model a process similar to the one for the Weibull case is pursued and a linear model for the logs of the durations is specified,

$$\log(T) = \beta x + \epsilon$$

(1.36)

with a disturbance term $\epsilon$ following an extreme value distribution with mean 0 and variance 1, which in this case is not affected by a scaling factor.

Lastly, regarding the estimation method, parametric EHA models are usually estimated using maximum likelihood. Censoring is controlled for with a likelihood function formed by two components: the density of failure times, $f(t)$, and the survivor function, $S(t)$. This way censored observations only contribute to the likelihood function through the survivor function, and non-censored observations through the density function.

### 1.2.2. Semi-Parametric Models

In some instances, parametric models can be too restrictive. First, it is necessary to allow the hazard rate to depend on time by specifying a distribution that might in fact be inappropriate. Second, these models do not allow for time varying regressors. In semi-parametric models these assumptions are relaxed. Examples from this family are the piecewise constant exponential and the Cox model.

In Chapter 4, I focus on the impact of ME in the latter, developed by Sir David Cox (1972, 1975), which can be understood as a simple generalization of the parametric models presented so far:

$$h(t|x) = h_0(t)\exp(\beta x)$$

(1.37)

The baseline hazard function is left completely unspecified whereas, as for the Weibull and exponential models, the regressors are assumed to have proportional effects on the hazard rate. The estimation of such a model is possible using partial likelihood, a variant of maximum likelihood where only the order of the failure-times is taken into account. As such, each interval between two failures is modelled
separately. For these two reasons the Cox model is often categorised as a semi-parametric model. Furthermore, because of its reliance on partial likelihood the Cox model can be extended easily to allow for explanatory variables that change in value over time.

On the other hand, the Cox model, like any other model assuming continuous time, is subject to problems derived from tied events, which arise when two or more subjects experience a transition at a given point in time. This problem is more frequent the wider the intervals used to collect the data are. It is in these situations, where event data is clearly discrete, that non-parametric methods can be more appropriate.

1.2.3. Non-Parametric Models

Non-parametric models do not examine duration, but rather event counts. Hence, prior to the estimation of non-parametric models, a person-time unit dataset needs to be formed. As opposed to datasets normally used in parametric and semi-parametric models, where each case is characterised by a variable denoting duration, $y_i$, and another indicating whether that event was right censored, $\delta_i$, datasets used in non-parametric models transform the duration of spells into a binary variable representing person-period cases. This binary variable now uses subindexes $i$ and $t$ to differentiate subjects and time-periods; it takes a value of 1 when the failure time is reached, otherwise it is coded as 0, regardless of whether the spell was censored. More formally:

$$y_{ti} = \begin{cases} 0 & t < y_i \\ 0 & t = y_i, \delta_i = 1 \quad i.e. \ censored \\ 1 & t = y_i, \delta_i = 0 \quad i.e. \ not \ censored \end{cases}$$ (1.38)

The discrete-time hazard for interval $t$ is the probability of an event during interval $t$, conditional on the event not having occurred in a previous interval and on the set of explanatory variables included in the model.

$$h(t) = Pr(T = t_i | T \geq t_i, x)$$ (1.39)

Since the dependent variable is binary, constructing models relating this variable to the explanatory variables involves selecting one from a variety of suitable distributions for binary data. Two commonly used link functions are the logistic
distribution and the standard normal distribution. The use of these distributions gives rise to the logit and the probit models, respectively. Here I will use the former, which transforms the response variable into the log of the odds-ratio of failure,

$$\log \left( \frac{y_{ti}}{1 - y_{ti}} \right) = \beta x_{ti} + \theta_{t-1} w_{t-1}$$  \hspace{1cm} (1.40)

where $y_{ti}$ represents the probability of observing $y_{ti} = 1$.

When using the logistic transformation this model is also known as the proportional odds (PO) model because, as in the PH model, changes in the regressors are assumed to induce proportional changes in the response variable. In addition, on the right hand side of equation 1.40, besides the set of explanatory variables of interest, a series of temporal dummies, $w_{t-1}$ (with their respective coefficients, $\theta_{t-1}$), are included in order to specify the baseline hazard function. Each of the dummy variables included represents a period of the time frame (except for the last period, which is not represented by a dummy variable to avoid a problem of perfect multicollinearity), in what is called a piecewise-constant hazard model. Just as for the PH Cox model, the flexibility to estimate the baseline hazard function eliminates the need to choose a distribution to specify that baseline hazard function. However, because of the additional set of parameters required to estimate the baseline hazard function the number of degrees of freedom is reduced, which affects precision, and can become a major weakness when the number of time-periods covered by the window of observation is high.

1.2.4. Single and Repeated Events

The descriptive statistics and models presented so far refer to a particular setting of EHA, one where only the first transition from the state of interest is contemplated. Some life-course events are bound to have such a design, e.g. the study of time to first marriage. However, there are other settings where the event of interest can be repeated. This is the case for the topic of study in this thesis: spells of unemployment.

The difference between designs for single and repeated spells is illustrated in Figure 4 below, which represents the three first cases from the sample of work histories captured in the retrospective question that I use in the analyses in Chapters 4 and 5.
Time in unemployment is represented by a continuous line. States different from unemployment (employed, out of the labour force, etc.) have been aggregated into a single category and they are denoted by a dashed line. The first case considered in Figure 4 shows a transition at day 170, while the third is right censored since it remained in unemployment until after the end of the window of observation. These two cases convey the same information regardless of whether we choose to use a model to account for single or repeated events since they only capture a unique spell of unemployment. However, the same cannot be said about case 2, for which a single events model would only consider the first spell of unemployment terminating at day 50, but a repeated events model would consider that first spell and the second spell spanning from days 120 to 240.

*Figure 4. Example of three work histories durations*

The consideration of multiple spells is generally desirable since studies that choose to look only at the first occurrence of an event ignore the repeated nature of the dataset. More precisely, discarding second and subsequent events implies, first, a reduction in the sample size of events, and second, the possibility of introducing selection bias since this process implicitly assumes that the time to the first event is representative of the time to all events.

However, the use of repeated events models is not always straightforward. In particular, careful consideration needs to be given to issues of dependency of events within subjects, and to whether the repeatable events follow a specific order as opposed to those that are out of order. For example, it is likely that the effect of
certain factors on the time to marriage changes substantially for subsequent marriages, for which we might want to consider some sort of order between spells in a repeated events model. Regarding the problem of dependency between spells of the same subject, two solutions can be explored. We can use marginal models, which treat this dependency as noise, or we can use random effects to try to model it.

Marginal models maintain the assumption of an independent correlation structure, regardless of what the true structure is. These models concentrate on adjusting the estimated variance of the regression estimates, which, in the presence of dependency between observations, would be systematically underestimated\(^7\). The most widely used variance adjustment method applied in marginal models is the “sandwich” variance estimator. Williams (2000) demonstrates that this is an unbiased estimator for cluster correlated data. It is calculated as the product of three matrices: the matrix capturing the observation-level score vectors\(^8\) (the meat of the sandwich), which is pre- and post-multiplied by the model-based variance matrix (the bread of the sandwich).

Multilevel models specify the dependence structure through the inclusion of random effects. Survival models with random effects are also known as frailty models, or shared frailty models when there is more than one observation per group or cluster. These models are often used when interest lies in modelling the heterogeneity arising from unobserved explanatory variables which tend to attenuate baseline hazard functions as subjects drop out of the risk set (i.e the set of those subjects that have not made a transition out of the state under analysis at a time \(t\)).

In its simplest form only one random term is introduced to represent the possibility of random intercepts\(^9\) (RI), that is, different intercepts for each subject. The RI is shared by all spells experienced by the same individual and can be interpreted as subject level unobserved heterogeneity. After controlling for the individual-specific unobservables represented by the random effect, it is assumed that the durations of episodes for the same individual are independent (Steele, 2005).

\(^7\) This is not necessarily the case for models using generalised estimating equations, which are semi-parametric and therefore less sensitive to variance structure specifications than likelihood based models.

\(^8\) Derived from the maximum likelihood estimation process in the parametric and non-parametric models, or from pseudolikelihood in the PH Cox model.

\(^9\) See Steele (2008), and Kelly (2004) for a review of how multilevel modelling can be applied to longitudinal data analysis, and the software available to estimate these types of models, respectively.
This extension can be formally illustrated using the PO logit model specified in equation 1.40 and replacing the temporal dummy variables, \( w_{t-1} \), by the RI term, \( \varphi_i \), which is assumed to be Normally distributed with a mean of zero and constant variance. So the RI PO logit model is defined as follows,

\[
\log \left( \frac{y_{ti}}{1 - y_{ti}} \right) = \beta X_{ti} + \varphi_i
\]

(1.41)

In this model, the log-odds of an event ending at a particular time \( t \) will be shifted up or down by a constant amount for all of the events experienced by that individual \( i \).

1.2.5. Event History Analysis Data

To conclude this presentation of EHA techniques, I review some of the most commonly observational schemes used to retrieve EHA data and the different types of ME mechanisms to which they are prone. In the introduction to this chapter I commented briefly on some of the differences between the two major schemes used to retrieve longitudinal data, the prospective and retrospective observational schemes. In terms of data quality, I noted that prospective designs are not affected by recall errors since respondents are contacted repeatedly across the window of observation and at each time they are only requested to report their current state. On the other hand, prospective schemes tend to suffer from problems of attrition and fail to capture short transitions that occur between periods when the respondent is contacted. Both of these issues could induce a problem of selection bias, but the latter will also generate a problem of ME in the form of omitted spells, which could have especially misleading implications when considering EHA models for repeated spells.

Focusing on retrospective schemes, it is important to differentiate between questions where the respondent is requested to report either the duration of a specific event or the number of times that event has occurred, from more complex questions where the respondent is requested to report entire life-course histories. The former can be reduced to a single variable capturing duration or count data, while the latter offers a richer pool of information that can be studied from different perspectives. One important difference between these two types of retrospective questions relates to the forms of ME that they are prone to. As we will see in Section 3.1, recall errors for the duration of an event or for the number of times that an event took place will be
reflected in misreports of the true values (mismeasurements of the durations or miscounts of the number of events). However, questions where entire life-course histories are reported will be prone to problems of MC of the status or omission of spells. These forms of ME might give rise to work histories in which the true and the reported values are unrelated, which will have important consequences on the configuration of a successful ME adjustment (this phenomenon will be reviewed in detail in Subsection 3.1.1).

To finish, we can make one last distinction regarding the type of retrospective questions used to retrieve entire life-course histories. Lawless (2003) differentiates between “event-occurrence” and “multi-state” designs. The former refers to questions where respondents are asked to identify and date the events experienced in order of occurrence. The latter uses a pre-established window of observation divided into a number of periods within which respondents are asked to identify their experienced state. The main advantage of “multi-state” questions resides in the lower cognitive demand placed on the respondent, who only has to identify the status experienced at different periods, but will not have to recall their dates of start and end. As such, these types of questions should be expected to be less prone to memory failures than those using an event-occurrence design. However, much like for the case of prospective schemes, it is possible that short events occurring within the given periods will be lost.

In principle this problem could be averted - and the accuracy with which the duration of spells is dated could be increased - by using an event-occurrence design. However, as we will see in Chapter 3, problems of either memory failures or interviewee fatigue seem to be the cause of the omission of a great proportion of short spells, while dates of start and end of spells tend to be rounded. In conclusion, we could think of the comparison between multi-state and event-occurrence designs as a good example of the type of compromise that survey designers often need to consider with respect to what we would ideally like to capture as opposed to what can be captured in reality.

In subsequent Chapters I use this classification of types of EHA observational schemes to categorize the datasets that will be used.
CHAPTER 2. DATA

The analyses presented in the following chapters are based on data from the “Longitudinal Study of the Unemployed” (LSA in Swedish), a research project designed by the Swedish Institute for Social Research (SOFI) at Stockholm University (Stenberg, 2010), where responses from retrospective questions on unemployment have been linked to official data from the Swedish register of unemployment (PRESO). Specifically, the collaboration of PRESO in the generation of this validation dataset involved the provision of individual-level data on the work status of those cases selected to participate in a longitudinal survey, two waves of which took place in 1993 and in 2001 (LSA-1993 and LSA-2001).

2.1. Survey Data from the LSA

The two survey waves are relatively similar with respect to the composition of the sample of participants and to the questionnaire. The sample was designed to capture 830 jobseekers randomly selected from those who were registered as such in the PRESO files on 28th February 1992. In addition, participants were selected if they met the following criteria: not employed on that day, willing to start work both immediately and full time, age between 25-54, Nordic nationality, and having no occupational disabilities.

The percentage of subjects responding to the surveys in 1993 and 2001 was 72% and 60%, respectively. Given the different sample sizes caused by attrition, there is a concern that the composition of the sample might have changed from one survey to the other affecting comparisons between the two. However, in terms of two key variables - age and gender –the composition of the sample did not change over time substantially. Men comprised 69% of the 1993 sample and 67% of the 2001 sample. The mean age went from 36.3 years old (with a standard deviation of 8.5) in 1993 to 45.4 (and standard deviation of 8.6) in 2001, but taking into consideration the eight year gap between surveys it appears that the age distribution has not changed either.

10 PRESO is a register from the Swedish employment office (Arbetsmarknadsstyrelsen: AMS).
Limiting the sample to those aged 25 to 54 (in February 1992) means the sample is not fully representative of the Swedish unemployed population. The reason behind this choice was to ensure that people with a higher probability of being economically active were captured\textsuperscript{11}. For the 2001 survey, the age range was pushed eight years forward, capturing respondents between the ages 33 and 63 and thus not being able to represent younger people at all.

The questions used to retrieve work histories in 1993 read as follows:

“Which of the alternative answers on the response card best describes your main activity the first week of 1992? When did this activity start? When did it end?

Which was the subsequent main activity? When did this activity start? When did it end?\textsuperscript{12}

In addition, the last line asking for ensuing activities was repeated twelve times, so, a total of thirteen\textsuperscript{13} “slots” could be examined. Since most of the interviews took place between March and April 1993, respondents needed to reflect about their work histories for a period of approximately fifteen months.

This question changed in the 2001 survey, where it reads:

“I would now like to review the work and other pursuits you have had since January 1990. Consider all the pursuits that lasted at least a month, not only jobs but also parental leave, unemployment, education and the like. Review these pursuits in chronological order until today”.

Both the 1993 and 2001 questions could be considered to employ an event-occurrence scheme. In the previous chapter I differentiated this scheme from the alternative multi-state design. This is an important difference to note since as we will see in the literature review in Section 3.3, these two types of questions have been used by the other two studies assessing the prevalence of ME in retrospectively reported work histories. The Panel Study of Income Dynamics used by Mathiowetz and Duncan (1988) employed an event-occurrence design, whereas the European

\textsuperscript{11} In the codebook “PRESO 1989-1993” it is argued that the age range 18-24 might include many students, whereas the risk for those over 55 years is that they would have left the workforce.

\textsuperscript{12} This and the following quote are translations from the original in Swedish.

\textsuperscript{13} The number of time-periods allowed seems to be enough for respondents to report their total activities experienced in one year since only three participants had ten or more activities to report.

Two main differences between the 1993 and 2001 surveys can be noted. Firstly, the recall time is vastly expanded in LSA-2001; from a time frame of little more than one year to one of eleven years, going back to Jan. 1990 rather than Jan. 1992 as in LSA-1993. Secondly, observations in LSA-2001 are dated on a monthly scale instead of the implied daily scale used in the 1993 survey. On the other hand, the different work status categories were the same in both surveys: “working”, “studying”, “jobseeker”, “unpaid parental leave”, “homeworker” (not employed), “pensioner”, “AMS-training”, and “other”. In addition to work status, two other variables in the 1993 survey are used in the analysis: gender of respondent and interview mode (indicating whether the interview was carried out by phone rather than face-to-face). Regarding the latter, in 1993, 84.7% of the interviews were carried out face-to-face at the respondent’s home, whereas the remaining 15.3% were conducted by phone. In 2001 these figures remained almost identical: 85.4% and 14.6%.

The work history variables that can be retrieved from these questions were unaffected by missing data in both LSA-1993 and LSA-2001. In event history analysis terms the two variables form a multistate multi-episode process. Multistate because respondents start from (and can make transitions to) different states; multi-episode because subjects can have different spells along the window of observation, that is, they are not dropped from the study after the occurrence of one transition to a specific state.

2.2. Register Data from PRESO

PRESO is the Swedish register of unemployment. It collects information from jobseekers on the last weekday of each month to be used for the statistical calculations of the AMS (Arbetsmarknadsstyrelsen / The National Labour Market Board).

The Swedish Employment Service consists of nine services for job seekers and employers: finding work, improving job search, guidance about work, training for
employment, entrepreneurial coaching, clarification of work conditions, work situation adaptation, recruitment, and pre-recruitment training. Because many people use the services of the unemployment office\textsuperscript{14} while at work, the PRESO register not only captures spells of unemployment, but it also offers some validation data on other work statuses such as employment, self-employment or training. Registration is a prerequisite for access to employment policy programmes and to collect unemployment benefits. Since people in Sweden who become unemployed register themselves at AMS in order to get unemployment benefits, the coverage of the PRESO register of all unemployment is close to complete. Korpi and Stenberg (2001) estimated that about 90\% to 95\% of the unemployed are registered as jobseekers in PRESO. In what follows, in order to represent PRESO as a gold standard for spells of unemployment, it is assumed that this proportion is 100\%.

A variable that codes the work status can be retrieved from PRESO on a monthly basis thus generating individual work histories. There are ten work status categories in PRESO and, although these are similar to the ones in the LSA, they are not exactly the same: “Trainee in replacement scheme”, “without job”, “part-time job”, “temporary job”, “permanent job”, “public temporary job”, “youth training scheme”, “AMI” (which also represents training), “labour market training”, and “other”. Because it was a sample design requirement, all the subjects have in common the fact they were \textit{registered} as unemployed in February 1992 (but not necessarily at the beginning of 1992, the start of the recall period in LSA-1993).

Contrary to what was found in analysis of the two LSA surveys, where there were no item-missing responses affecting work status, in PRESO the coverage of spells for statuses other than unemployment is not ideal. Since the decision to register at the unemployment office is a personal one, many of the participants decide to stop using the services of PRESO at different points. Intuitively, this situation could be expected to occur more often with subjects who are employed, especially if those employments are permanent. In addition, it could be argued that those respondents who are less embedded in the labour market, such as first-time workers (those who recently entered the labour market), those who have intermittent work histories, and the long-term unemployed, might tend to drop out of the PRESO file too. The first

\textsuperscript{14} This might be because some subjects are working part-time or on a fixed contract and they are looking for better positions through the guidance services of PRESO.
two might be affected by a lack of familiarity with the register, whereas the long-
term unemployed might be affected by both hopelessness and social stigma.

In general, the proportion of missing cases in the PRESO file is highly variable
across time. For the period 1990-1994 and from the 830 subjects originally selected
for the study, the average number of missing cases in 1990 was 640, decreasing to
564 in 1991, achieving a minimum of 75 in 1992, to increase steadily afterwards to
216 and 249 in 1993 and 1994 respectively. Again, this concentration of registered
cases around 1992 is not surprising since the sample was selected from the register

Another weakness stems from a relative lack of consistency. The PRESO file has
2000/November-2002) where its structure was slightly modified by the inclusion of
new questions and changes in the way they were worded. Work status and its
categories have not been modified though; however in 1997 there was a change in
the definition used by the Swedish government to measure unemployment which had
an effect on the way PRESO registered spells of unemployment.

A final remark that needs to be kept in mind relates to the consideration of the
PRESO file as a gold standard. One example where this assumption would be
violated is in the event of fraudulent behaviours, i.e. if participants are working in
the black market at the same time that they are registered in the PRESO file as
unemployed in order to collect their benefits. This issue is discussed in more detailed
in the conclusions (Chapter 6). For reasons of simplicity, across the analyses to be
presented in this thesis, I assume that the register captures spells of unemployment
perfectly.

Other variables that will be used from PRESO are: age, experience, spells of
unemployment, cumulative unemployment, spell length, and timespan. Experience
captures time-varying self-reported levels of experience for the type of work applied
for. The PRESO questionnaire considers three responses, low, medium and high.
Experience was reported by the subject and it is therefore likely to be subject to ME,
but age, cumulative unemployment and spells of unemployment were directly coded
at the register and can be assumed to be free of ME.
The *spells of unemployment* variable refers to the number of spells of unemployment experienced by the subject over the window of observation. *Cumulative unemployment* captures the number of days the subject has spent registered as unemployed during the window of observation. *Spell length* indicates the overall duration (according to the register) of the first spell reported that includes 1/1/92, and *timespan* captures the number of elapsed days between the day in question and the interview date.

## 2.3. Matched Samples

In this section I present the transformations to the work histories captured in LSA and PRESO that I have carried out in order to match the datasets. The analysis of ME in the reports of unemployment requires matching reported work histories with those found in the register for the same individual and period of time. The first consequence of this is the loss of subjects included in PRESO but dropped due to non-response in LSA. This reduced the initially selected sample of 830 subjects to just 594 in LSA-1993 and 500 in LSA-2001.

A second important deletion of cases stems from a problem found in the register data. I detected that, out of the 830 work histories captured in PRESO, 72 of them showed a problem of “nonsensical dates”. Specifically, those 72 cases contained at least one spell dated to have occurred after the start of the following spell, which generates incorrect overlaps between past and future spells. This problem illustrates the possibility of register data being affected by ME and, therefore, calls into question the validity of seeing PRESO as a gold standard (this issue is further discussed in Chapter 6). To limit the implications of the violation of PRESO not being a gold standard I dropped these cases, which reduced the final sample to 547 in LSA-1993 and 461 in LSA-2001.

Thirdly, during this matching process, I also detected a problem of incomplete dates in LSA-1993. Specifically, 15 of the remaining 547 LSA-1993 cases have at least the start or end of one spell wrongly dated. For example, some include non-existent days.

---

15 In the presence of register data showing similar incongruences for variables of education and nationality Wichert and Wilke (2012) imputed such cases. Here I decided to drop them to exclude the possibility of ME being due to problems in the imputation process.
such as 31st April, while others use days that have been coded as 88 or 99. Depending on how ME is defined (see Section 3.1) this issue of inadequate dates might not always be a problem. However, in some instances such as for the analysis of ME in the duration of spells of unemployment, I proceed by dropping these cases from the final sample. The different sample sizes that will be used in each part of the analysis are presented in the following subsections 2.3.1 and 2.3.2.

Another concern stems from the need to use the same time unit in both LSA and PRESO in order to be able to match their work histories. This is not a problem for LSA-1993 since it is reported in days, the same time unit used in PRESO. However, LSA-2001 demarcates events in months, and, therefore, the PRESO data need to be discretized into months to make the match possible. To do that I followed the procedure used in Pyy-Martikainen and Rendtel (2009). Months were coded as cases of unemployment if they contained at least 28 days of unemployment registered in the PRESO file.

This process of discretization could induce extra ME that is not derived from the interviewees’ responses, especially if reported and true occurrence times are close enough (separated by a few days) but they happen to fall in two different months. Such a case should not be understood as ME because the respondent offered a very accurate answer. However, when grouping events by months, it might appear that respondents did in fact provide a wrong answer. In addition, the coarseness of the time unit does not allow short spells to be captured. As a result, 13.7% of the spells of unemployment captured in PRESO were lost when matching them to LSA-2001 reports because they were shorter than 28 days.

The final transformation carried out on the datasets involves the recode of the different labour status into a simple binary variable. In LSA, spells categorised as “jobseeker” are considered as spells of unemployment, whereas all the other seven categories are grouped into the same status that could be defined as any work status other than unemployment. In PRESO the same process is carried out with the difference that the category taken to represent spells of unemployment is “without a job” with all the remaining nine categories being grouped as “other”.

During this recoding process I detected another issue in the LSA-1993 dataset; 24 subjects reported duplicate spells, i.e. two spells reported by the same individual,
taking place over the same time period, but indicating different work status. These cases probably reflect uncertainty of the respondent regarding the category that best defines their work status. However, all of these duplicate cases refer to statuses that I have recoded as different from unemployment. So, by keeping only the first of each duplicate spell, I do not incur any problems of MC in those work histories.

Every analysis presented in the following section is based on the comparison of survey (error-prone data) with register data which is assumed to be error-free. However, the problems seen above (incapacity to detect fraudulent behaviours, mistakes found in dating the event, omission of short spells due to discretization, and the choice of keeping the first reported spell when duplicate spells are captured in LSA) damage the validity of this assumption. In Chapter 6, the assumption of PRESO being a gold standard is further discussed but in what follows I maintain this assumption and consider discrepancies between the two datasets to be evidence of ME in the responses to the survey.

In the next two subsections I present the samples to be used in the analyses of ME carried out in this thesis. The first subsection covers the samples used in Chapter 3, where the overall prevalence of ME and the mechanisms that cause it are investigated. The second subsection focuses on the final sample used to assess the impact of ME in EHA and the effectiveness of different adjustments, to be presented in Chapters 4 and 5. For each of the samples presented I review their composition and show different estimates of ME that can be obtained, while in Appendix A I include some descriptive statistics of the explanatory variables that will be used in the respective regression analyses.

### 2.3.1. Samples Used in Chapter 3

In Chapter 3 I analyse ME from multiple perspectives. The samples that will be used in each part of this analysis are defined according to three criteria: i) the survey used: LSA-1993 or LSA-2001; ii) whether exact dates are necessary; and iii) whether the focus is on specific spells or entire work histories. Under these criteria I generated seven samples. Each one is named differently in alphabetical order. ‘Sample a’ to ‘sample e’ will be used in Chapter 3, while Chapters 4 and 5 only make use of samples ‘f’ and ‘g’.
Sample a: Section 3.4

Section 3.4 will be the only part of the analysis where the focus is on the assessment of ME at the spell level. In order to assess ME at the spell level I need to find a way of ensuring that the spells from LSA and PRESO to be compared refer to the same event (this requirement is discussed at the beginning of Section 3.4). This is achieved by restricting the analysis to the first spells reported and registered, which can be identified by the fact that they cover the date 1/1/1992 in both the survey and the register.

This restriction reduced the sample size from the 547 respondents in LSA-1993 to just 413 cases. The reason for this loss of 25% of cases stems from the fact that not all the subjects were registered in PRESO on 1/1/92. Remember that the original sample was designed to capture individuals who were unemployed on 28/2/1992, and participation in the register decays substantially when the subject is not unemployed.

Table 1. Status of the first spell reported in LSA and PRESO*

<table>
<thead>
<tr>
<th>Register (PRESO)</th>
<th>Survey (LSA)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unemployment</td>
<td>276 (66.8%)</td>
</tr>
<tr>
<td>Other</td>
<td>33 (8%)</td>
</tr>
<tr>
<td>Other</td>
<td>90 (21.8%)</td>
</tr>
<tr>
<td>Other</td>
<td>14 (3.4%)</td>
</tr>
</tbody>
</table>

*Percentages over the total sample size of 413 cases are included between brackets.

As shown in Table 1, using the recode of spells into ‘unemployment’ or ‘else’ described earlier, about 30% of the subjects included in this sample misclassified the work status of their first spell reported.

Sample b: Section 3.4

In order to estimate the prevalence of ME in spells of unemployment not as misclassifications of the work status, but as misdated durations, a subsample of the 413 cases that comprise ‘sample a’ will be used. I will first restrict the analysis to cases that were correctly classified, 290, and amongst those I will discard the 14 cases found to be referring to a work state different from unemployment, which leaves a total of 276 cases.

To assess the prevalence of misdated ends of first spells, I generate a new binary variable capturing whether the reported end of those spells fell within ±15 days of
the end date of the first spell seen in the register. As we will see in detail in Section 3.4, under this approach, 74% of the first spells reported were found to be misdated.

Sample c: Section 3.5.1

In Section 3.5 I seek to assess the prevalence of ME not in specific spells but in whole work histories. This is done in three subsections, one for each of the definitions of ME that I explore. In Section 3.5.1 I assess ME as the prevalence of miscounted number of spells of unemployment reported in the survey. The exploratory analysis of miscounted spells of unemployment is based on the 547 respondents. In addition, all spells different from unemployment in both the survey and the register were dropped, so the analysis could focus on ME as miscounts of spells of unemployment.

The total number of spells of unemployment captured in LSA-1993 is 793, whereas the register captured 924, which shows a general tendency to omit spells in the survey reports. In fact, despite the selection criterion applied in the original LSA sample design (restricted to subjects registered as unemployed; see Section 2.1), 10.5% of respondents reported no spells of unemployment. In particular, only 54% of the subjects reported the correct number of spells of unemployment.

After the exploratory analysis carried out in Section 3.5.1, this sample is modified to be able to study the mechanisms causing omission of spells of unemployment. The sample of 547 individuals is reduced to 480 after dropping those who had over-reported the number of spells of unemployment experienced. Using this new subsample, I model a binary variable that differentiates between subjects who reported the right number of spells of unemployment against those who omitted at least one such spell.

Sample d: Sections 3.5.2 and 3.5.3

Sections 3.5.2 and 3.5.3 use the same sample from LSA-1993. They study the prevalence of ME in the form of misreports of the overall time spent in unemployment and as mismatches of work-status in person-day cases, respectively. To carry out this analysis I dropped the 15 cases that reported incomplete or impossible dates, as ME will be assessed here at the day-level.

In addition, for the 532 subjects analysed, the start of the window of observation is extended to consider spells of unemployment reported in the survey that ended after
1/1/1992 but started before that date, whereas the day of the interview remains the end of the window of observation. The mean duration of the total time spent in unemployment according to the survey is 100 days shorter than in the register (271 and 371, respectively), while the standard deviation is 16 days lower (172 and 188). However, in spite of the observed tendency to under-report total durations of unemployment I also found that 132 subjects (25%) over-reported their time in unemployment.

For the study of mismatches in person-day observations I rely on the same 532 subjects. However, to avoid analysing periods covered by only a few work histories before Jan. 1992 I restricted the window of observation to an interval encompassing 500 days before the day of the interview. Under this configuration I analyse 245,606 person-day cases reported by the 532 subjects in the sample.

LSA and PRESO are linked at the person-period level and every unit for which there are LSA and PRESO values will be counted as a valid case. These are defined as ME if there are discrepancies (misclassification) between LSA and PRESO categories, regardless of whether the case is a false positive (PRESO registering ‘other status’ and LSA reporting unemployed) or a false negative (PRESO registering unemployed and LSA reporting ‘other status’).

Table 2. Status of person-day cases in LSA and PRESO*

<table>
<thead>
<tr>
<th></th>
<th>Survey (LSA)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Unemployment</td>
</tr>
<tr>
<td>Register (PRESO)</td>
<td>124,562 (50.7%)</td>
</tr>
<tr>
<td>Unemployment</td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>69,809 (28.4%)</td>
</tr>
</tbody>
</table>

*Percentages over the total sample size of 413 cases are included between brackets.

As shown in Table 2, under this definition the percentage of person-day cases correctly reported in LSA-1993 is 67.8%.

Sample e: Section 3.5.1 and 3.5.3

The major part of the analysis is based on LSA-1993, which uses the common recall period of (just over) one year. LSA-2001 comprehends a recall period of eleven years, and will be used only for specific analyses with the intention of illustrating the consequences of using such an extended recall period. Specifically, the samples used from LSA-2001 will be directed at the study of miscounts of the number of spells
and mismatches in person-month observations (Sections 3.5.1 and 3.5.3, respectively).

For these two studies I will use all the 500 subjects who responded to the LSA-2001 survey minus the 39 that contain incorrectly ordered spells in PRESO. For the final sample of 461 subjects, I observe a much higher propensity of omitting spells of unemployment than in LSA-1993 as the average number of spells reported over those registered was 3.2/8.1 in LSA-2001. In addition, only 7.5% of the subjects managed to report the right number of spells of unemployment, whereas the proportion of person-month cases correctly classified was 62.1% of the 21,240 person-months in the sample.

2.3.2. Samples Used in Chapters 4 and 5

In Chapters 4 and 5 I study spells of unemployment captured in LSA-1993 as the response variable in different EHA models, and the effectiveness of different statistical methods seeking to adjust for the impact of ME. In order to simplify these analyses, and to ensure that ME at the spell level can be identified, I make use of the fact that the sample design for LSA was configured as individuals registered as unemployed on 28th February 1992. I set the beginning of the window of observation at February 28th and consider only subjects who started from a state of unemployment in both the register and the survey.

Sample f: Section 4.2 and Chapter 5

After applying the above mentioned restriction a new sample is generated capturing 381 subjects out of the 532 LSA-1993 respondents used in ‘sample d’. In addition, the window of observation is shortened to encompass 381 spells of unemployment taking place from 28/2/92 to 30/03/93, where the ending date represents the earliest day interviews were taken. Under this configuration, 35% of the spells of unemployment in the register (133 in total) were right-censored, whereas in the survey it was only 6% (23 in total).

Sample g: Section 4.3

I will use ‘sample f’ to analyse the impact of ME in EHA models contemplating single spells in Section 4.2, and to assess the effectiveness of different adjustments in Chapter 5. However in Section 4.3 I also assess the impact of ME on EHA models
that allow for multiple spells. To do so, I use the same sample but I consider all of
the spells captured within the same window of observation. This way, whereas
‘sample f’ captures 381 spells both in the survey and in the register, in ‘sample g’, as
a result of the inclusion of repeated events and the effect of miscounting and
misclassification of spells, 559 spells of unemployment can be found in the register
and 706 in the survey.
A shorter version of the following chapter has been published in a peer-reviewed journal:


CHAPTER 3. PREVALENCE AND MECHANISMS OF MEASUREMENT ERROR IN RETROSPECTIVE REPORTS OF UNEMPLOYMENT

In this chapter I present the first of the three blocks of analyses covered in the thesis. The aim is to assess both the prevalence of measurement error in retrospective reports of unemployment and the mechanisms associated with the generation of such errors. Answers to the first point could help researchers understand the validity of data collected from similar survey questions, while the exploration of the second point could contribute to the design of better surveys in the future.

The chapter is organised as follows. First, in Section 3.1, I set out the different forms under which ME could arise when spells of unemployment are dated as part of the report of entire work-histories. The chapter continues with a brief review of the theories that have been employed to explain the mechanisms behind the generation of such ME. These mechanisms are summarised in four hypotheses that will be tested in the following section. Before doing that I present the empirical evidence found in the literature. As noted in this review, there are only two other studies that have analysed ME in retrospective reports of unemployment using register data, something which demonstrates the importance of the present study. One of those two studies is Pyy-Martikainen and Rendtel (2009), which has motivated many of the analyses presented here.

In Section 3.4 and 3.5 I will use data from LSA-1993 and LSA-2001 (the two survey questions presented in the previous section) linked to data from the PRESO register. As explained in the previous chapter, I assume that data from the PRESO register is
a gold standard. That is, I assume the register data are error-free and by comparing it with responses from the survey we are able to ascertain the extent of ME in those responses.

This analysis is presented in two sections: Section 3.4 deals with the study of ME at the spell level by focusing on the first spells reported in the survey and comparing them to what was recorded in the register for the same subject during the same period. Section 3.5 explores the forms of ME that can be identified in entire work histories captured by the survey against those available for the same subjects and period in the register. However, both Section 3.4 and 3.5 share the same two goals: to assess the prevalence of ME in its different forms using descriptive statistics, and to explore the ME-generating mechanisms operating in retrospective questions on work histories by the implementation of different regression models. Finally, in Section 3.6, I conclude with a summary of the main findings and a discussion of their implications and limitations.

3.1. Types of Measurement Error in Retrospective Reports of Unemployment

In spite of its apparent simplicity, the assessment of the prevalence of ME in survey reports using a gold standard is not straightforward, especially when the interest lies in retrospective reports of life cycle events such as work histories. First of all, it is important to differentiate between questions focused on retrieving the duration of a particular spell of unemployment from those that aim to retrieve information on work histories spanning a particular period of time. An example of the former are questions where subjects are asked to recall for how long they have been in their current work status or the day they entered that status (e.g. the day they became unemployed). The retrieval of entire work histories can be achieved either through event-occurrence or multi-state observation schemes (see Section 1.2.5), depending on whether respondents are required to identify and date spells or to simply report the work status experienced in different periods.

In retrospective questions where the interest lies in obtaining the duration of specific spells, the ME can only take two forms. The start of the spell or its total duration can
be misreported - depending on how the question was phrased. ME can also take the form of a misclassification (MC) of the reported work status - as is often the case between spells of unemployment and those that indicate being out of the labour force. Problems of MC can be represented by the specification of false negatives (FN) and false positive (FP) probabilities (equation 1.11), whereas the misreport of duration of spells could be expected to generate distortions of the true durations of the spell that are associated with the classical multiplicative ME (see Section 1.1.5) model.

There are two reasons why the classical multiplicative model is a good specification of misdating errors stemming from memory failures. First, it maintains the scale used in duration data \([0, \infty)\) by eliminating the possibility of observed durations being negative, as could occur under the specification of the more commonly used classical additive ME model. Second, the classical multiplicative model implies that it is more difficult to recall durations that are further away from the day of the interview. For example, we could expect that workers recently made unemployed will recall more accurately the date when that happened than those who have been unemployed for a longer time.

The analysis of ME in the form of misdating and MC of spells can also be studied in the first spells reported under an event-occurrence scheme. These two forms of ME observable at the spell level will be explored in Section 3.4. However, as a consequence of the risk of misclassifying or misdating the first spells, it is not possible to study the presence of ME in subsequent reported spells. In particular, it is not possible to identify whether a second or subsequent spell observed in the survey refers to the same spell captured in the register.

For example, imagine a work history where a subject reported two spells of unemployment but the register captured three; it could be argued that the respondent misclassified the missing spell of unemployment, however, this scenario could also be due to the mismeasurement of one of the two spells of unemployment that has ended up covering the period of the missing one. This and other scenarios are explored in detail in the second part of this section. For now it suffices to say that the ME found in the first reported spells can be propagated to subsequent spells making ME at the spell level unidentifiable and resulting in forms of ME that are only noticeable at the higher work-history level.
From the comparison of work-histories captured by the survey and the register, for the same subjects and periods, we can observe three different forms of ME. First, ME can be defined as mismatches between the number of times a particular spell has been reported and the number of times that spell occurred in reality. Entire events can be omitted and in some cases events that never happened might also be artificially reported. From here on I refer to this type of ME as miscounts in the number of spells of unemployment. Second, ME may be represented as differences between the total time reported as spent in a particular state and the actual time spent in that state. This type of ME will be denoted here as mismeasurements in the total time spent in unemployment. Third, we can think of ME as mismatches between the states reported in each of the time units to which the sample and the window of observation can be reduced. I refer to this form of ME as mismatches in person-day observations.

Each of these three forms of ME are a consequence of the misdating or MC of specific spells, which is what the questions under analysis require from the respondent. However, it is important to study them separately as each of these three ‘derived’ forms of ME relate to different levels of measurement that can be used to explore the phenomenon of unemployment. If we consider the total time spent in unemployment a duration variable can be retrieved; by considering the number of spells of unemployment a count variable can be retrieved; and person-day observations comparing unemployment versus any other possible work status can be used in non-parametric EHA models for discrete data. In order to assess the prevalence of ME in each of these measures they will be studied separately in Sections 3.5.1, 3.5.2, and 3.5.3, respectively.

3.1.1. Implications of the Forms of Measurement Error Arising from Event-occurrence Schemes

Before reviewing the psychological mechanisms producing these forms of ME in the next section I explore the implications that the MC or misdating of first spells have on subsequent spells of the work history. To do that, I assess different scenarios that could take place in different hypothetical work histories. This is an important exercise as it offers a first approximation to the behaviour of the ME arising in these types of questions, which could be used to approximate both the impact that ME will
have in EHA models (Chapter 4) and the best way to model that ME to generate adequate adjustments (Chapter 5).

The implications that result from the generation of different types of ME are illustrated in Figure 5. Here I present three work histories where every subject starts from the same category (unemployed) and only the first spells are considered in a potential EHA model (see Section 1.2.4). I assume that the spells of unemployment denoted by \( Y \) and presented in lines 1, 2, and 3, are the true ones. The error is represented by \( V \) and it is encompassed by the bracket immediately below, whereas the observed durations are represented by \( Y^* \).

When spells are misdated the observed durations can appear shorter or longer than the true ones. This is represented by the first case shown in Figure 5, where the only spell of unemployment experienced within the window of observation has been reported to be 70 days longer than it really was. Misdated errors can also occur amongst the simpler retrospective questions described at the beginning of this section, where subjects known to be in one particular state are requested to report the duration of that spell. As was anticipated before, this type of ME could be adequately specified under the classical multiplicative model.

\[ \text{Figure 5. Work histories affected by ME in a first spells setting} \]
Miscounting the number of spells can result in omitting or over-counting spells. Over-counting spells is not a problem when considering EHA models for single spells, since any reported spell beyond the first transition is not considered by the model. However, the omission of spells could distort estimations based on this data. Take the second work history presented in Figure 5: if the spell representing a status different from unemployment starting in day 30 was omitted, the two spells of unemployment that occur before and after would be linked, and the reported work history for that subject would look like a unique spell of unemployment. The consequences of these types of errors are twofold. The magnitude of ME as a proportion of the true duration is potentially larger than what is seen in errors derived from problems of misdating. Furthermore, the assumption of independence between the true duration and the ME used in the classical framework is violated. Given the fixed timespan covered by the window of observation, the length of the first spell acts as a constraint for subsequent spells. For example, the longer the first spell of unemployment is the lower the possibility of finding subsequent spells that are omitted or misclassified.

Finally, problems of misclassified spells taking the form of FP and FN need to be considered. In FP cases the observed duration is entirely formed by the error term, $Y^* = V$; this is represented by the third work history shown in Figure 5. The fact that the observed duration is not formed by a combination (either additive or multiplicative) of true duration and noise renders the classical framework inadequate here. In addition, FP cases represent an artificial increase of the number of unemployment spells (effectively, the sample size). Given a setting where all work histories start from unemployment and only first spells are considered, the effect of FN resembles one of missing data since FN durations are completely unobserved – the difference lies in the fact that in the presence of missing data we know which cases are missing. Importantly, as long as the probability of committing FN is independent of the duration of unemployment and other observed explanatory variables it will only reduce the precision of the EHA model estimates.

When repeated spells are considered, the implications of these types of retrospective errors become even more complex. Problems of misdating and FP are equivalent to what we have just seen. However, now FN cases would also be possible, having a similar effect to omitting cases before (see case 2 in Figure 5). Moreover, reports
resulting in omission or over-inclusion of spells could alter the sample size and produce different outcomes.

These scenarios are illustrated in Figure 6. The first case shows an example of over-inclusion. This situation would lead to wrongly considering an additional spell of unemployment. The second case shows the second spell of unemployment being omitted; in this case the repeated nature of the work history would be lost. The third case shows an omission of the spell different from unemployment, which would result in linking the first and second spell of unemployment in a single one encompassing the whole window of observation. So, just like for the FP case described in Figure 5, all of these problems of miscounting will now represent a misspecification of the classical model since they will give rise to cases where observed and true durations are entirely unrelated.

Figure 6. Work histories affected by ME in a multiple spells setting

To sum up, work histories captured by retrospective questions are problematic not only because of the higher prevalence of ME that could be associated with memory failures, but also because of the more complex forms that this ME can take. In particular, here I have shown that in event-occurrence schemes, where respondents are requested to identify and date life-course processes such as work histories,
different forms of ME arise that render the classical ME framework inadequate. Finally, I have pointed out that as a result of the effect of ME spreading to subsequent spells covered within a work history, the ME arising from these types of questions could be expected to have a bigger effect in EHA models for multiple spells than when only single spells are considered.

From here on, to differentiate between the types of ME seen in simpler retrospective questions where a single event is to be reported and that arising from event-occurrence questions where multiple spells need to be identified and dated, I will refer to the former as retrospective ME and to the latter as event-occurrence ME.

3.2. Measurement Error Mechanisms in Retrospective Reports of Unemployment

In this section I provide a theoretical review of the psychological mechanisms that have been argued to cause the types of ME that can be found in retrospective questions in general. The following section reviews the empirical evidence found in the literature regarding the ME mechanisms that are hypothesised here, and Sections 3.4 and 3.5 will test those hypotheses using the Longitudinal Study of the Unemployed.

The extent of ME in retrospective questions is mainly related to the saliency of the event and to recall time (Bound et al. 2001). Saliency refers to how much of an imprint the event of interest leaves in the respondent’s memory, recall time measures the time that has elapsed between the occurrence of the event and the date of the interview. The lower the saliency and the longer the recall time, the greater the expected ME. In turn, saliency is affected by interference. Interference as a source of ME arises from the difficulty of discerning the occurrence of specific events when several of them have taken place during the reference period (Mathiowetz et al. 2001).

Other sources of ME that affect reports of work histories (but that are not necessarily unique to the retrospective design) are ‘misunderstanding’ and social desirability. ‘Misunderstanding’, as noted by Tourangeau (1984) refers to the first step in the
cognitive process involved in answering a survey question. This source of ME appears when the question or its possible answers are not fully comprehended by the respondent. For work histories, it refers to the imperfect capacity to discriminate between two or more categories of work status. One example is the sometimes subtle distinction between being unemployed and being out of the labour force. It has been argued that being more embedded in the labour market is associated with being more familiar with its functioning and therefore to favour accurate reports (Bound, et al., 2001; Levine, 1993; Morgenstern and Barrett, 1974; and Paull, 2002).

Social desirability bias appears in value-laden topics. Socially undesirable events tend to be under-reported whereas socially desirable events are often over-reported (Edwards, 1957). Employment and unemployment are respectively the most and least desirable work status categories. Hence, the state of being unemployed might be more prone to under-reporting. “The unpleasantness or social undesirability of time spent looking for work may lead the respondent either to genuinely wipe such occurrences from memory or to consciously fail to reveal them” (Paull, 2002, p. 9).

In addition, Paull (2002) argues that the overall saliency of work status can be expected to be greater for men than for women because of the financial importance of being the prime household earner. In terms of misunderstanding, Bound et al. (2001) argue that population subgroups with lower labour force participation such as women and teenagers are likely to generate more ME because of their lower engagement with the labour market. With respect to social desirability, the long-term unemployed are, arguably, more stigmatised than people unemployed for a short period and are thus likely to generate more ME, especially if the interview uses a face-to-face format (Mathiowetz and Duncan, 1988).

One last cause of ME stems from the process of discretization. In retrospective designs, this mechanism is at work when the question forces respondents to use too coarse a time-unit, causing omissions of short spells, something which might be expected to affect every respondent equally. For example, when respondents are constrained to date spells in months instead of days, spells of unemployment shorter than a month cannot be reported. This type of problem is typical of multi-state observation schemes. Event-occurrence schemes such as the one analysed here are not affected since respondents are requested to date exactly the start and ends of spells.
Finally, there is an additional range of ME-generating mechanisms which are common to any kind of survey: coding errors, interviewer effects, self-acquiescence bias, etc. Figure 7 summarizes the different sources of ME that affect the collection of work histories using a retrospective design. These have been classified in three categories according to the parcel of survey research they stem from.

Figure 7. ME-generating mechanisms in retrospective reports of unemployment

In the analysis to be presented in this chapter I look at all of the above mentioned ME mechanisms except for those that can be generally found in any survey question. Hence I choose to focus on those ME mechanisms that affect retrospective questions or subgroups of the population systematically. To frame such analysis I start by specifying the principles of the error mechanisms affecting retrospective questions on work histories in terms of four hypotheses:

1. Recall time: the probability of generating ME of any form is positively associated with the elapsed time between the interview and the event reported.

2. Interference: the probability of misreporting spells and durations of unemployment is positively associated with the number of spells of unemployment experienced.

3. Misunderstanding: categories of the population that are relatively more embedded in the labour market (e.g. middle age men) produce fewer misclassifications of work status than less engaged groups (e.g. young people and women).

4. Social desirability: the number of spells and durations of unemployment will be under-reported by groups of the population more susceptible to the stigma
derived from being unemployed (e.g. long-term unemployed), and in interviews conducted face-to-face (as opposed to phone or web-based interviews).

3.3. Literature Review

There is a substantial literature on assessments of ME in the report of work histories. However, the relevance of the findings from these studies varies widely according to their design.

Two main research designs have been used to ascertain the presence of ME in surveys at the individual level: replication and validation studies. Replication designs identify ME from the variability in responses to identical questions taken from the same respondents at two or more points close in time. However, because none of the responses are free of ME, it is not possible to estimate the systematic component of the error. Validation designs use a gold standard, a dataset where the true measures for the same subjects are available, so an estimate of non-random variability in error-prone measures can be obtained.

Unfortunately, researchers’ access to official data on people’s work status has been restricted by concerns of confidentiality limiting the number of validation studies currently available. Consequently, despite their limitations, the bigger family of replicated studies generate the majority of useful insights into the presence of ME in retrospective reports of unemployment.

Within the family of replication studies, the most common design compares current work status with the status reported by the same person for the same time point, but derived from a question where respondents are asked to recall their work status during the previous year. Discrepancies between the two measures are seen as evidence of ME derived from memory failures. Most of the studies with this design have used the Current Population Survey\(^\text{16}\), a US panel that is run on a monthly basis. This survey introduced an annual extension, the Work Experience Survey, which included a retrospective question that asked the same subjects to report their work status during the last twelve months. Findings from the first of these studies were collected in the seminal work of Bound et al. (2001). Here, the author

\(^{16}\) [http://www.census.gov/cps/]
summarizes the empirical evidence that has been obtained in the literature for topics such as income, education or unemployment, with an entire section covering retrospective unemployment reports.

From reviewing the studies of Morgenstern and Barrett (1974), Horvath (1982), and Levine (1993), Bound et al. (2001) concluded that unemployment rates were being underestimated by retrospective questions. Furthermore, the rate of underreporting varied notably across different subgroups of the population. Morgenstern and Barrett (1974) estimated that groups whose part-time employment and movement into and out of the labour force is much greater than the average (women and young people), are more likely to omit spells of unemployment. In most cases they declared themselves out of the labour force during this time; “They resort to the social sanctity of cleaning the house, watching after the children, and attending school” (Morgenstern and Barrett, 1974, p. 357). On the other hand, whites, males 25 and over, and females 45 and over all show some tendency to overstate their periods of unemployment in the Work Experience Survey relative to current reporting. The greater attachment to the labour market of these groups might make them remember vacations and “relaxing” periods in between jobs as time spent in unemployment.

Jurges (2007) compares current reports with retrospective work histories for the last year from the German Socio-Economic Panel for the period 1985-2003. This paper specifically analysed the saliency of being unemployed, and arrived at similar conclusions to previous studies: groups that are less embedded in the labour market like women and young people tend to see unemployment as a less salient event than men do and thus their reports are more prone to ME. In addition, this study generated two original findings: first, the saliency of unemployment increased for both men and women during the observation period; second, unemployed respondents who said that they wanted to start employment as soon as possible were much more likely to recall unemployment than others.

During the last 10 years, an alternative replication study design has been more frequently used. Unlike the studies presented above, it doesn’t involve comparisons of current and retrospective questions, but it compares retrospective with retrospective. This design exploits the fact that surveys take a few months to complete. Some longitudinal studies include questions about work status starting from January of the previous year up to the month when the interview was taken,
hence the first months of the year are often captured twice, once when the respondent is referring to the previous year and another when considering her activity in the present. Figure 8 illustrates the period of study (i.e. the overlap), which is marked by the circle.

Figure 8. Comparing two retrospective designs

Discrepancies found between reports about the same period are associated with memory failures. Jacob (2002) used this design and data from the British Household Panel Survey to compare overlapping reports of employment. The author found similar evidence to previous replication studies. For example, the consistency of responses for employed men exceeds 90%, but drops to 69% for unemployed men and to 51% for unemployed women. What makes Jacob’s study unique is his complementary study of the difference in the reliability of employment histories when instead of being self-defined by respondents they are calculated according to the definition of unemployment given by the International Labour Organization\textsuperscript{17}. Jacob (2002) found that 86% of men who defined themselves as unemployed were also considered unemployed using the International Labour Organization approximation but only 44.5% of the women’s self-definition agreed.

Manzoni et al. (2010) used the Swedish Level of Living Survey, where retrospective information on work histories is collected with a time frame of 10 years, and where two surveys were run with the same subjects in the years 1991 and 2000. This particular setting made the period covered by the former (1981-1991) and the latter

\textsuperscript{17} Unemployment occurs when people are without jobs and they have actively sought work within the past four weeks.
(1990-2000) to overlap for the years 1990 and 1991 which allows responses to questions involving a recall period of one year to be compared with others that involve a 10 year recall. The authors found that employment careers appear less heterogeneous according to the report in the second interview (2000) than in the first (1991), that is, the number of episodes reported is much smaller the longer the recall period is extended. The difference is largest (43%) for unemployment episodes. Short spell duration does not make the report of unemployment any worse, while it increases the prevalence of ME in the other work status categories. “This shows that although unemployment spells are shorter and shorter spells contain more errors, unemployment and duration effects are not confounded” (Manzoni et al. 2010, p. 68).

The replication studies presented so far have been able to gauge the effect of memory bias by comparing questions using different recall times, but they cannot properly identify the other causes of ME (social desirability, misunderstanding, etc.) because the two measures involved are both prone to ME. For that, validation studies are needed. Such studies compare the data collected from a survey with a gold standard. However, only three studies on the topic have been published that I am aware of: Duncan and Hill (1985), Mathiowetz and Duncan (1988), and Pyy-Martikainen and Rendtel (2009).

Duncan and Hill (1985) used administrative files of the workers of an American manufacturing firm as a gold standard, and compared these values with the ones reported by the same employees in the Panel Study of Income Dynamics (PSID). They found little evidence for ME when unemployment spells were reported one year later. However, this sample is only made up of workers from one particular firm, something which severely limits the generalizability of the conclusion. It is unlikely that absence of ME holds for the US population at large or for the parts of the population less embedded in the labour market.

Mathiowetz and Duncan (1988) used the same dataset as Duncan and Hill (1985) and paint a rather different picture of ME and its implications, depending on what forms of ME are considered. Respondents offered very accurate answers when asked to report the total time spent unemployed in the last year. However, when required to identify and time each spell of unemployment, results are much less accurate: 66% of spells were omitted. In addition, Mathiowetz and Duncan (1988) modelled the
probability of misclassifying unemployment status for different socio-demographic groups and showed that associations with demographic variables such as ethnicity, education, age and gender were not statistically significant if other variables capturing saliency and interference were controlled for. Saliency was measured by the length of the spell and interference by the number of spells of unemployment that the respondent experienced during the period of analysis. The authors therefore speculate that it is not the condition of being younger or a woman that is associated with ME but their more complex work histories. Another finding from this paper challenges the hypothesis that accuracy deteriorates as the time between the period to be recalled and the date of the interview grows. The authors found that the effect of elapsed time is not linear but quadratic, with the probability of committing an error growing the closer the period is to the interview date up to a point, about five months, where the probability falls sharply.

Pyy-Martikainen and Rendtel (2009) is perhaps the most complete study on the topic. The authors assess the magnitude of ME in retrospective survey data on unemployment collected by the European Community Household Panel using a validation sample obtained from the Finnish Unemployment Office. This dataset, unlike the PSID, is not composed only of workers, which improves the external validity of its results. In addition, different forms of ME (omission of spells, underreported durations, misdated starts and ends of spells and misclassified status) were analysed, offering a very thorough review of the presence of ME in retrospective questions on work histories.

In a first descriptive stage the authors looked at misclassification and misdating in responses. Regarding misclassification, they found that 60.2% of spells in the survey ended in the respondent reported being employed and 2.1% in subsidised work, whereas the percentages in the register data were 53.5% and 11.9% respectively. Misdating was examined by plotting the starts and ends of the spells of unemployment for the survey and register data in a graph. Whereas data from the register seem to be uniformly spread along the calendar, periods from the survey were disproportionally reported as starting in January and ending in December, providing evidence of heaping effects.

In a second stage, the authors modelled both the probability of omitting a spell of unemployment, and the difference between the total times reported and registered in
unemployment. Their results indicate that being female increases the odds of omission by 24% with short spells being harder to remember. In addition, age has a quadratic effect, with the probability of omission decreasing until age 37 and increasing thereafter. Lastly, the authors also found that time spent in unemployment and the number of spells of unemployment generated under-reports, while female respondents over-reported their time in unemployment.

In the following Sections 3.4 and 3.5 I present results from the study of the five possible forms of ME using descriptive and inferential analyses. In doing so I complement Mathiowetz and Duncan (1988), who focused on the problem of misclassification, and Pyy-Martikainen and Rendtel (2009), who studied trends of misdating at the aggregate level and the probability of omission of spells and underreporting of durations at the person level, but did not consider a model-based analysis of MC.

The focus of the analyses will be on LSA-1993 since its time frame is much more common in other retrospective questions on work histories. LSA-2001 will be used in some instances to illustrate the different prevalence of ME when the recall time period is vastly extended. Finally, I determined the set of explanatory variables used in the analysis based on two criteria: I prioritised variables used in Mathiowetz and Duncan (1988) and Pyy-Martikainen and Rendtel (2009) to better replicate their analyses, but I also selected other variables that could help identify categories of the population and features of the interviews (see Sections 2.1 and 2.2).

### 3.4. Measurement Error at the Spell Level

As explained in Section 3.2, in event-occurrence questions ME at the spell level can arise either as a consequence of reports misclassifying work status, or misdating the start/end of spells. However, defining ME at the spell level is not straightforward. Misclassification or misdating first spells can be propagated to subsequent spells resulting in a problem of identification. This is illustrated in Figure 9 below.

Diagram A in Figure 9 represents a work history which is correctly reported since both survey and register capture two spells of unemployment (U) and a spell of employment (E) and they are all correctly dated. Diagram B represents a case where
ME affects the first and second spells reported but mapping each spell in the survey to a corresponding spell in the work history is still possible. In particular, B shows a problem of misdating that results in a shortening of the first spell of unemployment and an extension of the following spell of employment but with the second spell of unemployment correctly reported. Diagram C illustrates a more problematic case, one which prevents identification of the type of ME at the spell level beyond the first spell. In C we might assume that ME affects the first spell by extending it. However, it is impossible to tell whether the second spell (of employment) has been misclassified as unemployment (which would result in the first and third spells of unemployment being linked in error), or whether the end of the first spell was so badly misdated that it covered the whole work history.

**Figure 9. Problems of ME identification at the spell level**

Because of this problem of identification, I restrict the analysis of misdating to the first spells of unemployment in the register that included 1/1/92 and that were correctly reported in the survey on 1/1/92. For the study of misclassification I will use all the first spells reported regardless of their registered work status. This leads to the omission of 22% of the sample since not all the subjects were registered in PRESO on 1/1/92. In addition, to simplify the analysis only two categories will be considered: “unemployment” and “other”; the latter grouping all status categories that are not unemployment.
In order to estimate the prevalence of errors stemming from misdated spells of unemployment I ascertain whether the end of the first spell reported in the survey fell within ±15 days of the end date of the first spell seen in the register. Using this approach, 74% of the first spells reported were found to be misdated.\textsuperscript{18}

This high figure of misdated spells might be partly due to a typical problem regarding misdating in responses to surveys known as heaping effects. Torelli and Trivellato (1993) define these effects as “\textit{abnormal concentrations of responses at certain durations (for questions about elapsed time in a state) or at certain dates (for questions asking when an event took place).}” (p. 189-190). The presence of these errors is assessed graphically in Figure 10, which summarizes the proportion of starts of all spells of unemployment reported at each day of the month in the survey (dashed line), and the ones captured by the register (solid line) across the whole window of observation. The diagram shows that survey participants have a propensity to report the first day of the month as the starting day for their spells of unemployment: 33\% of all spells, much more than in the register (8\%). A second day that stands out is the 15\textsuperscript{th} with 7\%.

\textit{Figure 10. Frequencies of the starts of spells of unemployment by day of the month: LSA 1993}

The presence of heaping effects might suggest that misdated spells are due solely to rounding error (of dates). This potential explanation is investigated by estimating the

\textsuperscript{18}For the analyses of misclassified and misdated spells carried out in this section I use samples a and b, defined in Section 2.3.1.
proportion of reported ends of first spells falling in a wider interval of ±31 days of the registered date. Although this interval eliminates discrepancies due to wrongly reported days, I still find that 57% of those dates remain misdated. So it seems that: i) the problem of misdated spells is very widespread, and ii) it takes the form of both rounded days and mistaken months.

So far I have shown estimates of the prevalence of ME in the form of misdates and misclassifications at the spell level. In order to test the hypotheses presented in Section 3.2, and to explore the mechanisms generating these errors in detail two logit models are specified. The first uses a binary variable capturing misdated spells that fall outside of the ±15 day interval, the second uses another binary variable indicating whether spells were misclassified, and both models use the same set of explanatory variables: age, female, phone interview, experience, cumulative unemployment, spells of unemployment, and spell length. These variables have been presented in Sections 2.1 and 2.2: remember that spells of unemployment records the registered number of spells of unemployment experienced by the subject over the window of observation; cumulative unemployment captures the number of days the subject has spent registered as unemployed during the window of observation; and spell length indicates the overall duration (according to the register) of the first spell reported that includes 1/1/1992.

These two models assume that there is a latent variable \( m_i \) describing the propensity of a person \( i \) to misclassify the work status of the first spell (first model) or to misdate the end of the first spell (second model). The latent variable formulation of the logistic regression model assumes

\[
m_i = \beta x_i + \varepsilon_i
\]  

(3.1)

for subjects \( i = 1, 2, \ldots, n \), where \( x_i \) is a column vector of explanatory variables (including a constant), \( \beta \) is a row vector of the parameters to be estimated, and the error terms \( \varepsilon_i \) are i.i.d. logistic zero-mean variables (with variance \( \pi^2/3 \).) The observed variable is the binary

\[
m_i = \begin{cases} 
1 & \text{if } m_i > 0 \\
0 & \text{if } m_i \leq 0
\end{cases}
\]  

(3.2)

Results for these two models are presented in Table 3. Cumulative unemployment is statistically significant and negative in both models, indicating that the longer a
person stays in unemployment the lower is the propensity both to misdate and to misclassify the first spell. Hypothesis 3 on the propensity to misclassify being inversely related to the level of embeddedness in the labour market is supported by the significant and positive estimate found for female in the MC model. On the other hand, gender does not predict misdating which suggests that the hypothesized problem of these less embedded workers is limited to problems of misunderstanding the differences between work status, and not so much derived from general recall errors.

Table 3. Estimates (standard errors) for the logit models for misdating and misclassification of the first spells reported*, **

<table>
<thead>
<tr>
<th></th>
<th>Misclassification</th>
<th>Misdating</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>-.014 (.014)</td>
<td>.014 (.015)</td>
</tr>
<tr>
<td>Female</td>
<td>.606 (.248)</td>
<td>-.296 (.312)</td>
</tr>
<tr>
<td>Phone interview</td>
<td>.396 (.314)</td>
<td>-.374 (.376)</td>
</tr>
<tr>
<td>Experience</td>
<td>-.303 (.207)</td>
<td>-.322 (.259)</td>
</tr>
<tr>
<td>Cumulative unemployment</td>
<td>-.002 (.001)</td>
<td>-.002 (.001)</td>
</tr>
<tr>
<td>Spells of unemployment</td>
<td>-.114 (.093)</td>
<td>.147 (.046)</td>
</tr>
<tr>
<td>Spell length</td>
<td>-.003 (.001)</td>
<td>.002 (.001)</td>
</tr>
<tr>
<td>LR chi²(7)</td>
<td><strong>50.4</strong></td>
<td><strong>22.84</strong></td>
</tr>
<tr>
<td>Sample size</td>
<td>413</td>
<td>276</td>
</tr>
</tbody>
</table>

* The regression estimates represent the effect on the log-odds of misdating or misclassification.
** Statistically significant results on the 5%-level appear in bold in all tables.

The propensity to misdate the end of the spell increases with the number of spells of unemployment in the register, which corroborates hypothesis 2 on interference. On the other hand, the longer the first spell the lower the propensity to misclassify it. As with gender, this finding suggests that the increased saliency of a spell helps respondents to remember which kind of work status it was but not so much when it was dated. These two results are in line with the main argument posited in Mathiowetz and Duncan (1988) on the difficulty of the task being the main mechanism generating ME. However, the findings shown here serve to nuance that claim by saying that the error mechanisms operating in situations of interference are
specifically reflected in misdated spells, while differences in saliency seem to be particularly associated with the misclassification of spells.

By modelling the propensity to misdate or misclassify spells, direct insights into the error generating mechanisms of interest can be obtained. However, because of problems of identification, I have had to restrict this analysis to the first spells reported. In order to study the presence and nature of ME throughout the whole window of observation I proceed to analyse forms of ME at the work history level. This analysis is particularly relevant because, as described at the beginning of this section, ME in the first spell can be spread to subsequent spells.

3.5. Measurement Error at the Work History Level

Here ME in reports of unemployment at the work history level is studied by comparing the number of spells of unemployment and the number of days spent in unemployment in the register with survey reports for that same period and person. These two contrasts can be used to assess the prevalence of ME in survey reports of unemployment taking the form of count or duration data. That is, they offer insights into the accuracy of the data derived from retrospective reports of unemployment at the level of measurement that is most often used in the study of unemployment. In addition, to deepen the study of the ME generating mechanisms, I also explore ME as mismatches between the survey and register for person-day observations.

3.5.1. Miscounts of the Number of Spells of Unemployment

I start by assessing differences between the number of spells reported and registered. Since this analysis involves the study of the number of spells and not their specific dating, LSA-2001 data (where an 11 year recall period was used) can also be analysed.\(^{19}\)

Despite the selection criterion applied in the sample design (restricted to subjects registered as unemployed; see Section 2.3.2) 10.5% of respondents reported no spells

\(^{19}\) For the analyses of miscounted number of spells carried out in this section I use samples c and e, defined in Section 2.3.1.
of unemployment in the 1993 survey and 16.1% in the 2001 survey. The mean number of spells of unemployment reported by subject over what was registered was 1.4/1.7 in LSA-1993 and 3.2/8.1 in LSA-2001\textsuperscript{20}. These comparisons of means show a tendency to omit spells, especially marked in LSA-2001. The differences in error rate is notable considering that 54% of the subjects reported the correct number of spells of unemployment in LSA-1993 and only 7.5% managed to do so in LSA-2001. These tendencies towards the omission of spells of unemployment can be seen graphically in Figure 11, where histograms for the number of spells reported vs those registered both in LSA-1993 and LSA-2001 are compared.

\textit{Figure 11. Number of spells of unemployment in LSA and PRESO-1993 (top) and in LSA and PRESO-2001 (bottom)}

\textsuperscript{20}t-tests for the differences in spells of unemployment in LSA and PRESO showed a p-value < .0001 in both datasets.
Some of the differences between the two surveys may be accounted for by their main distinguishing features: the longer recall period and the use of months as time units. The effect of the latter was estimated as the percentage of spells recorded as shorter than 28 days, which amounted to 13.7% of the spells of unemployment that were omitted in LSA-2001. However, the difference in the percentage of cases that report the correct number of spells between the two surveys is much greater than that (46.5%), which supports hypothesis 1 on the impact of extended recall time.

To explore the other hypotheses, I estimate a logit model like the one previously used in Pyy-Martikainen and Rendtel (2009), focusing on the LSA-1993 data. The latent variable $m_i^t$ of equation 3.1 now corresponds to the propensity of a person to omit at least one spell of unemployment. The 9.4% of subjects who over-reported their number of spells of unemployment were excluded from the model so I only study omission of spells.

**Table 4. Estimates (standard errors) for the logit model for omission**

<table>
<thead>
<tr>
<th></th>
<th>Omission of spells</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>-.004 (.013)</td>
</tr>
<tr>
<td>Female</td>
<td>.175 (.240)</td>
</tr>
<tr>
<td>Phone interview</td>
<td>.726 (.287)</td>
</tr>
<tr>
<td>Experience</td>
<td>-.252 (.171)</td>
</tr>
<tr>
<td>Cumulative unemployment</td>
<td>-.002 (.001)</td>
</tr>
<tr>
<td>Spells of unemployment</td>
<td>1.285 (.156)</td>
</tr>
<tr>
<td>LR chi²(6)</td>
<td>98.7</td>
</tr>
<tr>
<td>Sample size</td>
<td>480</td>
</tr>
</tbody>
</table>

* The regression (logistic) estimates represent the effect on the log-odds of omission.

Table 4 shows that *spells of unemployment* is positive and significant while *cumulative unemployment* is also significant but negative, which implies that the more spells of unemployment and the shorter they are the higher the probability of omitting them. The former corroborates once again hypothesis 2 on interference, but the latter is an unexpected result since from a social desirability standpoint
(hypothesis 4) the opposite could be expected. This result supports the proposition introduced in the previous section indicating that the long term unemployed might offer more accurate reports because the saliency of unemployment is relatively high. Phone interview also has a significant positive effect, meaning that spells of unemployment are less often omitted in face-to-face interviews, which runs counter to the social desirability hypothesis.

3.5.2. Misreports of Total Durations in Unemployment

A second way of assessing ME at the work history level is by contrasting the total amount of time reported in unemployment against the total registered.\textsuperscript{21} To do this, I extend the window of observation to consider spells of unemployment reported in the survey which had started before 1/1/1992, and the day of the interview remains the end of the window of observation. The mean duration in the survey is 100 days shorter than in the register (271 and 371, respectively), while the standard deviation is 16 days lower (172 and 188)\textsuperscript{22}.

However, in spite of the observed tendency to under-report total durations of unemployment I find that 132 subjects (25\%) over-reported their time in unemployment. These are the cases in Figure 12 that lie above the dashed diagonal line. In addition, Figure 12 shows that longer durations of unemployment are the most severely underreported. The lowess curve (the non-dashed line) shows a divergence between reported and registered durations that becomes especially pronounced after a point around 400 reported days. Notice as well how the 10.5\% of respondents that did not report spells of unemployment seen in Figure 11 are also depicted here. Interestingly, these cases tend to be concentrated amongst those with short durations of unemployment in the register.

\textsuperscript{21} For the analysis of misclassified and misreported durations carried out in this section I use sample d, defined in Section 2.3.1.
\textsuperscript{22} A t-test for the difference of mean durations in LSA and PRESO showed a p-value < 0.0001. An F-test on the difference of variances in LSA and PRESO was also significant, with a p-value of .045.
The error generating mechanisms are explored by modelling the aggregated misreported durations. Specifically, the square roots of the absolute differences between reported and registered cumulative times are taken to eliminate their otherwise right-skewed distribution, and to be able to include cases where unemployment was over-reported.\footnote{The distributions of the absolute differences before and after taking the square roots are shown in Appendix B.}

More formally, the response variable of the model, \( d_i \), is defined as follows:

\[
d_i = \sqrt{\sum_{s=1}^{S_i} y_{si}^* - \sum_{r=1}^{R_i} y_{ri}}
\]  

(3.3)

where \( y_{si}^* \) and \( y_{ri} \) are the durations of a particular spell, \( s \), of reported and \( r \) registered unemployment for subject \( i \); and \( S_i \) and \( R_i \) are the total number of spells reported and registered for person \( i \). A regression model \( d_i = x_i\beta + \epsilon_i \) is assumed for the misreports, where the error terms \( \epsilon_i \) are i.i.d. Normally distributed.
Table 5. Estimates (standard errors) for misreported total durations in unemployment

<table>
<thead>
<tr>
<th>Misreported durations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
</tr>
<tr>
<td>Female</td>
</tr>
<tr>
<td>Phone interview</td>
</tr>
<tr>
<td>Experience</td>
</tr>
<tr>
<td>Cumulative unemployment</td>
</tr>
<tr>
<td>Spells of unemployment</td>
</tr>
<tr>
<td>Constant</td>
</tr>
<tr>
<td>Sample size</td>
</tr>
<tr>
<td>R²</td>
</tr>
</tbody>
</table>

The results in Table 5 show that Phone interview has a positive and significant estimate, which indicates that conducting the survey face-by-face improves the quality of reports in general. However, unlike the model for omission, we cannot reject hypothesis 4 on how the stigmatizing effect of unemployment promotes more under-reporting in face-to-face interviews because here both under and over-reported durations are being modelled. Experience was negative and significant. This result both corroborates and extends hypothesis 3, since it could be interpreted as subjects more embedded in the labour market report more accurate durations in unemployment. That is, their reporting faculties are not just limited to better differentiation of work status.

The positive effect for cumulative unemployment is out of line with what was found in all the previous models where time spent in unemployment was associated with more accurate reports. With respect to the model on misdates we have to take into account that its response variable captured ends of spells misdated by ±15 days, whereas here the actual extent of the misdate is modelled, and as we saw in Figure 12, the magnitude of misreports is particularly large in the cases with the longest
durations in unemployment. So it seems that hypothesis 4 on the effect of social desirability holds for those who have been unemployed for over a year.

3.5.3. Mismatches of Work Status in Person-Day Observations

Finally, I examine ME taking the form of mismatches between the survey and the register in each of the person-day observations covered in the window of observation. Note that this analysis, like those presented in Sections 3.5.1 and 3.5.2, is based on a form of ME derived from misdates and MCs of specific spells. I start the analysis with a simple crosstabulation for the original categories that were used in LSA-1993 and PRESO. The entire crosstabulation for LSA and PRESO status is presented in Appendix C. Here, in order to facilitate its interpretation, the most informative sections of that crosstabulation are presented. This is shown in Tables 6 and 7, where cells report the raw number of cases and the percentages over the column total in brackets; that is, the percentage of person-days found in each cell over the total in that category for PRESO.

In Table 6 it can be seen how a category that is clearly defined such as ‘without job’ engenders a better recall than another one that is somehow more open to interpretation such as ‘replacement scheme’, which stands for full time but temporary employment. The better recall observed for the category without job (64%) than for replacement scheme (54%) should not be explained by social desirability bias since none of these categories reflects a position of special prestige. Nor could it be argued that problems of interference affect recall since replacement scheme occurred less frequently (1.7% of the total cases) than unemployment (75.7%). It seems that the different recalls are related to how well respondents identify the definition of each category, which relates to hypothesis 2 ‘misunderstanding’.

---

24 For the analyses of mismatches of person-day cases carried out in this section I use samples c and e, defined in Sections 2.3.1 and 2.3.2, respectively.
Table 6. Crosstabulation of person-day cases of PRESO and LSA-1993 for the PRESO categories without job and replacement scheme.

<table>
<thead>
<tr>
<th>PRESO LSA</th>
<th>Without job</th>
<th>Replacement scheme</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jobseeker</td>
<td>124,562 (64%)</td>
<td>659 (21%)</td>
</tr>
<tr>
<td>Employee</td>
<td>43,253 (22%)</td>
<td>1,709 (54%)</td>
</tr>
<tr>
<td>Job training</td>
<td>6,646 (3%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>Entrepreneur</td>
<td>7,504 (4%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>Homemaker</td>
<td>1,542 (1%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>Parental leave</td>
<td>4,379 (2%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>Employment development</td>
<td>53 (0%)</td>
<td>781 (25%)</td>
</tr>
<tr>
<td>Other</td>
<td>6,432 (3%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>Total</td>
<td>194,371 (100%)</td>
<td>3,149 (100%)</td>
</tr>
</tbody>
</table>

As opposed to being without job, which is a relatively specific category, replacement scheme can be harder to define. Specifically, 54% of the time spent in replacement scheme was correctly reported as being employed, however, up to 25% was wrongly classified as some sort of subsidized work (employment development).

Another interesting pattern in Table 7 can be identified from the better recall seen in permanent full-time jobs (63%) than in part-time employment (58%). Presumably this difference would be derived from social desirability or saliency reasons; although it is hard to distinguish which of the two sources has a bigger effect.

The crosstabs presented in this section can be used to shed some light on ME generating mechanisms such as those of misunderstanding or social desirability. However, as explained in Section 3.4, it needs to be taken into account that ME on earlier spells can be propagated across the reported work history making impossible the identification of subsequent spells and how they are affected by ME. In consequence, a proportion of the mismatches observed in Tables 6 and 7 should be understood as evidence of ME that is not necessarily reflecting misclassification between different work statuses. To facilitate such comparisons results from the first and last spells reported in LSA-1993 are now presented.
Table 7. Crosstabulation of person-day cases of PRESO and LSA-1993 cases for the PRESO categories permanent job and part-time employed.

<table>
<thead>
<tr>
<th>PRESO LSA</th>
<th>Permanent job</th>
<th>Part-time employed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jobseeker</td>
<td>304 (19%)</td>
<td>3,004 (31%)</td>
</tr>
<tr>
<td>Employee</td>
<td>1,019 (63%)</td>
<td>5,657 (58%)</td>
</tr>
<tr>
<td>Job training</td>
<td>190 (12%)</td>
<td>404 (4%)</td>
</tr>
<tr>
<td>Entrepreneur</td>
<td>96 (6%)</td>
<td>181 (2%)</td>
</tr>
<tr>
<td>Homeworker</td>
<td>0 (0%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>Parental leave</td>
<td>0 (0%)</td>
<td>342 (3%)</td>
</tr>
<tr>
<td>Employment development</td>
<td>0 (0%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>Other</td>
<td>0 (0%)</td>
<td>94 (1%)</td>
</tr>
<tr>
<td>Total</td>
<td>1,609 (100%)</td>
<td>9,682 (100%)</td>
</tr>
</tbody>
</table>

Table 8 gives the percentage of subjects reported to be in each category of LSA in January 1st 1992 over the total number of subjects that were registered as unemployed by PRESO. Table 9 represents that same ratio but for the interview dates of each subject (from March to April 1993). By choosing these two important dates that refer to the first spell to be reported and to the current status, we can be certain that mismatches in person-days observations reflect genuine misclassifications between spells.

Table 8. Classification of spells of unemployment in January 1992*

<table>
<thead>
<tr>
<th>PRESO LSA</th>
<th>Without job</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jobseeker</td>
<td>87%</td>
</tr>
<tr>
<td>Employee</td>
<td>7%</td>
</tr>
<tr>
<td>Training</td>
<td>2%</td>
</tr>
<tr>
<td>Entrepreneur</td>
<td>2%</td>
</tr>
<tr>
<td>Homeworker</td>
<td>0%</td>
</tr>
<tr>
<td>Parental leave</td>
<td>0%</td>
</tr>
<tr>
<td>Employment development</td>
<td>0%</td>
</tr>
<tr>
<td>Other</td>
<td>2%</td>
</tr>
<tr>
<td>Total</td>
<td>100%</td>
</tr>
</tbody>
</table>

*The sample size of subjects available in in PRESO in January 1992 was 209.
The difference in recall accuracy is counterintuitive and marked: 87% for the first day of the time frame and 55% for the last. One possible explanation could be that during the cognitive process of answering the question the moment of highest concentration would be directed towards recalling the status at the beginning of the period of observation. In addition this first spell is not affected by errors in the report of subsequent spells. This could be a common effect of event-occurrence frameworks where the questionnaire emphasises the recalls of the status for the first spell, while simply asking to date the start and report the status of following events.

**Table 9. Classification of spells of unemployment for the day of the interview***

<table>
<thead>
<tr>
<th>PRESO LSA</th>
<th>Without job</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jobseeker</td>
<td>55%</td>
</tr>
<tr>
<td>Employee</td>
<td>25%</td>
</tr>
<tr>
<td>Training</td>
<td>7%</td>
</tr>
<tr>
<td>Entrepreneur</td>
<td>5%</td>
</tr>
<tr>
<td>Homeworker</td>
<td>2%</td>
</tr>
<tr>
<td>Parental leave</td>
<td>3%</td>
</tr>
<tr>
<td>Employment development</td>
<td>0%</td>
</tr>
<tr>
<td>Other</td>
<td>3%</td>
</tr>
<tr>
<td>Total</td>
<td>100%</td>
</tr>
</tbody>
</table>

*The sample size of subjects available in PRESO for the dates of their interviews was 236.

Another explanation might be derived from reasons of social desirability. It could be argued that a social desirability bias grows in intensity the closer the reported time is to the date of the interview. Perhaps it is easier to report unemployment spells dated a year before than to report being currently unemployed. In general, this difference between levels of misclassification between the first and last periods could be used to argue that recall accuracy does not necessarily decay with time, which would refute hypothesis 1. However, for that it would be better to control for other potential effects.

As in the analysis of misclassification at the spell level to do so I will consider only two categories: “unemployment” and “other”, the latter grouping all categories that are not unemployment. LSA and PRESO are linked at the person-period level and every unit for which there are LSA and PRESO values counts as a valid case. These are defined as ME if there are discrepancies (MC) between LSA and PRESO categories, regardless of whether the case is a false positive (PRESO registering
other and LSA reporting unemployed) or a false negative (PRESO registering unemployed and LSA reporting other).

Under this definition the percentage of cases correctly reported in LSA-1993 and LSA-2001 is 67.8% and 62.1%, respectively, which contributes to the validation of hypothesis 1 indicating that recall accuracy deteriorates with time. However, that rate only decayed by six points, when the time frame is about ten times longer, making the effect less substantial. In addition, just as was anticipated in Tables 8 and 9, the MC rate varies widely across the reporting period. This is depicted in Figure 13, which shows quadratic relationships for LSA-1993 and LSA-2001 between timespan (that is, the time from the first reported day until the interview) and the proportion of person-days mismatched across their respective windows of observation.

*Figure 13. Scatter plot of the proportion of misclassified time-units and time-span from the interview in LSA-1993 (above) and LSA-2001 (below)*

*The timespan units are days (top graph) and months (bottom graph).*

In order to analyse this effect in more detail together with the study of other error generating mechanisms I proceed to model the probability of mismatches using

92
person-day observations of LSA-1993. To do this I implement a random effects logistic model, where the probability of mismatch is specified in terms of the latent variable

\[ \theta_{ij}^* = \beta x_{ij} + \xi_i + \epsilon_{ij} \]  

(3.4)

with \( \theta_{ij} \) defined as \( m_i \) in equation 3.2, except for the new term \( \xi_i \), \( (i = 1, \ldots, n) \), which captures the unexplained variability between the level 2 units (persons), and the subscript \( j \) \( (j = 1, \ldots, J_i) \) which is now used to differentiate amongst person-day units. The error term \( \epsilon_{ij} \) is again assumed to be independent and to follow a logistic distribution. Following standard practice in multilevel modelling (Snijders and Boskers, 1999) the person specific error terms \( (\xi_i) \) are assumed to be i.i.d. Normal with mean zero and variance \( \sigma^2_{\xi} \). 25 The variability of the response variable with elapsed time is modelled by including timespan as a predictor.

Results from the model are presented in Table 10. Higher levels of work experience are associated with lower propensities of observing mismatches between the survey and the register, which supports hypothesis 3 on the level of embeddedness in the labour market. On the other hand, cumulative unemployment is associated with an increase in the propensity to mismatch. This finding reinforces the contradiction seen before, warning us against making any strong claims on the effect of the stigma of being unemployed (hypothesis 4) on the accuracy of reports. Finally, the estimates for timespan show the expected quadratic effect anticipated in Figure 13, which supports the argument on the propagation of errors from previous spells, contradicts other views on the topic (such as Bound et al. (2001), or Solga (2001)) that assumed independence of errors across the window of observation26, and raises some questions about whether hypothesis 1 (on the effect of time on ME) can be supported as it is currently worded.

25 I used the Gauss-Hermite quadrature approximation for obtaining the maximum likelihood estimates. The random intercepts models were replicated using MCMC estimation using the MLwiN software in order to assess their robustness. Similar regression coefficients were obtained in both models, although standard errors were higher in the MCMC models. Mathiowetz and Duncan (1988) adopted a similar approach in their analysis of ME although they used jacknife replications of their sample in order to calculate the variance of the regression coefficients across all replicates and thereby adjusting their standard errors.

26 “The extent to which classification error in one month biases estimates of transitions between statuses depends on whether the errors are persistent or independent from one month to the next. Lacking direct evidence on this score, analysts assume that the errors in one month are unrelated to errors in the next.” Bound et al. (2001, p. 68)
Table 10. Estimates (standard errors) from the random effects logit model for mismatch*

<table>
<thead>
<tr>
<th>Mismatch (person-days)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
</tr>
<tr>
<td>-.006 (.016)</td>
</tr>
<tr>
<td>Female</td>
</tr>
<tr>
<td>.106 (.302)</td>
</tr>
<tr>
<td>Phone interview</td>
</tr>
<tr>
<td>.290 (.382)</td>
</tr>
<tr>
<td>Experience</td>
</tr>
<tr>
<td>-.569 (.034)</td>
</tr>
<tr>
<td>Cumulative unemployment</td>
</tr>
<tr>
<td>.004 (.001)</td>
</tr>
<tr>
<td>Spells of unemployment</td>
</tr>
<tr>
<td>-.107 (.190)</td>
</tr>
<tr>
<td>Timespan</td>
</tr>
<tr>
<td>.019 (&lt;.001)</td>
</tr>
<tr>
<td>Timespan2</td>
</tr>
<tr>
<td>&lt;.001 (&lt;.001)</td>
</tr>
</tbody>
</table>

| Intra-cluster correlation                     |
| .756 (.009)                                   |
| Level 1 units                                 |
| 245,606                                       |
| Level 2 units                                 |
| 532                                           |
| Wald chi²(8)                                  |
| 8,678.6                                       |

* The estimated regression estimates represent the effect ceteris paribus of a unit change in the variable on the log-odds of mismatch.

The argument that due to memory failures the extent of ME is positively related to length of the timespan between the time of the interview and that when the reported event took place is self-evident. However when reporting work histories (or any other series of life-course histories), errors from the same subject are positively correlated in a manner that past errors increase the probability of committing errors in subsequent periods. Hence, when looking at the probabilities of person-period cases being misclassified in a particular time frame two simultaneously occurring error generating mechanisms might be at work: i) a “memory failure” mechanism as defined in hypothesis 1 that increases the deterioration of reports the further they are from the date of the interview, and ii) a “ME propagation” mechanism increasing the probability of wrongly reporting spells as a result of preceding events being affected by ME. Further research on this topic is required.
3.6. Discussion of the Measurement Error Mechanisms

In this Chapter I have assessed both the prevalence and nature of the underlying mechanisms associated with ME in retrospective reports of unemployment. This has been done by implementing an original approach which analyses different ways of operationalizing ME when validation data is available. For the survey question under consideration, ME can either occur as a result of the MCs or misdates of reported spells. However, because of the identification problem derived from the propagation of errors across time, the analysis of these forms of ME had to be restricted to the first spells reported.

In Section 3.5 the analysis was extended to the rest of the spells found within the window of observation by using forms of ME that can be observed at the work history level. These are: omission of spells, misreports of aggregated durations, and mismatches of person-day cases. In doing so I have combined and elaborated Mathiowetz and Duncan’s (1988) analysis, where only the prevalence of mismatches of observations was modelled, and Pyy-Martikainen and Rendtel’s (2009) study where only omission of spells and underreported durations were modelled.

For the LSA-1993 question, which involves recalls of 12 to 15 months, I found that 74% of the end dates of first unemployment spells reported were misdated by more than ±15 days, while 30% of the subjects misclassified the work status of all the first spells. A tendency to omit the number of spells experienced was detected; in particular only 54% of the subjects reported the correct number of spells of unemployment, with that figure plummeting to 7.5% when the recall time was extended to 11 years. Finally, the mean duration of the total time reported in unemployment in the survey was 73% of the length of the time captured in the register while the percentage of person day cases reported that matched the status found in the register was 67.8%.

Some of the forms of ME analysed here use the same level of measurement that are typical of variables derived from retrospective reports of events. For example, the miscount of the number of spells has direct implications if work histories are to be used in count data analyses (e.g. a Poisson model), misreporting spell lengths affects models relying on duration data (e.g. event history analysis for continuous data such
as the accelerated failure time Weibull model set out in Section 1.2.1), and mismatches of person-period observations affects data in categorical form (e.g. event history analysis for discrete data such as the proportional odds model set out in Section 1.2.3). Hence, quantifying the prevalence of ME in multiple forms serves the purpose of making users of this type of data aware of the magnitude of the problem.

However, the main contribution of this chapter stems from a better understanding of the mechanisms underlying the generation of ME in retrospective questions. This study represents the first analysis using a validation design that models forms of ME directly operating at the spell level. Moreover, the inclusion of models specifying additional forms of ME operating at the work history level makes the analysis more comprehensive than previous studies of the topic.

Returning to the hypotheses set out in Section 3.2, the following can be concluded:

Hypothesis 1 specifies that the probability of generating ME in any form is positively associated with the elapsed time between the interview and the event reported. I find that while this statement is true between questions, it is not necessarily so within questions. For example, when comparing the main question (LSA-1993) with another one (LSA-2001) that uses a recall period 10 times longer in the study of the number of spells of unemployment reported, much greater prevalence of ME in the form of omission of spells of unemployment is found in the question using an extended recall period. However, when modelling the effect of time on the probability of misclassifying person-day observations for the shorter recall period I show that its effect is not linear but quadratic. That is, periods that are further away from the interview date are not necessarily associated with higher probabilities of being misclassified.

Hypothesis 2 states that the quality of the recall is negatively associated with the number of spells experienced. In the analyses presented here, the number of spells of unemployment increased the propensity to misdate the ends of first spells and to omit spells of unemployment, but it is not significantly associated with all the other forms of ME. These results serve both to corroborate the hypothesis and to identify the particular forms of ME which are affected by interference. In particular, it is interesting to note that the lower saliency that is assumed to result from interference does not affect the probability of misclassifying the first spell, whereas the length of
that spell, which could also be understood as a proxy for saliency, reduces the probability of misclassifying but does not generate misdates.

Hypothesis 3 indicates that groups of the population relatively more embedded in the labour market differentiate better between work status categories. I find some evidence pointing in that direction, although in many instances results are inconclusive. Women have a higher propensity to misclassify first spells, while no effect is found for the other forms of ME. This suggests that the error mechanism derived from lower levels of embeddedness specifically affects the capacity to distinguish between categories of work status, whereas other problems affecting the capacity to date spells correctly are common across population groups. Age is not found to be significant for any of the forms of ME. However, the validity of my findings regarding age is limited because younger and older subgroups of the population were deliberately omitted in the sample design.

Alternative evidence to test this hypothesis can be derived from the inclusion of a variable capturing the work experience of respondents. This allows ascertaining whether subjects who are well embedded in the labour market actually make better reports without relying on using age and gender as proxies. I find that the level of experience is negatively associated with misreported durations and the probability of finding mismatched cases, thus extending the error generating mechanism postulated in this hypothesis to other forms of ME that were not previously associated with it.

Hypothesis 4 states that being unemployed has a stigmatizing effect that leads to under-reports of unemployment. The evidence obtained here support but also nuance this hypothesis. I find that the longer the time spent in unemployment the lower the propensity to misdate and misclassify first spells, and to omit spells in general. Moreover, regarding the two survey modes, interviews by phone are found to increase the propensity to omit spells. However, it is important to bear in mind that survey mode was not randomly assigned to subjects. Also, most importantly, it needs to be considered that the model predicting misdates of first spells only discriminate between cases misdated by more or less than 15 days. When considering the absolute magnitude of misdates I find evidence of strong under-reports of time in unemployment for persons who have been unemployed longer than a year.
Finally a caveat regarding the validity of the analysis needs to be made. In theory, the use of validation data represents an improvement in the study of ME in surveys compared to studies relying on replicated data. In practice, however, that depends on how close the validation data is to the true values. Here I have assumed that the register data from PRESO are a gold standard, but there are reasons to suppose that this assumption might not always hold. Administrative data will be affected by coding errors. Moreover, the definition of unemployment for the AMS changed in 1997, which might produce artefactual variations when analysing ME in the 2001 survey. Finally, PRESO might be prone to systematic errors in that some persons registered as unemployed might in fact have casual employment. The assumption of PRESO being a gold standard is further discussed in Chapter 6.
A shorter version of the following chapter has been published in a peer-reviewed journal:


**CHAPTER 4. IMPLICATIONS OF MEASUREMENT ERROR IN EVENT HISTORY ANALYSIS**

In this chapter I study the consequences of using durations of unemployment prone to event-occurrence ME as the response variable in EHA models such as those seen in Section 1.2. The impact of ME is assessed by comparing estimates obtained from models that are specified using durations of unemployment derived from LSA against those obtained using durations from PRESO (the gold standard) in the same model.

In choosing to study the consequences of ME in the response variable of EHA models I address an area which has been under-researched. In the analysis of ME a majority of studies have focused on settings where the explanatory variables are prone to ME, in what is known as the “errors in variables” problem. This focus on the predictors can be explained from the widespread belief that ME affecting the response variable only has an impact on the precision of the model and therefore it is a lesser problem (equations 1.19 and 1.21). In addition, the study of ME was until recently restricted to analyses using linear models, with the seminal work of Fuller (1987) as the main reference. In the last decade the study of ME has been extended to other non-linear models, especially as a result of the publication of Carroll et al. (2006). However, as has been noted by many authors, within the study of ME in non-linear models, EHA is still an area in need of further contributions (Augustin (1999), Pyy-Martikainen and Rendtel (2009), Skinner and Humphries (1999), and Jäckle (2008)).

This chapter is structured as follows. In the next section I present a review of the literature. The analysis is presented in Sections 4.2 and 4.3. In the former I explore
the effect on EHA models where only the first reported spell of unemployment is considered. To detect possible differences in the impact of ME due to the EHA model implemented I carry out the analysis using four models: an AL exponential (equation 1.36), AL Weibull (equation 1.34), PH Cox (equation 1.37), and PO logit (equation 1.40). That is, at least one model for each of the EHA families (parametric, semi-parametric and non-parametric) will be explored. In Section 4.3 the same analysis is replicated with the difference that here the models will be extended to account for multiple spells experienced by the same subject. Finally, Section 4.4 concludes with a summary of the results obtained in Sections 4.2 and 4.3, and with an explanation on how these results relate to previous findings from the literature.

4.1. Literature Review

According to the research design used we can identify two main groups of studies that have assessed the impact of ME in EHA. These can be analytical or empirical. The former imply tracing out the impact of ME in EHA models algebraically. However, because of the greater complexity of EHA models, the number of settings explored is much more limited than for the case of linear models. In fact, until the 1990s, research was concentrated on classical ME affecting explanatory variables in the PH Cox model. Some examples are Prentice (1982) and Nakamura (1992) who presented an analytical development of the bias found in the parameter estimates of PH Cox models with classical ME in the explanatory variables. In this context, both authors found attenuation bias in all the regression estimates.

The only studies that have explored analytically the impact of ME on the response variable in EHA models that I am aware are the working papers by Augustin (1999) and Dumangane (2007). They used AL Weibull models and assumed classical multiplicative errors affecting the recall of durations. In this case, ME in the response was found to produce an attenuation bias in the regression estimates.

This type of setting where ME is assumed to take the form of misdated durations may be realistic for state-based observation schemes. Holt et al. (1991) define such schemes as those used to collect the spell duration for subjects known to be in a particular state. However, as we saw in Section 3.1, event-occurrence questions can
generate different forms of ME such as omission of spells, or misclassification of statuses, which are not accounted for by the particular setting assumed by Augustin (1999) and Dumangane (2007). In addition, Augustin (1999) requires the assumption of no right censoring in the data and Dumangane (2007) assumes that the true duration and error distributions are independent. The set of assumptions used in these papers shows both the difficulty of studying the effect of ME in the response variable of EHA models analytically, and how the general expressions developed so far are not really representative of the problems found in retrospective data, which are prone to other types of ME besides mismeasured durations.

Another group of studies assessing the impact of ME in EHA are those that have carried out an empirical analysis. These studies compare estimates derived from a model that uses prone to ME data against the estimates obtained from replicating the same model but using data free of ME. Korn et al. (2010) studied the effects of ME in a PH Weibull model by means of simulating multiplicative log-normal errors in the response, which captured times until death of breast cancer patients. The authors found small downward biases in the hazard rate as long as the ME remains non-differential. However, the authors also demonstrated that the degree of attenuation increased substantially when the ME is differential. This scenario was explored using the only variable included on the right hand-side of the model, a binary variable indicating whether observations stem from the control or treatment arm. Specifically, ME was made non-differential by increasing the size of the ME in the survival times taken from the experimental arm.

Considering non-parametric models for discrete data, Meier et al. (2003) assess the bias in the regression estimates produced by simulating different levels of non-differential false positive (FP) and false negative (FN) in the response variable (that is, probabilities of FP and FN independent from the values of the explanatory variables). The authors conclude that the bias is always toward the null, and that FP cases induce greater bias than FNs when the hazard rate is low. Moreover, for this case of EHA (a non-parametric model using discrete data), additional findings can be obtained from studies on the impact of misclassification (MC) in the response variable in more general models for categorical data. In this respect, Magder and Hughes (1997) show how a response variable subject to MC could generate bias in the regression estimates of a logistic regression. Neuhaus (1999) derived general
expressions for the magnitude of the bias due to MC in the response for different types of regression models for binary data (logistic, probit, complementary log-log). The authors found that ignoring MC in the response variable leads to attenuation bias effects when the errors are independent of the explanatory variables. However, when MC probabilities depend on the explanatory variables, ignoring these errors can lead to bias away from the null.

An early attempt to look at the effect of retrospective ME in the response in EHA models was Holt et al. (1991). Here, the authors compare different AL Weibull models where duration of unemployment was regressed on age. A sample of durations was simulated and different types of differential and non-differential multiplicative ME were superimposed on them. The comparison of free-from-ME with prone to ME models showed substantial biases in the estimate of the constant but much smaller biases for the estimate of age and the power parameter of the baseline hazard function. Much larger biases were found when the ME was correlated with age (the explanatory variable), except for the power parameter of the baseline function, which remained relatively unbiased across the different scenarios studied.

Skinner and Humphreys (1999) used another case presented in Holt et al. (1991), where data capturing age at first menstruation in women with no right-censored cases was simulated, to study alternative types of ME. The authors simulated different types of non-differential errors (additive vs multiplicative, homoscedastic vs heteroscedastic, in different combinations) on both the durations of spells of unemployment and for age at menarche. Their findings showed that under the assumption of no censoring, the regression estimates of a Weibull model are approximately unbiased when MEs affecting spells are independent of each other, of the spell durations and of the explanatory variables. The estimator of the shape parameter that determines the duration dependence of the hazard is, however, biased. The authors traced the effects analytically and noted that if ME is related to explanatory variables then the estimators of the corresponding estimates are likely to be biased.

These last two studies contribute to the understanding of the effect of ME affecting the response variable in duration models. However, just like the studies of Augustin (1999), Dumangane (2007), and Korn et al. (2010), they do not consider the fact that
ME also takes the form of omitted spells and misclassified status. Hence their studies do not entirely reflect the consequences of using retrospectively reported work histories.

Other interesting studies that represent the effect of ME stemming from event-occurrence questions more closely than those reviewed so far are Courgeau (1992), Peters (1998), and Jäckle (2008). The former studied reports of dates when couples moved to their current residency, finding – in general – no substantial impacts in the regression coefficients of parametric, semi and non-parametric models when the reported durations where used as the response variable instead of register data. Peters (1998) compares survey data captured from prospective (measuring current states) and retrospective questions on different life-cycle events: time to first marriage and time to first divorce. The durations of the two events are specified using a PH Weibull model. The regression estimates for the models that were run on the retrospective data differ only slightly from the ones obtained using prospective data. Jäckle (2008) used retrospective data and compared it to a gold standard (obtained from the “Improvement of Survey Measurement of Income and Employment” study in Jäckle et al. (2005)). The author found that ME in the reporting and dating of receipt of unemployment benefits attenuated both the duration dependence and the regression estimates from a Weibull model. The recall period used by the retrospective question was four months though, which is perhaps not long enough for the typical memory failures that characterize retrospective data over longer time periods.

One last study that used the more common recall period of one year was Pyy-Martikainen and Rendtel (2009). Here the authors compare data derived from a retrospective question on work status from the European Community Household Panel against a gold standard obtained from the Finnish register of unemployment. PH Cox and Weibull models for unordered repeated events were specified for the duration of unemployment and both attenuation and augmentation bias were found in the regression estimates. None of these biases changed the survey estimates by more than 30%, and they were found in the same direction and with a similar magnitude for both the Cox and Weibull models. Moreover, the comparison of the Cox and Weibull models shows that the baseline hazard was more accurately estimated by the former. The survey baseline hazard from the Weibull model is nearly constant while
the register baseline hazard shows positive duration dependence leading to erroneous conclusions about the duration dependence whereas the Cox baseline hazards from survey and register both display positive duration dependence.

In summary, it seems that when the response variable of EHA models, regardless of how it is defined (log duration, hazard rates, or person period cases), is affected by non-differential ME, the regression estimates of the model are attenuated. On the other hand, when the ME is associated with some of the explanatory variables, the direction of the bias in the estimates cannot be anticipated. Finally, because of the complexity of tracing the impact of ME in EHA models analytically, more empirical studies using validation datasets are necessary in order to assess both the peculiarities of retrospective ME and the consequences of these types of errors. Currently I am only aware of Jäckle (2008), and Pyy-Martikainen and Rendtel (2009).

4.2. Impact in Single Spell Models

I start the analysis that will be developed here and in Section 4.3 by simplifying the observation scheme considered in the survey and in the register. I use an important feature of the sample design, the fact that only subjects who were registered as unemployed on 28th February 1992 were contacted, to ensure that ME at the spell level (see Section 3.2) can be identified. Put it differently, the aim is to make sure that the spell of unemployment from the same subject captured by the survey and the register refer to the same event, which facilitates the definition of the ME term simply as the difference between PRESO and LSA durations.

This is achieved by setting the beginning of the window of observation at February 28th and considering only subjects who started from a state of unemployment in both the register and the survey. Such a strategy could be considered the most cautious approach to follow by researchers who only have access to survey data and who are aware of both the high prevalence of ME in the retrospective data and of the above mentioned peculiarity of the sample design. That is, it could be expected that in order
to reduce the impact of ME researchers would discard subjects who appeared to have misclassified their work status on 28th February.\textsuperscript{27}

Under this restriction, the sample that I use shares the structure seen in state-based sample designs (Holt, et al. 1991), where the sample frame is created out of individuals who are known to be in a particular state. The final sample size is 381 individuals (out of a total of 532 captured by both survey and register), and the window of observation encompasses spells from 28/2/92 to 30/03/93, where the ending date represents the earliest day interviews were taken. Right censoring is present in both datasets. The explanatory variables to be used in the models are age, experience, and their interaction term (see Sections 2.1 and 2.2 for a description of these variables).

I start by looking at the descriptive statistics of the first spells of unemployment. In this part of the analysis the sample used contains the same number of subjects and spells, in both the register and the survey datasets, i.e. 381. Figure 14 shows the Kaplan-Meier estimate (equation 1.30) of the survivor functions for the registered and reported time in unemployment. The two datasets show a similar path for the first 30 days; from that point until about day 100 the two measures diverge due to an accelerated failure rate in the survey; from then on the two survivor functions behave roughly similarly and the gap between them is maintained. At the end of the window of observation 35\% of the spells of unemployment in the register, 133 in total, were right-censored, whereas in the survey this was only 6\% of the sample, 23 in total.

\textit{Figure 14. Survivor function for the register and survey data}

\textsuperscript{27} For the analysis of the impact of ME in EHA models for single spells carried out in this section I use sample f, defined in Section 2.3.2.
To study the effect of ME on the durations of unemployment in more detail I present two more graphs grouped under Figure 15 below. The first is a scatterplot of LSA and PRESO durations which can be used to explore the effect of ME in each case. The second shows the probability density functions for the durations of LSA, PRESO, and their difference, which could be understood as the distribution of the error term.

Using these two plots we can identify three important features of the ME process. First, we can see how an important proportion of cases are relatively unaffected by ME. This is reflected by the points lying around the diagonal line of the scatterplot, where PRESO equals LSA. In particular, 162 subjects, 42% of the sample, reported durations within ±15 days of what was captured in PRESO. This pattern is also manifested by the green line depicting differences between PRESO and LSA in the second plot, which shows a majority of cases for which the observed ME is Normally distributed and centred around zero, as could be expected from classical ME (point 3 in equation 1.2). However, the same density function also shows a substantial part of the sample for which durations have been markedly underreported.

This overall underreporting of durations is the second ME feature that needs to be highlighted. In fact, by comparing the distributions of PRESO and LSA durations we can see how remarkably different they are. The durations of PRESO are concentrated at the end of the window of observation (most of them being right-censored), whereas the durations of LSA peak around values lower than 100 days. Specifically, the median durations in PRESO and LSA are 253 and 92 days, respectively. The magnitude of the underreporting can also be assessed by looking at the censored observations. Only 4 cases (1% of the sample) were wrongly reported to be right censored, whereas 115 (30% of the sample) were wrongly reported not to be right censored. That is, regarding censored durations, reports from LSA show a much higher prevalence of being FN than FP.
The third feature to notice from the plots in Figure 15 relates to the ME mechanism responsible for the overall underreport of durations, which seem to be unrelated to either the observed or true durations besides the fact of being underreported. For example, aside the process resembling classical ME observed in the distribution of the differences between PRESO and LSA durations (green line), we can observe that the rest of differences are roughly uniformly distributed across the window of observation. Similarly in the scatterplot are evenly scattered across the PRESO scale and relatively so in LSA too, although here they are less concentrated in durations higher than 300.

Finally, one last feature of the ME that cannot be inferred from the plots included in Figure 15 is its non-differentiality (point 6 from equation 1.2). Non-differentiality is one of the key assumptions of classical ME. For the typical case of “errors-in-variables” it refers to the independence between the ME and the response variable included in the model of interest. In the case explored here, where the response
variable is the one affected by ME, non-differentiality refers to the independence between the error term and the explanatory variables included in the model: age, work experience, and their interaction term (described in Section 2.1 and 2.2). To explore the assumption of non-differentiality I estimated the Pearson correlation between the ME - defined as the difference between PRESO and LSA - and both age and experience. These correlations were .07 and -.01, respectively, which show at least linear independence between the explanatory variables and the ME.

In short, the ME that I study here could be defined as: non-differential but highly complex due to the presence of two ME mechanisms that are acting simultaneously. Specifically, the complexity of the ME stems from the argument that certain cases appear to be affected by a classical ME process, while others are defined by a mechanism generating overall underreported durations that are unrelated to the true durations. Formally, this ME process can be defined as an extension of the classical ME model (equation 1.1) where the second ME mechanism is defined by a variable, $\Phi$, that is unrelated to the true durations, and bounded by 0 and 395 (the censoring time). This could be expressed as follows,

$$X^* = (1 - \pi)X + U + \pi \Phi$$  \hspace{1cm} (4.1)

where $\pi \in \{0,1\}$ is a parameter that determines whether the component $\Phi$ is present or not. From the different forms of ME reviewed in Section 3.1, $\Phi$, could be used to reflect FP spells of unemployment (case 3 in Figure 5) or as the linking of two different spells of unemployment due to the omission of a spell different from unemployment between them (case 2 in Figure 5).

### 4.2.1. Tools for the Assessment of the Impact of Measurement Error in Event History Analysis Models

In order to assess the impact of retrospective ME affecting the response variable in EHA models I use a research design similar to Jäckle (2008) and Pyy-Martikainen and Rendtel (2009). I specify EHA models using duration of spells of unemployment derived from the retrospective question and compare their estimates to the ones obtained by specifying the same type of models, for the same subjects, time frame, and explanatory variables, but using durations derived from the register. This register is assumed to be a gold standard; consequently differences in the estimates of the
models using survey data with respect to those obtained using register data are understood as evidence of the impact of ME.

For the sake of completeness the effect of ME on four different models will be analysed. These are: an AL Weibull and an AL exponential representing parametric models, a PH Cox from the semi-parametric models, and a PO logit representing non-parametric models. In addition, each of these models is analysed for single and repeated events. The latter are specified using marginal models, except for the PO logit, which is also explored using a random intercepts (RI) specification (see Section 1.2.4).

I use four different measures to assess the differences found in the regression estimates when the models are specified using the survey and the register data. The simplest of the four is the bias, calculated as the difference between the regression estimate obtained from the model using survey data and the same obtained using register data,

\[
BIAS = \hat{\beta}_s - \hat{\beta}_r
\]  

(4.2)

where \(s\) stands for survey and \(r\) for register. A second measure particularly useful for making comparisons between models and between explanatory variables using different scales is the relative bias,

\[
R.BIAS = \frac{|(\hat{\beta}_s - \hat{\beta}_r)100|}{|\beta_r|}
\]  

(4.3)

In order to take into account impacts on the precision of the estimates I also use the root mean squared error, which is the square root of the sum of the squared bias and the variance of the regression estimate obtained from the survey,

\[
RMSE(\hat{\beta}_s) = \sqrt{Var(\hat{\beta}_s) + (BIAS)^2}
\]  

(4.4)

Finally, in order to facilitate comparisons between models in terms of the RMSE, the relative root mean squared error will also be used,

\[
R.RMSE = \frac{\left|RMSE(\hat{\beta}_s) - RMSE(\hat{\beta}_r)\right|100}{RMSE(\hat{\beta}_r)}
\]  

(4.5)

where \(RMSE(\hat{\beta}_r)\) is just the standard error of \(\hat{\beta}_r\).
The impact of using this prone to error data in EHA models for single spells is analysed next. A separate subsection is included for each family of EHA models starting from the parametric family, and finishing with a summary of the relative performance of the different EHA models in the presence of ME.

### 4.2.2. Impact on the Accelerated Life-Time Weibull and Exponential Models

The results obtained when comparing the AL Weibull models using register and survey data are shown in Tables 11 and 12. In the model using the register data the main effects for both *age* and *experience* are negative and statistically significant, while their interaction effect is also significant but positive. So, considering the main effects of *age* and *experience*, the younger and less experienced the subjects the longer it takes to make a transition out of unemployment. However, this effect is nuanced by the positive interaction term, which indicates that the negative effect of *age* on the duration is attenuated the more experienced the subject is.

#### Table 11. Single event AL Weibull model using register and survey data*

<table>
<thead>
<tr>
<th></th>
<th>Regression Estimate</th>
<th>95% Confidence Interval</th>
<th>Standard Error</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Register</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>-.087</td>
<td>-.163-.012</td>
<td>.039</td>
</tr>
<tr>
<td>Experience</td>
<td>-1.38</td>
<td>-2.38-.38</td>
<td>.51</td>
</tr>
<tr>
<td>Age*Exp</td>
<td>.038</td>
<td>.010-.065</td>
<td>.014</td>
</tr>
<tr>
<td>Constant</td>
<td>9.08</td>
<td>6.34-11.81</td>
<td>1.40</td>
</tr>
<tr>
<td>α</td>
<td>.98</td>
<td>.88-1.10</td>
<td>.05</td>
</tr>
<tr>
<td>LR Chi² (3)</td>
<td>11.34</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Survey</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>.001</td>
<td>-.055-.058</td>
<td>.029</td>
</tr>
<tr>
<td>Experience</td>
<td>-.09</td>
<td>-.835-.645</td>
<td>.377</td>
</tr>
<tr>
<td>Age*Exp</td>
<td>.001</td>
<td>-.019-.022</td>
<td>.010</td>
</tr>
<tr>
<td>Constant</td>
<td>5.07</td>
<td>3.06-7.07</td>
<td>1.02</td>
</tr>
<tr>
<td>α</td>
<td>1.11</td>
<td>1.02-1.21</td>
<td>.05</td>
</tr>
<tr>
<td>LR Chi² (3)</td>
<td>.98</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Here and in the rest of the tables estimates in bold indicate that they were significantly different from zero at the 5% level.

Considering the impact of ME, the first result to note is the attenuation of all the regression estimates as a consequence of using survey data. It can be argued that attenuation bias represents the least bad type of bias since it only buffers the
estimated effect size, therefore leading to type II errors (Korn et al., 2010). However, the substantial size of the biases found here makes them non-negligible. Except for the constant term, all the regression estimates are not statistically significant when survey data is used. Furthermore, the estimate for age changes sign and becomes positive. Standard errors of the regression estimates have also been underestimated, although this is to be expected given the attenuation of the regression estimates, which now represent smaller effects.

Table 12 shows the four measures set out in the previous section to assess the impact of ME. The relative biases in the estimates of the explanatory variables age and experience are very large, 101.1% and 97.4%, respectively. These results indicate that the size of the bias is roughly the size of the true estimate. The interaction effect suffers from a similar effect, with a R.BIAS of 93.1%; only in the constant term, 44.2%, is the R.BIAS more moderate. The lower standard errors observed in the naïve (i.e. survey) model contributed to a reduction in the RMSE. However, the bias in the estimates is so large that the R.RMSE goes from 137.6% for the estimate of age to 196.6%. That is, the RMSE in the naïve model is at least twice as big as that found in the true model.

<table>
<thead>
<tr>
<th></th>
<th>BIAS</th>
<th>R.BIAS</th>
<th>RMSE</th>
<th>R.RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>.088</td>
<td>101.1%</td>
<td>.093</td>
<td>137.6%</td>
</tr>
<tr>
<td>Experience</td>
<td>1.29</td>
<td>93.1%</td>
<td>1.34</td>
<td>161.9%</td>
</tr>
<tr>
<td>Age*Exp</td>
<td>-.037</td>
<td>97.4%</td>
<td>.038</td>
<td>173.8%</td>
</tr>
<tr>
<td>Constant</td>
<td>-4.01</td>
<td>44.2%</td>
<td>4.14</td>
<td>196.6%</td>
</tr>
<tr>
<td>(\alpha)</td>
<td>.130</td>
<td>13.3%</td>
<td>.139</td>
<td>178.6%</td>
</tr>
</tbody>
</table>

The impact of ME on \(\alpha\), the parameter used in the Weibull model (see equations 1.33 and 1.34) to estimate the shape of the baseline hazard function, might seem relatively unimportant compared to what has been seen in the other estimates since the true estimate is .98 and the one found using survey data is 1.11. However, as Skinner and Humphreys (1999) point out, in some settings, there is interest not only in the size of this estimate but also in the distinction between \(\alpha<1\), \(\alpha=1\), and \(\alpha>1\), or equivalently between a decreasing, constant or increasing hazard function, respectively. For example, Chesher and Dumangane (2002) indicate that in the analysis of unemployment durations it is well known that uncontrolled across-individual
heterogeneity in hazard functions can lead to the appearance of negative duration dependence. In the case under analysis we observe a positive duration dependence when the model is specified using survey data, while the model using register data shows no duration dependence. Here the impact of ME differs from what we have seen for the rest of estimates, indicating a positive effect where there is none, which represents a type I error.

The graphs included in Figure 16 show the shapes of the baseline hazard functions for the register and the survey data. In spite of the different signs of the slopes, it is worth noting that the shape of the baseline hazard function from the survey data mimics quite well the one from the register data. To a certain extent this similarity could be due to the constraints of the Weibull model, where only one shape parameter is used, bounding the baseline hazard functions to either be monotonically increasing or decreasing.

*Figure 16. Weibull baseline hazard function for the register and survey data*

Another characteristic to note from the graphs in Figure 16 is the flatness of both hazard functions, which are almost constant across the window of observation. This feature suggests the possibility of using a simpler model to parameterize the baseline hazard function. In particular, the AL exponential appears to be a good alternative because it assumes a constant baseline hazard function.

A likelihood ratio test between the two models using register data (taking the exponential model to be nested in the Weibull) corroborates this intuition. The test shows that the difference in deviances (.13) for 1 degree of freedom is not statistically significant (p-value=.72). Results are shown in Tables 13 and 14.
In addition, the exponential model seems to perform marginally better at buffering the effects of ME; at least in terms of R.BIAS which is now lower for all the estimates. It is possible that parametric EHA models are more sensitive to ME in the response when the baseline hazard function is misspecified.

**Table 14. Bias in the single event AL Exponential model**

<table>
<thead>
<tr>
<th></th>
<th>BIAS</th>
<th>R.BIAS</th>
<th>RMSE</th>
<th>R.RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>.084</td>
<td>96.6%</td>
<td>.090</td>
<td>136.5%</td>
</tr>
<tr>
<td>Experience</td>
<td>1.25</td>
<td>91.7%</td>
<td>1.32</td>
<td>163.9%</td>
</tr>
<tr>
<td>Age*Exp</td>
<td>-.035</td>
<td>94.6%</td>
<td>.037</td>
<td>164.3%</td>
</tr>
<tr>
<td>Constant</td>
<td>-3.95</td>
<td>43.7%</td>
<td>4.12</td>
<td>200.9%</td>
</tr>
</tbody>
</table>

### 4.2.3. Impact on the Proportional Hazards Cox Models

I now turn to explore the effect of ME on the PH Cox model. Estimates from the PH Cox model are often presented on a hazard rate scale; however, here I present the untransformed estimates to facilitate comparisons between models.\(^{28}\) Tables 15 and 16 below show the results of the PH Cox model using the register and survey data.

---

\(^{28}\) That is I report \(\hat{\beta}_i\) instead of \(\exp(\hat{\beta}_i)\)
Results regarding the impact of ME on the PH Cox model show a very similar picture to what was found in the previous models. The regression estimates are again heavily attenuated. Interestingly, in terms of RMSE the PH Cox model performs similarly - or even slightly better - than the AL exponential and Weibull models. This is surprising given the higher precision expected from parametric models. This result suggests that, in spite of using an optimal parametric form for the true durations (from the register data), the baseline hazard function will probably change when there is ME, and less restrictive models such as the PH Cox are then a better choice.

The Cox baseline functions for the survey and register data are displayed in Figure 17. These are calculated using the ‘stcurve, hazard’ procedure in STATA 11, where the baseline hazard ratio at each transition is estimated first, and then kernel density estimation is used to smooth them. From the comparison of the two functions it can be seen that the former is overestimated, just like it was in the Weibull model. Also, now that the baseline function is freely estimated, I can confirm that the baseline function from the register data is truly constant, which corroborates the adequacy of the exponential model as opposed to the Weibull one. In addition, we can see that when using survey data there are a couple of bumps (at about day 220 and 330),
which could lead to slightly misleading time-dependence inferences. These shocks were not captured by the exponential nor the Weibull baseline functions for survey data because of their parametric restriction. However, since the Cox baseline function using survey data remains roughly constant when the Weibull one shows a positive slope, we might say that the former reflects the true function more faithfully.

**Figure 17. Cox baseline hazard function for the register and survey data**

So, when considering which model to use in the presence of ME in the response variable, I agree with Pyy-Martikainen and Rendtel (2009) in saying that the flexibility of the Cox model makes it a better choice than a parametric approach. The only exception to this would be where the true baseline hazard function can be properly approximated by a parametric form, as was shown for the case of the AL exponential model. In those cases the restrictive form of a parametric function could be beneficial. However, knowing the true baseline function conditional on a set of explanatory variables represents a major challenge, and it becomes even harder in the presence of ME.

### 4.2.4. Impact on the Proportional Odds Logit Model

Lastly, I present the effect of event-occurrence ME in the response variable on a model from the non-parametric family, a PO logit model. Here, a series of temporal dummies are included in the model in order to specify the baseline logit-hazard function. Each of the dummy variables represents a period of the time frame, in what is called a piecewise-constant hazards model. This is a reasonable solution when coarse time units relative to the window of observation are used. However, for the durations of unemployment analysed here this model raises some complications. First, the degrees of freedom are drastically reduced from the inclusion of 394
dummy variables, one for each day (except for the first one). Second, some of the days capture the same number of failures, which produces a problem of perfect multicollinearity in the model. In order to prevent these two problems I used temporal dummies that aggregate failures by weeks.

The results for the PO logit model are shown in Tables 17 and 18. The dummy variables representing the 56 weeks considered in the window of observation are not included in the tables for reasons of space, but they are shown in Figure 18 below as the dots composing the baseline hazard functions. In addition, the sample size is now 89,842 person-day cases in the register, and 50,366 in the survey.\(^{29}\)

### Table 17. Single event PO logit model using register and survey data

<table>
<thead>
<tr>
<th></th>
<th>Regression Estimate</th>
<th>95% Confidence Interval</th>
<th>Standard Error</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Register</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>.086</td>
<td>.012</td>
<td>.161</td>
</tr>
<tr>
<td>Experience</td>
<td>1.36</td>
<td>.38</td>
<td>2.35</td>
</tr>
<tr>
<td>Age*Exp</td>
<td>-.037</td>
<td>-.064</td>
<td>-.010</td>
</tr>
<tr>
<td>Constant</td>
<td>-9.41</td>
<td>-12.24</td>
<td>-6.57</td>
</tr>
<tr>
<td>LR Chi(^2) (61)</td>
<td>77.29</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Survey</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>.002</td>
<td>-.061</td>
<td>.065</td>
</tr>
<tr>
<td>Experience</td>
<td>.14</td>
<td>-.69</td>
<td>.97</td>
</tr>
<tr>
<td>Age*Exp</td>
<td>-.003</td>
<td>-.025</td>
<td>.020</td>
</tr>
<tr>
<td>Constant</td>
<td>-5.76</td>
<td>-8.09</td>
<td>-3.42</td>
</tr>
<tr>
<td>LR Chi(^2) (61)</td>
<td>161.33</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The outcomes of the two models are again very similar to what was found in the previous EHA specifications. In fact, the same standard errors as in the AL exponential and the PH Cox were obtained for age, experience, and the interaction effect when the register data is used.

### Table 18. Bias in the single event PO logit model model

<table>
<thead>
<tr>
<th></th>
<th>BIAS</th>
<th>R.BIAS</th>
<th>RMSE</th>
<th>R.RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>-.084</td>
<td>97.7%</td>
<td>.090</td>
<td>136.5%</td>
</tr>
<tr>
<td>Experience</td>
<td>-1.22</td>
<td>89.7%</td>
<td>1.29</td>
<td>157.0%</td>
</tr>
<tr>
<td>Age*Exp</td>
<td>.034</td>
<td>91.9%</td>
<td>.036</td>
<td>157.5%</td>
</tr>
<tr>
<td>Constant</td>
<td>3.65</td>
<td>38.8%</td>
<td>3.84</td>
<td>165.4%</td>
</tr>
</tbody>
</table>

\(^{29}\) The two datasets differ in their sample size because of the transformations required in the specification of EHA models for discrete data, which involved going from a dataset capturing one case for each subject to another capturing person-week cases.
Because of the effect of aggregating days into weeks, both baseline hazard functions differ from those obtained in the PH Cox models. The function for the register data is not constant. Also, unlike in the previous cases where the effect of ME was expressed as a higher baseline function, here what we see is more volatility between time-periods (weeks), resulting in a more jagged baseline function.

**Figure 18. PO baseline hazard function for the register and survey data*,**

*The hazard is measured in odds ratios.
**Values for the first week were omitted to prevent multicollinearity in the model.

4.2.5. Summary of the Impact on Models for Single Spells

In this section we have seen that the consequences of using retrospective data in EHA models for single spells are not negligible and are very similar across different models. Strong attenuation effects were found in all the regression estimates. In Table 19 these results are summarised by taking the mean R.BIAS and R.RMSE over the estimates of age, experience and their interaction term, for each of the four models studied.

**Table 19. EHA models’ performance in the presence of event-occurrence ME**

<table>
<thead>
<tr>
<th></th>
<th>R.BIAS</th>
<th>R.RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>AL Weibull</td>
<td>97.2%</td>
<td>157.8%</td>
</tr>
<tr>
<td>AL exponential</td>
<td>94.3%</td>
<td>154.9%</td>
</tr>
<tr>
<td>PH Cox</td>
<td>94.3%</td>
<td>152.7%</td>
</tr>
<tr>
<td>PO logit</td>
<td>93.1%</td>
<td>150.4%</td>
</tr>
</tbody>
</table>

In addition to the strong attenuation effects, illustrated by measures of R.BIAS not lower than 93%, it is striking to see the similarity of the effects across models. None

---

30 In order to make comparisons possible I excluded the constant term from this analysis since the PH Cox model does not estimate it.
of the models seems to buffer the effects of ME better than the others. In fact, the between model variability in terms of R.BIAS and R.RMSE is 1.8% and 3.1% respectively, while the average effect within models is 94.7% for the former and 153.9% for the latter.

The analysis presented so far has focused on assessing the impact of ME in each model separately. However, it could be argued that some EHA models are superior to others for the type of data used here. In order to assess which model performs better in the presence of ME they need to be compared against a common benchmark. That is, the use of a common reference allows us to analyse not only comparisons between the same models using error-free and prone-to-error data, but also comparisons between different models when prone-to-error data is used.

Here, I take results from the PH Cox model based on register data as that benchmark. There are both empirical and theoretical reasons for this choice. First, the PH Cox model has, along with the AL exponential and the PO logit, the lowest standard errors in their regression estimates (see Tables 13, 15, and 17). Second, since the baseline hazard function is freely estimated it cannot be misspecified. In addition, tied events, which can affect models in continuous time such as the Cox model (see Hosmer and Lemeshow, 1999, Chapter 3) are not a major issue here. The window of observation covers 395 days, which makes the time-unit approximately continuous, and rarely do two spells or more end on the same day. Finally, I checked that the proportional hazards assumption for the explanatory variables age and experience has not been violated in the model using register data by means of creating interactions of the two explanatory variables with time and comparing the Pearson product-moment correlation between the scaled Schoenfeld residuals and time. The Pearson correlations for age and experience were -.07 and -.08, with p-values of .263 and .226, respectively.

This process to assess the relative impact of ME on the different EHA models using survey data implies that the Cox model using register data produces the true estimates. Comparisons can be formally defined by equations 4.3 and 4.5, where \( \hat{\beta}_r \) is now substituted by \( \hat{\beta}_{r,\text{cox}} \).
Results are shown in Table 20, where we can see that the PO logit performs marginally better than the rest. It is also interesting to note that the AL Weibull offers the worst performance. These results reinforce the idea put forward when discussing the effect of ME in the baseline function: EHA models that do not make use of a restrictive parametric form seem to do better at buffering the effect of ME in the response variable, although the differences are very small. This seems to be especially true when the parametric form used is not the most appropriate, as shown by the worse performance by the Weibull than the exponential model.

I proceed with the analysis by looking at the implications derived from using retrospective data for the same EHA models when multiple spells are considered.

### 4.3. Impact in Multiple Spells Models

The consideration of repeated spells makes both the specification of the durations of unemployment, and the types of ME affecting them, more complex. Considering the three types of ME affecting retrospectively reported work histories discussed in Section 3.2 (miscounting, mismeasurement, and misclassification), and the setting used in the previous section, where every subject starts from the same category (unemployed) and only first spells are contemplated, the observed ME could only be due to mismeasurements of the duration of reported spells of unemployment. That is, the possibility that the survey data contain misclassified cases is ruled out since cases starting from a different status were discarded, while the effect of omitted or over-represented spells is non-existent since only first spells were considered.

However, when interest lies in modelling not only the first but also subsequent spells of unemployment, more complex types of ME affecting the reported durations should be expected. For example, the omission of spells of unemployment will reduce the sample size and probably bias results since, as we saw in Section 3.5.1,
shorter spells have a higher probability of being omitted. Similarly, we should expect results to be distorted from the inclusion of misclassified cases in the form of FP spells of unemployment.

In the analysis presented in the previous section both the register and survey datasets contained a sample size of 381 spells. Now, as a result of the inclusion of repeated events and the effect of miscounting and misclassification of spells, there are 559 spells of unemployment in the register, and 706 in the survey. This difference in the number of spells captured by the survey and register samples can be observed in the form of histograms in Figure 19.\textsuperscript{31} Notice that this Figure differs from Figure 11 since here the analysis is based on Sample g, which only considers subjects who had correctly reported to be unemployed in Feb. 1992.

\textit{Figure 19. Probability mass functions of the number of spells in the register and the survey}

The median durations of spells in the register and the survey are 180 and 89 days, respectively. These median durations are lower than the ones in the previous section, which implies that the additional spells included in this setting are shorter than the first ones.

Figure 20 displays the probability density functions for the spells of unemployment in both the survey and the register. The density function for the survey peaks for spells lasting around 50 days, whereas the register function has two peaks, one at day

\textsuperscript{31} For the analysis of the impact of ME in EHA models for repeated spells carried out in this section I use sample g, defined in Section 2.3.2.
50 and another where spells are right-censored\textsuperscript{32}. Similar shapes were found in Figure 15, where the probability density functions for the case of single spells was shown. However, it seems that now the register and survey functions have converged, as a result of which we might predict a lower effect of ME on the regression estimates. On the other hand, an analysis based solely on shapes of the probability density functions can be misleading, since these functions compare proportions, that is, they do not take into account the higher number of spells recorded in the survey.

\textit{Figure 20. Probability density function of repeated spells}

\begin{center}
\includegraphics[width=\textwidth]{Fig20.png}
\end{center}

In order to make a better graphical assessment of the impact of ME in EHA models for repeated events a scatter plot capturing the number of subjects unemployed at each period of the window of observation (Figure 21) can be used. The first part of the graph shows a similar picture to that of the survivor functions from Figure 14; the rates of unemployment for both the register and the survey fall in a similar pattern for the first 30 days, subsequently the survey shows a sharper decay, widening the difference between the survey and register rate more than in the single spells setting. In addition, after approximately day 90, when some of the first spells have failed and repeated spells are entering the study, the two functions remain relatively stable and they even show both growth and a degree of convergence.

\textsuperscript{32}This second bump at the end of the window of observation is however an artefact of having treated the right censored cases as if they were equal to 395. Had we observed the entire duration of such cases the bump would appear flattened.
The mild convergence between the two functions at the second half of the timespan can only be due to the higher number of cases re-entering unemployment in the survey, since as we saw before, the mean length of spells is shorter when considering repeated events, indicating that the first spells are longer. In Figure 22 survivor functions for the second spells found in the register and the survey are plotted. Here we observe an accelerated failure rate in the survey, similar to what was found in Figure 14 for the case of first spells. The survivor function for the register is based on 159 spells with a median length of 120 days, while the one from the survey had 238 spells with a median length of 89.5 days. Unlike the first spells, these start at different times during the window of observation and, because of that, censored cases are seen at different times for the survey and the register.
In what follows I assess the impact of using the survey dataset on different EHA models when repeated events are considered. In all of the following models it is assumed that the spells of unemployment were of the same type and unordered. That is, spells of unemployment were not treated differently because of their location in the window of observation or their position with respect to their order of appearance. To account for within subject dependencies robust standard errors are used. Specifically I use the sandwich estimator, which is the default option used in STATA version 11 under the procedure ‘cluster(id)’. This choice of model for repeated events replicates what has been used before in the literature on labour market studies (Gash, 2008; Pyy Martikainen and Rendtel, 2009). Finally, the impact of ME in the correlations of spells within subjects is briefly explored at the end by comparing random intercpets PO logit models. As before, I start with the parametric family of EHA models.

4.3.1. Impact on the Accelerated Life-Time Weibull and Exponential Models

The estimates from the Weibull model for the register data are now slightly different than for the single spells case; all the regression estimates remain significant, while age, experience, their interaction, and all the standard errors are now smaller (see Table 21). The lower effect size of the explanatory variables can be interpreted as the length of second and subsequent spells being more weakly associated with age, experience or their interaction effect, while the smaller standard errors represent the improvement in precision after the increase in sample size.
Table 21. Repeated events AL Weibull model using register and survey data

<table>
<thead>
<tr>
<th></th>
<th>Regression Estimate</th>
<th>95% Confidence Interval</th>
<th>Standard Error</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Register</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>-.084</td>
<td>-.150 -.017</td>
<td>.034</td>
</tr>
<tr>
<td>Experience</td>
<td>-1.22</td>
<td>-2.10 -.34</td>
<td>.45</td>
</tr>
<tr>
<td>Age*Exp</td>
<td>.033</td>
<td>.009 .058</td>
<td>.012</td>
</tr>
<tr>
<td>Constant</td>
<td>8.91</td>
<td>6.51 11.3</td>
<td>1.22</td>
</tr>
<tr>
<td>(\alpha)</td>
<td>1.01</td>
<td>.92 1.10</td>
<td>.04</td>
</tr>
<tr>
<td>LR Chi² (3)</td>
<td><strong>8.43</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Survey</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>-.005</td>
<td>-.054 .043</td>
<td>.025</td>
</tr>
<tr>
<td>Experience</td>
<td>-.15</td>
<td>-.77 .47</td>
<td>.32</td>
</tr>
<tr>
<td>Age*Exp</td>
<td>.003</td>
<td>-.015 .020</td>
<td>.009</td>
</tr>
<tr>
<td>Constant</td>
<td><strong>5.36</strong></td>
<td>3.66 7.05</td>
<td>.86</td>
</tr>
<tr>
<td>(\alpha)</td>
<td><strong>1.10</strong></td>
<td>1.01 1.19</td>
<td>.04</td>
</tr>
<tr>
<td>LR Chi² (3)</td>
<td>.78</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 22 summarizes the impact of ME in this model. Similar to the case of single spells all the estimates are attenuated, and they are not significant. The impact in terms of R.BIAS is slightly smaller than for the single spells model for all estimates, whereas in terms of R.RMSE the effect is lower for age and the constant, and bigger for experience and the interaction term. The similarity of these results when compared to the single spells model is rather unexpected, since in section 3.1.1, I anticipated that the forms of ME could be more complex in models accounting for multiple spells.

Table 22. Bias in the repeated events AL Weibull model

<table>
<thead>
<tr>
<th></th>
<th>BIAS</th>
<th>R.BIAS</th>
<th>RMSE</th>
<th>R.RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>.08</td>
<td>94.0%</td>
<td>.083</td>
<td>143.7%</td>
</tr>
<tr>
<td>Experience</td>
<td>1.07</td>
<td>87.9%</td>
<td>1.12</td>
<td>149.5%</td>
</tr>
<tr>
<td>Age*Exp</td>
<td>-.03</td>
<td>90.9%</td>
<td>.031</td>
<td>161.0%</td>
</tr>
<tr>
<td>Constant</td>
<td>-3.56</td>
<td>39.9%</td>
<td>3.66</td>
<td>199.3%</td>
</tr>
<tr>
<td>(\alpha)</td>
<td>-.090</td>
<td>8.2%</td>
<td>.088</td>
<td>119.1%</td>
</tr>
</tbody>
</table>

Finally, unlike what we saw in the previous section, the shape parameter is now significant and positive when survey data is used. However, its estimate is still very close to 1, which suggests once again the appropriateness of using an exponential
specification. Tables 23 and 24 show the results for the AL exponential model.33

**Table 23. Repeated events AL Exponential model using register and survey data**

<table>
<thead>
<tr>
<th></th>
<th>Regression Estimate</th>
<th>95% Confidence Interval</th>
<th>Standard Error</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Register</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>-0.084</td>
<td>-0.151 - 0.017</td>
<td>0.034</td>
</tr>
<tr>
<td>Experience</td>
<td>-1.22</td>
<td>-2.11 - 0.34</td>
<td>0.45</td>
</tr>
<tr>
<td>Age*Exp</td>
<td>0.034</td>
<td>0.009 - 0.058</td>
<td>0.012</td>
</tr>
<tr>
<td>Constant</td>
<td>8.93</td>
<td>6.53 - 11.33</td>
<td>1.23</td>
</tr>
<tr>
<td>LR Chi² (3)</td>
<td>8.40</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Survey</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>-0.006</td>
<td>-0.056 - 0.043</td>
<td>0.025</td>
</tr>
<tr>
<td>Experience</td>
<td>-0.16</td>
<td>-0.80 - 0.48</td>
<td>0.33</td>
</tr>
<tr>
<td>Age*Exp</td>
<td>0.003</td>
<td>-0.015 - 0.021</td>
<td>0.009</td>
</tr>
<tr>
<td>Constant</td>
<td>5.39</td>
<td>3.64 - 7.13</td>
<td>0.89</td>
</tr>
<tr>
<td>LR Chi² (3)</td>
<td>0.71</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Results from the exponential model are very similar to the ones found in the Weibull model, with the exponential model performing slightly better for all the estimates in terms of R.BIAS and R.RMSE, except for the interaction term. This result backs up the hypothesis that correctly specified models for the true data perform better in the presence of ME. In addition, compared to the model for single spells, the R.BIAS and R.RMSE show the milder impact seen for the Weibull model, supporting the intuition that the use of multiple spells does not necessarily imply a stronger impact of ME.

**Table 24. Bias in the repeated events AL Exponential model**

<table>
<thead>
<tr>
<th></th>
<th>BIAS</th>
<th>R.BIAS</th>
<th>RMSE</th>
<th>R.RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>.078</td>
<td>92.9%</td>
<td>.082</td>
<td>140.9%</td>
</tr>
<tr>
<td>Experience</td>
<td>1.07</td>
<td>87.0%</td>
<td>1.11</td>
<td>147.7%</td>
</tr>
<tr>
<td>Age*Exp</td>
<td>-.031</td>
<td>91.2%</td>
<td>.032</td>
<td>169.0%</td>
</tr>
<tr>
<td>Constant</td>
<td>-3.54</td>
<td>39.7%</td>
<td>3.65</td>
<td>198.0%</td>
</tr>
</tbody>
</table>

### 4.3.2. Impact on the Proportional Hazards Cox Model

Results for the PH Cox model are included in Tables 25 and 26. The estimates for both the model using register and survey data are again very similar to what has been seen in this section for the case of the AL Weibull and exponential models.

33The use of the exponential model is justified by a likelihood ratio test (p-value=.86) used to compare it against the Weibull specification.
Table 25. Repeated events PH Cox model using register and survey data

<table>
<thead>
<tr>
<th></th>
<th>Regression Estimate</th>
<th>95% Confidence Interval</th>
<th>Standard Error</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Register</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>.083</td>
<td>.017 - .149</td>
<td>.034</td>
</tr>
<tr>
<td>Experience</td>
<td>1.21</td>
<td>.34 - 2.08</td>
<td>.45</td>
</tr>
<tr>
<td>Age*Exp</td>
<td>-.033</td>
<td>-.057 - -.009</td>
<td>.012</td>
</tr>
<tr>
<td>LR Chi² (3)</td>
<td>8.32</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Survey</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>.007</td>
<td>-.044 - .059</td>
<td>.026</td>
</tr>
<tr>
<td>Experience</td>
<td>.17</td>
<td>-.48 - .84</td>
<td>.34</td>
</tr>
<tr>
<td>Age*Exp</td>
<td>-.003</td>
<td>-.022 - .015</td>
<td>.009</td>
</tr>
<tr>
<td>LR Chi² (3)</td>
<td>.76</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The impact both in terms of R.BIAS and R.RMSE is smaller than in the AL Weibull and exponential models from this section, reinforcing the idea that freely estimated EHA models perform better at buffering the effect of ME. In addition, compared to the PH Cox model for single spells, we can see here the same milder impact observed in the previous AL Weibull and exponential models for repeated events.

Table 26. Bias in the repeated events PH Cox model

<table>
<thead>
<tr>
<th></th>
<th>BIAS</th>
<th>R.BIAS</th>
<th>RMSE</th>
<th>R.RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>-.076</td>
<td>91.6%</td>
<td>.080</td>
<td>136.2%</td>
</tr>
<tr>
<td>Experience</td>
<td>-1.03</td>
<td>85.5%</td>
<td>1.09</td>
<td>144.1%</td>
</tr>
<tr>
<td>Age*Exp</td>
<td>.030</td>
<td>90.9%</td>
<td>.031</td>
<td>161.0%</td>
</tr>
</tbody>
</table>

4.3.3. Impact on the Proportional Odds Logit Model

In the estimation of the PO logit for the repeated events 110,563 and 82,599 person-day cases were used in the register and survey models respectively. Results for the PO logit are presented in Tables 27 and 28. Regression estimates are similar to the ones found in the rest of the models from this section.
Table 27. Repeated events PO logit model using register and survey data

<table>
<thead>
<tr>
<th></th>
<th>Regression Estimate</th>
<th>95% Confidence Interval</th>
<th>Standard Error</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Register</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>.083</td>
<td>.016</td>
<td>.150</td>
</tr>
<tr>
<td>Experience</td>
<td>1.21</td>
<td>.32</td>
<td>2.09</td>
</tr>
<tr>
<td>Age*Exp</td>
<td>-.033</td>
<td>-.057</td>
<td>-.009</td>
</tr>
<tr>
<td>Constant</td>
<td>-9.26</td>
<td>-11.84</td>
<td>-6.68</td>
</tr>
<tr>
<td>LR Chi$^2$ (61)</td>
<td>88.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Survey</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>.005</td>
<td>-.047</td>
<td>.056</td>
</tr>
<tr>
<td>Experience</td>
<td>.15</td>
<td>-.50</td>
<td>.81</td>
</tr>
<tr>
<td>Age*Exp</td>
<td>-.002</td>
<td>-.021</td>
<td>.017</td>
</tr>
<tr>
<td>Constant</td>
<td>-6.13</td>
<td>-8.07</td>
<td>-4.19</td>
</tr>
<tr>
<td>LR Chi$^2$ (61)</td>
<td>207.7</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Unlike what we saw in the models for single spells, the impact of ME in terms of R.BIAS is now slightly larger than in any of the other models for repeated events. However, this is not the case when the impact is measured in terms of R.RMSE, which is very similar to what we have seen in the previous parametric models for repeated events.

Table 28. Bias in the repeated events PO logit model

<table>
<thead>
<tr>
<th></th>
<th>BIAS</th>
<th>R.BIAS</th>
<th>RMSE</th>
<th>R.RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>-.078</td>
<td>94.0%</td>
<td>.085</td>
<td>141.8%</td>
</tr>
<tr>
<td>Experience</td>
<td>-1.05</td>
<td>87.4%</td>
<td>1.10</td>
<td>145.1%</td>
</tr>
<tr>
<td>Age*Exp</td>
<td>.031</td>
<td>93.9%</td>
<td>.032</td>
<td>169.0%</td>
</tr>
<tr>
<td>Constant</td>
<td>3.13</td>
<td>33.8%</td>
<td>3.28</td>
<td>149.3%</td>
</tr>
</tbody>
</table>

4.3.4. Impact on the Random Intercepts Proportional Odds Logit Model

Here, the effect of event-occurrence ME in a random effects model is explored. To study the impact of ME on estimates such as the between subjects variance (denoted in Tables 29 and 30 as Var(RI), an acronym for variance of the random intercepts term), and to compare the impact in the regression estimates with that observed in the previous PO logit model, I will use a RI PO logit model (defined in Section 1.2.4). This model is estimated using Markov chain Monte Carlo, and unlike in the
previous PO logit model estimated in STATA, here I will use MLwiN\(^34\). In order to assure that convergence was achieved and to reduce simulation error I used 5,000 iterations after burning-in another set of 5,000 iterations. Results are presented in Table 29 and 30 below.

**Table 29. Random intercepts PO logit model using register and survey data**

<table>
<thead>
<tr>
<th></th>
<th>Regression Estimate</th>
<th>95% Credible Interval</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Register</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>.068</td>
<td>.036</td>
<td>.100</td>
</tr>
<tr>
<td>Experience</td>
<td>1.05</td>
<td>.62</td>
<td>1.43</td>
</tr>
<tr>
<td>Age*Exp</td>
<td>-.029</td>
<td>-.039</td>
<td>-.017</td>
</tr>
<tr>
<td>Constant</td>
<td>-8.81</td>
<td>-10.12</td>
<td>-7.41</td>
</tr>
<tr>
<td>Var(RI)</td>
<td>.083</td>
<td>.002</td>
<td>.280</td>
</tr>
<tr>
<td>DIC</td>
<td>4281.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Survey</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>.010</td>
<td>-.010</td>
<td>.036</td>
</tr>
<tr>
<td>Experience</td>
<td>.23</td>
<td>-.049</td>
<td>.552</td>
</tr>
<tr>
<td>Age*Exp</td>
<td>-.004</td>
<td>-.014</td>
<td>.003</td>
</tr>
<tr>
<td>Constant</td>
<td>-6.40</td>
<td>-7.45</td>
<td>-5.48</td>
</tr>
<tr>
<td>Var(RI)</td>
<td>.017</td>
<td>.001</td>
<td>.093</td>
</tr>
<tr>
<td>DIC</td>
<td>6308.7</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Compared to the previous PO logit model for repeated events all the estimates from the fixed part of the model are lower when using register data, but higher for survey data. This smaller attenuation effect is manifested in the measures of R.BIAS which are now lower than in any of the previous models.

Another distinctive result of using random effects as opposed to robust standard errors to model the within subject’s episodes dependency is that the standard errors, when using both register and survey data, are now substantially smaller, about half the size than before. This higher precision is more pronounced than the reduction of bias in relative terms, which makes the impact in terms of R.RMSE about twice the size than before, with the only exception of the RI variance. For this estimate, the reduction of the standard errors is so large (70.1%) that it makes the impact of ME in terms of R.RMSE almost negligible.

\(^34\) Although the model was called from the STATA interface using the command runmlwin developed by Leckie and Charlton (2013).
### Table 30. Bias in the random intercepts PO logit model

<table>
<thead>
<tr>
<th></th>
<th>BIAS</th>
<th>R.BIAS</th>
<th>RMSE</th>
<th>R.RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>-.058</td>
<td>85.3%</td>
<td>.059</td>
<td>230.2%</td>
</tr>
<tr>
<td>Experience</td>
<td>-.826</td>
<td>78.3%</td>
<td>.842</td>
<td>282.5%</td>
</tr>
<tr>
<td>Age*Exp</td>
<td>.025</td>
<td>86.2%</td>
<td>.025</td>
<td>322.0%</td>
</tr>
<tr>
<td>Constant</td>
<td>2.41</td>
<td>27.4%</td>
<td>2.46</td>
<td>278.1%</td>
</tr>
<tr>
<td>Var(RI)</td>
<td>-.066</td>
<td>79.5%</td>
<td>.070</td>
<td>9.2%</td>
</tr>
</tbody>
</table>

The variance of the RI term is significant in both models but it is much smaller in the one using survey data. Precisely, in terms of R.BIAS, there is an attenuation of 79.5%. This result indicates that the between subjects unobserved heterogeneity shrinks in the presence of event-occurrence ME. Similarly, it could be argued that the impact of ME makes work histories experienced by different subjects less distinguishable.

#### 4.3.5. Summary of the Impact on Models for Repeated Spells

As was the case for models using single spells, the estimates for age, experience and their interaction effect were significant in all the models using register data and became non-significant when the survey data is used. Very similar attenuation effects to those seen in models for single spells have been shown. However, contrary to the hypothesis that more complex forms of ME would be expected as a result of considering repeated events, we have seen that the average R.BIAS is lower than before, in each of the models analysed.

Similarly, the impact of ME in terms of R.RMSE for models considering repeated events is not higher than for the single spells case. The only exception in this respect is the RI PO logit model, which shows almost twice the impact than any other model. This is shown in Table 31, where the average results for age, experience and their interaction term are compared. Regarding the comparison of models for repeated events is interesting to note that the PH Cox comes the first and second in terms of R.RMSE and R.BIAS, respectively. The lowest impact in terms of R.BIAS comes from the RI PO logit. However, this might be due to the smaller effects found when register data is used, which leaves less room for attenuation towards the null.
In order to properly assess not only the impact of ME in each of the models, but which one performs better in the presence of ME, the comparisons need to be made against a common benchmark. For that I use the PH Cox model based on the register data as in Section 4.2.5. This is the model less prone to problems of misspecification, and as we just saw the model that suffers the least when both R.BIAS and R.RMSE are contemplated. Results from these comparisons are presented in the following Table 32.

Table 32. EHA models’ performance compared to the PH Cox

<table>
<thead>
<tr>
<th></th>
<th>R.BIAS</th>
<th>R.RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>AL Weibull</td>
<td>90.88%</td>
<td>151.77%</td>
</tr>
<tr>
<td>AL exponential</td>
<td>90.18%</td>
<td>153.28%</td>
</tr>
<tr>
<td>PO logit</td>
<td>91.78%</td>
<td>161.48%</td>
</tr>
<tr>
<td>RI PO logit</td>
<td>85.63%</td>
<td>91.50%</td>
</tr>
</tbody>
</table>

Here, we see that the three models using robust SEs perform very similarly in terms of both R.BIAS and R.RMSE, which reinforces the finding from the previous section regarding the similarity of the impact of ME across models. That pattern is broken for the RI PO logit model, which on the one hand shows lower attenuation when survey data is used than the same PH Cox model, and on the other shows higher precision than any other model using survey data, which results in lower R.RMSE.

4.4. Discussion of the Implications of Measurement Error in Event History Analysis

In this chapter I have explored the implications of using EHA models where the response variable is affected by ME derived from an event-occurrence type of retrospective question. Evidence of large attenuation biases in the regression
estimates is found across different EHA models. These findings go against the common belief that ME in the response variable only affects the standard errors of the model’s estimates, and also contrast with what has been seen so far in the literature.

In particular, my results contrast with those of Skinner and Humphreys (1999) where, after simulating classical multiplicative errors, no bias was found in the regression estimates. The difference between these and the results presented here stems from the sequential component embedded in retrospective questions on work histories. The type of errors simulated by the authors can be used to replicate the ME processes found in simpler retrospective questions, e.g. number of spells of unemployment experienced in the last twelve months, or number of sexual partners in a lifetime. However, we have seen that when a sequential component is required – in the case analysed here, when spells need to be dated - the ME generating mechanisms might not be appropriately specified by the classical multiplicative model.

The findings obtained here also disagree with Pyy-Martikainen and Rendtel (2009), which is the most similar study in the literature since retrospectively reported spells of unemployment are also compared to data from a register of unemployment. The authors found both attenuation and augmentation biases affecting the regression estimates. Arguably, the mix of biases that they found might be related to the ME being associated with some of the explanatory variables. The ME analysed here was non-differential with respect to the two regressors that were used (age and experience), and the direction of the biases was always towards the null. Moreover, these results are consistent with all the other studies that I am aware of that have assessed the impact of non-differential ME in the response in EHA: Augustin (1999), Dumangane (2007), Korn et al. (2010), Magder and Hughes (1997), Meier et al. (2003), and Neuhaus (1999).

Another substantive difference between the results presented here and those from Pyy-Martikainen and Rendtel (2009) is the bigger size of the biases found here. In Table 22 I showed that the average R.BIAS in the regressors of the Weibull model for multiple events was 90.9%, whereas the biggest bias found in Pyy-Martikainen and Rendtel (2009) for the same model was 30% of the true estimate. These differences might be due to both the use of months as time-units in Pyy-Martikainen
and Rendtel (2009), which limit the appearance of ME, and to the much bigger sample size both in terms of individuals (1,482) and window of observations (responses to five years were pooled), which reduced the share of cases that were right-censored.

One original feature of the study presented here is the assessment of different families of EHA models. Very similar results were found for the four types of EHA models that were studied (AL Weibull, AL exponential, PH Cox and PO logit), which implies that the way the response variable capturing lifecourse events is defined (duration data, hazard rates, or person-period cases) is not related to the effect of ME on the model estimates. In fact, all the models performed similarly for the different comparisons carried out: when true data is used, when estimates of true data are compared against the ones obtained using survey data, when the Cox model is used as a benchmark, and for both the case of single and repeated spells. Perhaps the PO logit model showed the biggest differences. This might be due to the inclusion of temporal dummies which were discretized to capture weeks instead of days.

Using the Cox model as a benchmark I found that the exponential model is less affected by ME than the Weibull both in terms of R.BIAS and R.RMSE and for both the case of single and multiple spells. It seems that parametric forms that are correctly specified for the true data might buffer the effect of ME better than when they are misspecified. Moreover, semi and non-parametric models perform similarly to the exponential model, even in terms of R.RMSE. This is an interesting result since parametric models, when correctly specified, are expected to obtain more precise estimates. Moreover, ascertaining the shape of the baseline hazard function is complicated, and in general, it could be expected that parametric forms will perform worse than shown here. Hence, when the shape of the baseline hazard function is unknown, as is the case in settings that use durations measured with errors, the use of semi- and non-parametric forms should be recommended.

Similarly, here I have found that inferences about the time-dependency of the event derived from the PH Cox or the PO logit model are less misleading than those obtained from the AL Weibull model, which indicated wrongly that the probability of making a transition out of unemployment increased with time. This result corroborates Pyy-Martikainen and Rendtel (2009), where the authors posited that freely estimated baseline functions offer better results than those which impose a
parametric form. An exception to this precept might be cases where the parametric form perfectly maps the form of the baseline function. This is what has been observed here for the case of the AL exponential. However, in most cases, previous knowledge about the shape of the baseline function conditional on a set of regressors is not available, let alone when the durations are affected by ME. Hence, for the estimation of time-dependencies in the event of duration data prone to ME the use of semi-or non-parametric models can be recommended.

Comparisons of the effects of ME for single and repeated events specifications showed an average lower attenuation when multiple spells are considered. This result contrasts with my a priori expectations. I anticipated that more complex forms of event-occurrence ME would be found when repeated events are considered, which would make the impact of ME stronger than in the case for single spells.

Finally, I explored the impact of ME in hierarchical models using a RI PO logit model. Here, the variance of the RI term was as strongly attenuated as all the other regression estimates, indicating that the unobserved heterogeneity between work histories was diluted in the presence of the event-occurrence ME. However, taking the estimates of the PH Cox model using register data as a benchmark I found that the impact of ME is lower for the RI PO logit than for any other model, both in terms of R.BIAS and R.RMSE for all estimates.

In this study I have used data derived from a retrospective question on work histories for a period of 395 days, yet the findings obtained here could be generalized to other cases where retrospective data is used to derive different lifecourse events. In particular, this would be the case for events that, because of their relatively low saliency, can be subject to recall errors in the form of mismeasuring, miscounting, and misclassification of spells in the same way as spells of unemployment are.

However, more research is necessary since some of the findings presented here need to be tested. In particular, I would like to extend this study to cases where the ME affecting the response variable in EHA is associated with the explanatory variables. Non-differential ME can generate biases that are not necessarily towards the null, but it is not clear what the levels of association are that could cause a change in the direction of the bias. Another setting of interest would be the extensions of the
models seen here to the case of competing risks (Prentice et al, 1978). This would allow contemplating the impact of retrospective data in EHA in greater detail.
CHAPTER 5. ADJUSTMENT OF EVENT-OCCURRENCE MEASUREMENT ERROR

The last 30 years have seen a remarkable development of methods for the adjustment of ME. Initially limited to simplistic settings of classical ME affecting the only explanatory variable in a simple regression model (Fuller, 1987), the range of adjustments available nowadays covers problems of systematic errors in models for cluster data (Zidek et al., 1998) or differential errors in causal analysis (Imai and Yamamoto, 2010). However, as has been noted by different authors (Augustin, 1999, Jäckle, 2008, Pyy-Martikainen and Rendtel, 2009), their application to event history models remains relatively understudied.

Furthermore, all the methods designed to adjust for ME from retrospective questions in event history analysis that I am aware of have used simulated data. That is, they have been applied to studies where the specification of the errors and the degree to which they affect the true values was hypothesised (Holt et al., 1991, Augustin, 1999, Skinner and Humphreys, 1999, Cole et al. 2006, and Dumangane, 2007). As pointed out in Section 4.1 these studies reflect the type of ME that could be expected to arise from retrospective questions enquiring about the duration of one particular spell. However, they fail to reflect the type of ME arising from questions about the order and durations of a series of spells (what I have called here event-occurrence ME).

In this chapter I assess and compare the effectiveness of a selection of methods for the adjustment of ME when the response variable in different event history models is affected by event-occurrence ME. But, unlike what has been done in the past, here I use real data from survey responses that have been linked to register data on unemployment, which is assumed to be a gold standard. In this way, the

35 “Little is however known about the nature of errors in event history data from panel surveys or their effects on estimates, let alone about ways of mitigating these” (Jäckle, 2008, p. 2), “In contrast to its practical importance, ME has not yet attracted much attention in duration analysis.” (Augustin, 1999, p. 2), “Despite the recognition of the existence of measurement errors in survey-based data on event histories, little is known about their effects on an event history analysis.” (Pyy-Martikainen and Rendtel, 2009, p. 140).
effectiveness of different adjustment methods can be explored in a way that is not reliant on an appropriate simulation of errors.

5.1. Methods for the Adjustments of Measurement Error

A great variety of methods for the adjustment of ME have been developed in different disciplines. The taxonomy of these methods has traditionally been based upon two major defining characteristics: the extent to which the methods rely on assumptions about the distributional specification of the error model (classical additive, multiplicative etc.), and the extent to which additional data is available to explore the structure of the ME model (Carroll et al., 2006 and Freedman, 2008).

Regarding the first feature, a majority of methods are built on the hypothesis that the distribution of the ME is classical (see Section 1.1.1). Other methods, such as those based on likelihood adjustments, are more flexible and can produce effective adjustments for errors that are distributed differently, so long as the ME process is correctly specified. This first differentiating feature is linked to the second one; the extent to which the methods rely on additional sources of data to perform the adjustments.

As a rule of thumb, we can say that methods that do not rely on the correct specification of the error model tend to rely more heavily on exploiting additional information of some kind. This information could be: i) a measure that is associated with the true values but is not related to the ME so it can be used as an instrumental variable (Wright, 1928, and Theil, 1953); ii) alternative measures of the same concept that could be used in data reduction techniques such as factor analysis (Harman, 1960); iii) identical replicate measures so the common variability of the replicated measures can be differentiated from the ME variability\(^{36}\); iv) a subsample of validation data so that some of the observations of the mismeasured variable can be compared with their true values in order to detect both systematic and random components of the ME.

\(^{36}\) See Dunn (2004) for a review of adjustment methods that can be used in the presence of instrumental variables, alternative or replicated measurements are available.
Finally, a third feature that is normally used in the literature (Fuller, 1987, Carroll et al., 2006, and Buonaccorsi, 2010) differentiates between methods that require a specification for the true variable, known as functional methods, and those where such a requirement is not necessary, known as structural methods. More formally, functional methods are those in which the true values are seen as a sequence of unknown fixed constants, whereas structural methods they are considered random values arising from a probability function that needs to be specified.

With this classification in mind I explore the effectiveness of various ME adjustment methods. The target of these adjustments is to reduce the differences observed between the vector of regression estimates from EHA models obtained using the survey data, $\hat{\beta}_s$, and those obtained using register data, $\hat{\beta}_r$. Specifically, to compare the relative effectiveness amongst the adjustments that are successful (those that are demonstrated to be at least partially beneficial), I use the ratio of the R.BIAS found in the naïve analysis over the R.BIAS obtained after the adjustment, denoted as R.R.BIAS, or more formally as,

$$R.R.BIAS = \frac{R.BIAS_{adj}}{R.BIAS_{naive}}$$  \hspace{1cm} (5.1)

The denominator of that ratio was defined in equation 4.3; the R.BIAS from the adjustments ($R.BIAS_{adj}$) is obtained using the same expression but substituting the regression estimates obtained using the survey data, $\hat{\beta}_s$, by the estimates obtained from the adjustment, $\hat{\beta}_{adj}$. Hence, the R.R.BIAS, can be interpreted as the effectiveness of an adjustment in reducing the bias produced by ME. In addition, to assess the effectiveness of the adjustment while taking into account the impact on the SEs I use the R.R.RMSE, a similar ratio between the naïve and adjusted models but using the R.RMSE statistics, defined in equation 4.5.

From the family of functional methods I study the implementation of simulation-extrapolation (SIMEX), regression calibration (RC) and multiple imputation (MI). The first of these methods assumes classical ME, while the other two make fewer assumptions about the ME process but rely on validation subsamples. The functioning, implementation and results from SIMEX are presented in Section 5.2, while RC and MI are presented together in Section 5.3. From the family of structural
methods I implement different likelihood-based methods using a Bayesian approach, which are presented in Section 5.4.

5.2. SIMEX

SIMEX was first presented by Cook and Stefanski (1994) and refined in the following years by Stefanski and Cook (1995) and Carroll et al. (1996). From the different methods that have been developed to adjust for problems derived from errors in variables, SIMEX is one of the simpler to implement and less reliant on additional assumptions or additional variables in the form of validation or replicated data. In addition, SIMEX can be extended to outcome models with increasing levels of complexity, and it does not require the distribution of the true values to be specified (it is a functional model). However, SIMEX also has some weaknesses: it is only applicable to simple cases of ME, and an estimate of the variance of the ME is needed.

“The key idea underlying SIMEX is the fact that the effect of measurement error on an estimator can be determined experimentally via simulation” (Carroll et al., 2006, p. 98). SIMEX exploits the relationship between the size of the ME affecting one variable and the size of the bias in the regression estimates in the model of interest. For the application implemented in the following subsection it is important to note that SIMEX can be used for cases of ME affecting either the response or an explanatory variable in a model of interest, so long as increasing levels of ME affecting that variable are positively associated with the impact (the bias) in the regression estimates to be adjusted.

In order to unveil the specific pattern of association, SIMEX simulates new variables that reproduce the observed variable plus increasing levels of the type of ME to which it was originally affected. The outcome model is rerun a number of times using these new variables affected by increasing levels of ME. New biased estimates from the outcome model can be retrieved for each of those times the model is rerun, and by pairing the biased estimates to the levels of ME used, a function can be defined. Using this function we can estimate the true parameter of interest that the
estimator from the naïve model fails to converge to. This is done by extrapolating the function to a value where the level of ME is zero.

The target in this chapter is to adjust for error in the response variable when estimating the parameters of an event history model. However, to facilitate the understanding of SIMEX I proceed now to review the different steps involved in its implementation using a simple example of bias in the slope of a simple linear regression as a result of the only explanatory variable being affected by classical ME.

5.2.1. SIMEX for Measurement Error in the Explanatory Variable of a Simple Regression Model

In Section 1.1.3, the outcome model was defined (in equation 1.13) as follows,

\[ Y = \beta_0 + \beta_1 X^* + \epsilon \] (5.2)

where the only explanatory variable, \( X^* \), is prone to classical additive error (equations 1.1 and 1.2) and \( Y \) is measured without error.

In this case we know (Fuller, 1987) that the unadjusted estimator of the slope, \( \hat{\beta}_1 \), does not converge asymptotically to the parameter \( \beta_1 \) but to:

\[ \hat{\beta}_1^* = \beta_1 \left( \frac{\sigma_X^2}{\sigma_X^2 + \sigma_\epsilon^2} \right) \] (5.3)

where \( \sigma_X^2 \) and \( \sigma_\epsilon^2 \) represent the variance of the true explanatory variable and the error term. In other words, the estimator of the slope is biased downwards in absolute terms by the reliability ratio, \( \rho \) (defined in equation 1.5), of the observed variable, \( X^* \). In this situation, and if \( \rho \) or \( \sigma_\epsilon^2 \) is known, it would not be practical or efficient to use SIMEX, since the adjustment would simply be achieved by substituting the variance terms in the equation above. However, I will use this simple setting for illustrative purposes.

The implementation of SIMEX is divided into six phases:

1) The first step involves simulating additional variables with increasing levels of ME. These new variables are generated in a way that emulates the classical ME model, but with successively larger values of \( \sigma_\epsilon^2 \) affecting \( X \). Specifically, \( K \) new explanatory variables \( X^*_k(\lambda_k) \) are generated by the rule:
\[ X_k^*(\lambda_k) = X^* + \sqrt{(\lambda_k)}U \] (5.4)

with \( k = 0,1, ..., K \), the simulated error, \( U \), is Normally distributed with mean 0 and variance \( \sigma_U^2 \), and \( \lambda_0 < \lambda_1 < \ldots < \lambda_K \) are a set of parameters that are used to amplify the ME variance (often these are (.5, 1, 1.5, 2)).

2) Once the different variables with added ME have been generated, the outcome model is re-estimated using this new data, and the values of the estimator of interest (i.e. \( \beta_1 \)) for the different levels of ME (\( \lambda_k \)) are saved. In particular, for the case of a simple linear model with the explanatory variable affected by classical ME, and the data generating rule described in equation 5.4, the estimator of the slope will now take the following form,

\[ \hat{\beta}_{1k}^* = \beta_1 \sigma_X^2 / (\sigma_X^2 + (1 + \lambda_k) \sigma_U^2) \] (5.5)

where the bias increases monotonically as \( \lambda_k \) increases.

3) In order to reduce the Monte Carlo error associated with simulations steps 1 and 2 are repeated \( B \) times so a mean estimate of \( \hat{\beta}_{1b}^* \) for \( b = 1, ..., B \) can be computed,

\[ \bar{\hat{\beta}}_{1k}^* = \frac{\sum_{b=1}^{B} \hat{\beta}_{1kb}^*}{B} \] (5.6)

where the rule of thumb\(^{37}\) is to use \( B = 100 \) iterations.

4) At this stage the \( \bar{\hat{\beta}}_{1k}^* \) and \( \lambda_k \) values can be paired considering the former as a function of the latter, \( G \left( \bar{\hat{\beta}}_{1k}^*, \lambda_k \right) \), known as the extrapolation function, which should be plotted in order to obtain a first insight of its shape.

5) The extrapolation function is estimated using a regression model, with data \( \left( \bar{\hat{\beta}}_{1k}^*, \lambda_k \right) \). Carroll et al. (2006) recommend the use of one of three types of simple functional forms\(^{38}\).

a) linear,

\[ G \left( \bar{\hat{\beta}}_{1k}^*, \lambda_k \right) = \zeta_1 + \zeta_2 \lambda_k \] (5.7)

b) quadratic,

\[ G \left( \bar{\hat{\beta}}_{1k}^*, \lambda_k \right) = \zeta_1 + \zeta_2 \lambda_k + \zeta_3 \lambda_k^2 \] (5.8)

\(^{37}\)This is the number of iterations used by default in the SIMEX packages in STATA and R.

\(^{38}\)From the different functional forms that have been suggested, Carroll et al. (2006) indicate that the quadratic function often produces the most conservative corrections, but they are also the least variable and more robust to misspecifications.
c) non-linear or ratio-linear, \[ G \left( \tilde{\beta}_{1k}, \lambda_k \right) = \zeta_1 + \zeta_2 / (\zeta_3 + \lambda_k) \] (5.9)

For our example, and if the extrapolation function is well approximated by the chosen functional form, we would find the following function,

\[ E \left( \tilde{\beta}_{1k} | \lambda_k \right) = G \left( \tilde{\beta}_{1k}, \lambda_k \right) = \beta_1 \sigma^2_X / (\sigma^2_X + (1 + \lambda_k)\sigma^2_U) \] (5.10)

6) From here, the SIMEX estimator, \( \hat{\beta}_{SIMEX} \), can be calculated by extrapolating \( G \left( \tilde{\beta}_{1k}, \lambda_k \right) \) to \( G \left( \tilde{\beta}_{1k}, \lambda_k = -1 \right) \). Note that from equation 5.5 when \( \lambda_k = -1 \) the bias is cancelled out.

I have used the case of ME affecting the predictor in a simple linear regression for heuristic reasons, but the method works equally well for other outcome models regardless of how complex they are as long as a monotonically increasing function derived from increasing levels of ME can be associated with increasing levels of bias in the estimate that needs to be adjusted. In those situations, the same logic applies since the \( \tilde{\beta}_k^* \) are calculated from measurement having variance \( (1 + \lambda_k)\sigma^2_U \), and for \( \lambda_k = -1 \) we have \( (1 + \lambda_k)\sigma^2_U = 0 \).

The SIMEX method can also be represented graphically. This is done in Figure 23, where the solid line denotes the part of the extrapolation function that can be approximately observed through the regression estimates resulting after the outcome model is specified using simulated predictors with increasing levels of ME, and the dashed line represents the extrapolation to the case of no ME, which gives the adjusted estimate.

Figure 23 also shows some of the limitations of SIMEX. The entire extrapolation function cannot be observed; hence it is hard to assess the quality of the adjustment. In addition, the extrapolation function needs to be approximated using a simple functional form. Therefore adjustments are only approximated, and their effectiveness depends on how well the extrapolation function is estimated, for which the choice of the functional form is crucial. In the case depicted by Figure 23 it seems clear that a quadratic function is the best approximation, but in other instances the choice might not be so clear. In those cases, a likelihood ratio test comparing different functional forms could offer a more formal solution.
This extrapolation step is commonly criticised by researchers not for the problem related to finding an appropriate extrapolation function - mentioned above in step 5 - but as a result of sceptical views towards extrapolating functions to unknown values. In general, predictions based on values of the explanatory variables outside the range of observations used in the model need to be well justified. It needs to be accepted that the same data-generating processes defined in the model are operating for the observations that are extrapolated. This is often not the case in Social Sciences where relationships between variables often reach thresholds after which the nature of those relationships changes. For example, height increases with age until a saturation point is achieved at about 21 (for men), so, if a model had been estimated only using observations from the first five years of life, predictions to older ages would overestimate height. However, in SIMEX, the extrapolation to \( \lambda_k = -1 \) is reasonable because it still lies within the data generating mechanism modelled by the extrapolation function (equation 5.10). Specifically, the error-generating mechanism considered in SIMEX encompasses values of \( \sigma_0^2 \in [0, \infty) \), with \( \sigma_0^2 = 0 \) when \( \lambda_k = -1 \).

Another cause of concern stems from the accuracy of the estimate of \( \sigma_0^2 \) that is used in the simulations. For example, considering the case depicted in Figure 23, if \( \sigma_0^2 \) is underestimated, the extrapolation function will have a lower slope and the adjustment would only be partial. That is, for an underestimated \( \sigma_0^2 \), lower values of \( \hat{\beta}_k^* \) would have been generated for \( (1 + \lambda_k) = (1.5, 2, 2.5, 3) \), which would have made the estimated extrapolation function shallower, and the extrapolation step to
$\lambda_k = -1$ would have generated a bigger adjusted estimate, one which would still be upwardly biased. Such suboptimal adjustment is illustrated in Figure 24 where I compare the extrapolation function shown in Figure 23 with a similar one that would be obtained if $\sigma_U^2$ had been underestimated.

**Figure 24. Comparison of extrapolation functions**

The process set out here has focused on retrieving adjusted regression estimates, or $\hat{\beta}_{adj}$. For the estimation of the SEs of those estimates different methods with varying levels of complexity can be used. “The ease with which estimates can be obtained via SIMEX, even for very complicated and nonstandard models, is offset somewhat by the complexity of the resulting estimates, making the calculation of SE difficult” (Carroll et al., 2006, p. 392). Three methods can be used for the estimation of the standard errors of the adjusted regression estimates.

Stefanski and Cook (1995) suggest a method to obtain the asymptotic distribution of the SIMEX estimator that, just like the SIMEX algorithm, splits into an analysis of the simulation and extrapolation steps. This method is only approximate in the sense that it is generally valid in large samples with small ME. Carroll et al. (1996) also suggest using a method based on the sandwich estimator and on the theory of M-estimators$^{39}$ to obtain an asymptotic covariance estimator. In particular, this method is based on the asymptotic equivalence of $\hat{\beta}(\lambda_k)$ and an M-estimator, producing a

---

$^{39}$ M-estimator is a term due to Huber (1964) that refers to the type of robust estimators that are obtained as the minima of sums of functions of the data. For example the method of least squares is a type of M-estimators since the estimator is defined as a minimum of the sum of squares of the residuals.
closed form equation from which the SEs can be directly derived, which avoids the process of replicating the SIMEX procedure with all its limitations seen above. On the other hand the Carroll et al. (1996) method is algebraically complicated and requires additional programming.

These two methods have been developed for the specific case of classical additive ME with known (or at least estimated) \( \sigma^2 \). New procedures need to be developed when the method is to be extended to different ME processes, or when SIMEX is used as a sensitivity analysis procedure where different levels of \( \sigma^2 \) are tried. In these circumstances, standard resampling methods such as bootstrap or jackknife become a natural strategy to analyse the variability of the adjusted estimates, \( \hat{\beta}_{adj} \). These are non-parametric methods that can be applied regardless of the ME mechanism that is considered or the accuracy with which \( \sigma^2 \) is estimated. The bootstrap is described in more detail in Appendix D.

The main problem though is time: jackknife and bootstrap\(^{40}\) are computationally intensive techniques and so is SIMEX, so their combination can sometimes become too demanding in terms of clock time, depending on the sample size and on the complexity of both the outcome and the SIMEX model. Nonetheless, given my interest in using additional ME models other than the classical additive, the bootstrap was used in all the SIMEX applications carried out here.

5.2.2. Implementation of SIMEX

Some examples of studies applying SIMEX are: Stefanski and Cook (1995), who used SIMEX in a logistic regression where coronary heart disease was regressed on systolic blood pressure; Hardin et al. (2003), who implemented SIMEX in a logistic regression where the appearance of breast cancer was modelled by a set of explanatory variables, including recalls for individual saturated and caloric fat intake; He et al. (2007) applied SIMEX to an accelerated failure Weibull model with classical errors in explanatory variables (cholesterol level and systolic blood

\(^{40}\) Jackknife is the method used by default in the SIMEX package in R, while bootstrap is the one used by STATA.
pressure); and Battauz et al. (2008), who adjusted an ordinal probit model with ME affecting one explanatory variable capturing educational achievement.

However, in spite of its simplicity and applicability the method has not been as widely used as other more common methods for dealing with ME (likelihood-based methods, instrumental variables or regression calibration). This could be due to limitations in its original formulation where it was only applicable to problems of classical additive ME. However, new extensions have been developed that allow the application of SIMEX to different ME models. “SIMEX is ideally suited to problems with additive measurement error, and more generally to any problem in which the measurement error generating process can be imitated on a computer via Monte Carlo methods” (Carroll et al., 2006, p. 97).

In principle, SIMEX can be extended to any situation where cumulative levels of ME measured by increasing variance of the ME term produce a monotonically increasing bias in one of the estimates, and as long as the ME mechanism can be replicated accurately. Some extensions that have proven to be robust are: i) SIMEX for classical multiplicative ME (Carroll et al., 2006, and Biewen et al. 2008); ii) MC-SIMEX (Kuchenhoff et al. 2006), or the application of the SIMEX methodology to problems of misclassification of either the response or an explanatory variable in the outcome model; and iii) the Berkson-SIMEX adjustment (Althubaiti and Donev, 2011), which can be used to adjust the regression estimate of a simple linear model with Berkson Gamma distributed errors in the explanatory variable.

In this study, as well as exploring the performance of the standard SIMEX method for classical additive ME, adjustments assuming classical multiplicative errors are also carried out. Thus I assume that \( Y^* = YV \), where \( V \) is distributed according to a log-normal distribution with mean one. As was discussed in Section 1.1.1, the classical multiplicative model might offer a more appropriate representation of the ME observed in the LSA sample. Furthermore, the classical multiplicative ME model has often been suggested in the literature as an appropriate specification to model ME arising from retrospective questions (Holt et al., 1991, Skinner and Humphreys, 1999, Augustin, 1999, and Dumangane, 2007). However, the

---

41 Battauz, et al. (2008) is the only article that I am aware of that used SIMEX within the social sciences.
effectiveness of adjustments using such a ME specification has not been assessed using validation data.

In order to adapt the original SIMEX to classical multiplicative ME Carroll et al. (2006) propose a change in the way the simulated variables that are increasingly affected by ME, $X^*_k$, are generated in step 1. In particular, equation 5.4 is substituted by

$$Y^*_k(\lambda_k) = \exp\{\log(Y^*) + \sqrt{\lambda_k}\log(V)\}$$  \hspace{1cm} (5.11)

to represent the multiplicative relationship between the observed durations and the simulated noise, while the rest of the six-step algorithm is implemented as before.

The correction process works just as for the classical additive model. Providing the observed durations are actually affected by classical multiplicative ME, increasing levels of simulated noise using equation 5.11 will result in more extreme biases from which an extrapolation function can be specified.

Below I report the results obtained after using both classical additive and multiplicative SIMEX for the problem of a PH Cox model for single spells with its response variable being affected by event-occurrence ME (the naïve model using survey data presented in Section 4.2.3). In both adjustments, the simulations of increasing levels of noise affecting the observed durations, $Y^*(\lambda_k)$, were bounded at 0 and 395 to constrain them to the window of observation used in LSA and PRESO.\footnote{For the assessment of the effectiveness of SIMEX, RC&MI, and the Bayesian adjustments carried out in this section, in 5.3, and in 5.4, I use sample f, defined in Section 2.3.2.}

The variance of the error term, $\sigma^2_v$, required for the SIMEX adjustment, can be estimated as the difference between the variances from LSA and PRESO durations $\sigma^2_v = \sigma^2_{v^*} - \sigma^2_{v}$, assuming classical additive ME, and as the exponential of that difference in logs, $\sigma^2_v = e^{\log(\sigma^2_{v^*}) - \log(\sigma^2_{v})}$, if we assume classical multiplicative ME. However, due to censoring in both LSA and PRESO, the variance of the error term cannot be estimated without resorting to parametric assumptions about the distribution of the censored cases. Instead I will carry out various adjustments assuming different values of $\sigma^2_v$. For the classical additive adjustment these are: 400, 1600 and 6400, which correspond to standard deviations of 20, 40 and 80. The
variances used for the classical multiplicative adjustment are: .01, .04, and .12, or .10, .20, and .35 in terms of standard deviations.

**Standard SIMEX**

I start by presenting the results obtained for the classical additive model (the code used to program in R the SIMEX process assuming classical additive ME has been included in Appendix E). Figure 25 - with \( Y^* \) on the x-axis and \( Y^*(\lambda_1 = 1) \) on the y-axis - illustrates the effect of the second set of simulations \( (\lambda_1 = 1) \) in increasing the level of ME found in the observed LSA durations, \( Y^* \), for the three levels of \( \sigma_w \) used. We can see that as the standard deviation of the simulated error increases the correlation between \( Y^* \) and \( Y^*(\lambda_1 = 1) \) becomes smaller.

*Figure 25. Classical additive simulations for \( \lambda_1 = 1 \) compared to LSA durations*

To compare the distribution of the simulated durations (for the three levels of \( \sigma_w \)) against the shape of the distribution of the true and observed durations I plot them all together in Figure 26. SIMEX is based on the requirement of being able to reproduce the ME generating mechanism with increasing levels of intensity \( (Y^*(\lambda_1 = 1) \) and \( Y^*(\lambda_1 = 2) \), which are designated as “noise.1” and “noise.2” in the legend in Figure 26). So, ideally in Figure 26, the green and blue lines (the simulated durations)
should be able to replicate and augment the original ME process that shifted the distribution of true durations (black line, PRESO) to that of the observed durations (red line, LSA). That is, we would expect an even lower proportion of right censored cases and a higher proportion of durations around 50 days, where LSA peaked.

This is not the case here, however. In fact, we can see that in the different scenarios studied the effect of simulated classical additive ME is to reduce the concentration of durations smaller than 100 days observed in the LSA data, which becomes more evident for the higher levels of $\sigma_p$.

*Figure 26. Observed and true durations compared to the first two sets of classical additive simulations*

*The black line denotes $Y$, the red line $Y^*$, and the green and blue line for the first and last level of simulated durations, $Y^*(\lambda_1 = 1)$ and $Y^*(\lambda_2 = 2)$, respectively.

The reason why standard SIMEX failed to reproduce the ME mechanisms observed in LSA (compared to PRESO) stems from the assumption that the ME is classical, whereas in Section 4.2 we saw that two different ME mechanisms were likely to be acting simultaneously in the report of event-occurrence spells of unemployment. Furthermore, from these two ME mechanisms, the systematic underreport of long
durations surely has a bigger impact in biasing regression estimates given the much bigger effect it has in distorting the true durations.

The failure to reproduce accurately the ME mechanism results in the incapacity to trace the extrapolation function that links growing levels of simulated ME with increased bias in the regression estimates. Specifically, the classical additive SIMEX process implemented here has not generated the expected monotonic pattern for values of age, experience and their interaction term across the levels of \( \lambda_k \). This is shown in Figure 27 where I plotted the average values (over 1000 iterations) obtained for the regression estimates of age (the figures for experience and the interaction term are shown in Appendix F) when rerunning both the PH Cox model for the different levels of \( \lambda_k \) (0.5, 1, 1.5, 2) and the naive estimate obtained when using the LSA durations (\( \lambda = 0 \)).

Since none of the three extrapolation functions recommended offer an adequate fit the adjustment was carried out using the simplest of them all, the linear function. This is represented by the red line, where the black circle at \( \lambda = -1 \) indicates the adjusted estimate of age, experience, and the interaction term. It is important to underline that the problem observed here (across the three levels of \( \sigma_u \) that I used) is the generation of a non-monotonic pattern between the added levels of ME and the bias of the estimates. This is different from the problem shown in Figure 24, which pointed at the adjustment being ineffective as a result of having either under or overestimated the variance of the error term. Arguably the problem of non-monotonicity seen here might stem from the failure of the simulated errors to reproduce the original ME seen in LSA (Figure 26).
As becomes clear from Figure 27 the standard SIMEX adjustment assuming classical additive ME has failed. To assess the degree of ineffectiveness of the adjustment I report in Table 31 the adjusted estimates of *age*, *experience* and their interaction term, for the three levels of \( \sigma_v \) used, together with the true estimates (from PRESO), and those obtained from the naïve analysis using LSA durations. Here we can see that, although some estimates were marginally adjusted, most were not, increasing the bias seen in the naïve analysis. Interestingly, similarly unpromising results can be observed for each of the three levels of \( \sigma_v \), which reinforces the view that the use of a particular value of \( \sigma_v \) might not be as important as other major problems that render the SIMEX method inadequate in the context under study.

**Table 33. Regression estimates using the standard SIMEX adjustment**

<table>
<thead>
<tr>
<th></th>
<th>PRESO</th>
<th>LSA</th>
<th>SIMEX (( \sigma_v = 20 ))</th>
<th>SIMEX (( \sigma_v = 40 ))</th>
<th>SIMEX (( \sigma_v = 80 ))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>.086 (.038)</td>
<td>.002 (.032)</td>
<td>-.017 (.036)</td>
<td>.013 (.046)</td>
<td>-.062 (.067)</td>
</tr>
<tr>
<td>Experience</td>
<td>1.35 (.50)</td>
<td>.13 (.42)</td>
<td>.15 (.035)</td>
<td>.10 (.051)</td>
<td>.19 (.071)</td>
</tr>
<tr>
<td>Age*Exp</td>
<td>-.037 (.014)</td>
<td>-.002 (.012)</td>
<td>-.052 (.031)</td>
<td>.092 (.048)</td>
<td>-.056 (.070)</td>
</tr>
</tbody>
</table>

*SEs are included within brackets
To estimate the SEs for the adjustment a bootstrap process with 100 samples (see Appendix D) was used. Figure 28 shows the sampling distributions estimated using bootstrap for the three estimates, from which I calculated their SEs, included within brackets in Table 31. Compared with the SEs obtained after using the PRESO and LSA durations we can see that the adjusted SEs are higher for the estimates of age and the interaction term, and that they increase for higher levels of $\sigma_u$.

*Figure 28. Bootstrapped sampling distributions for the regression estimates*

![Sampling Distributions](image)

**SIMEX Classical Multiplicative ME**

To extend the study of SIMEX, I assess the effectiveness of the method when classical multiplicative ME is assumed. Figure 29 plots the mismeasured durations from LSA and those obtained after simulating classical multiplicative ME for different levels of $\sigma_u$. Now the effect of the ME increases proportionally (rather than linearly as in Figure 25) with the size of the mismeasured durations. This effect can turn observed durations into extreme durations of length 395. However, short durations are relatively unaffected in the ME simulations.
In Figure 30 I compare the distribution of the simulated durations for levels of $Y^*(\lambda_1 = 1)$ and $Y^*(\lambda_2 = 2)$ with the distributions from LSA and PRESO for the three levels of $\sigma_u$. Here we can see that the simulated durations under the classical multiplicative model did not manage to replicate the original ME problem observed when comparing LSA to PRESO. In particular, we can see that the effect of increased levels of multiplicative noise on LSA is almost imperceptible at the distribution level. Contrary to what we observed in Figure 26 the concentration of durations around 100 days is not substantially reduced after the classical multiplicative ME is simulated, which is probably due to the previously mentioned lesser effect of multiplicative ME on the shorter durations.
Figure 30. LSA and PRESO compared to the first two sets of classical multiplicative simulations*

The inadequacy of the adjustment assuming classical multiplicative ME is also manifested by the non-monotonic pattern observed in the extrapolation functions for the estimates of age, experience and their interaction term. These extrapolation functions are shown in Figure 31, while the specific results from the adjustments in comparison with the true and naïve model and the standard SIMEX adjustment are reported in Table 32.

*The black line denotes $Y$, the red line $Y^*$, and the green and blue line for the first and last level of simulated durations, $Y^*(\lambda_1 = 1)$ and $Y^*(\lambda_2 = 2)$, respectively.
As it was the case for the standard SIMEX adjustment, the lack of a monotonic relationship between the increased levels of ME and the bias in each regression estimate renders the adjustment ineffective. As before, we can see a partial adjustment for some estimates while others are further biased.

Table 34. Regression estimates and standard errors using the classical multiplicative SIMEX adjustment

<table>
<thead>
<tr>
<th></th>
<th>PRESO</th>
<th>LSA</th>
<th>SIMEX ($\sigma_v = .10$)</th>
<th>SIMEX ($\sigma_v = .20$)</th>
<th>SIMEX ($\sigma_v = .35$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>.086 (.038)</td>
<td>.002 (.032)</td>
<td>-.022 (.028)</td>
<td>.016 (.042)</td>
<td>.011 (.061)</td>
</tr>
<tr>
<td>Experience</td>
<td>1.35 (.50)</td>
<td>.13 (.42)</td>
<td>.107 (.028)</td>
<td>.005 (.042)</td>
<td>.211 (.059)</td>
</tr>
<tr>
<td>Age*Exp</td>
<td>-.037 (.014)</td>
<td>-.002 (.012)</td>
<td>.033 (.032)</td>
<td>.071 (.037)</td>
<td>.048 (.056)</td>
</tr>
</tbody>
</table>

*SEs are included within brackets

5.2.3. Conclusions from SIMEX

In conclusion, SIMEX does not seem an appropriate method for the adjustment of the type of ME found in event-occurrence questions on work histories. The
effectiveness of SIMEX relies on the use of an accurate estimate of the variance of the ME, which cannot be easily obtained for duration data that is right censored but also given the peculiarities of the type of ME seen here. More specifically, SIMEX has failed to produce adequate adjustments because I have not been able to reproduce the ME mechanisms affecting the observed durations of unemployment using classical additive or multiplicative ME models. Some imprecisions in the adjustment could be expected given the limitations of the method (e.g. the need to approximate the real extrapolation function) or because of the peculiarities of longitudinal data that contains right censored cases. However, not being able to reproduce the ME mechanisms represents a central problem of the SIMEX process. As a result, the adjustments were ineffective for both the classical multiplicative and classical additive models.

It is possible that SIMEX might work under a more appropriate ME model, for example, we could explore adding a systematic component to the classical additive ME model to imitate the shorter durations observed in the LSA sample. However, given the complexity of the ME processes (see Section 4.2) I am sceptical about the ability of SIMEX to produce any substantial adjustments. In addition to discourage the use of SIMEX for complex ME problems, I believe that the analysis carried out here can also be used to rule out the use of other methods that rely on the assumptions of classical ME. In the next section I implement adjustment methods that do not require the specification of a ME model; the regression calibration (RC) and the multiple imputation (MI) methods.

5.3. Regression Calibration and Multiple Imputation

RC was simultaneously proposed around two decades ago by Carroll and Stefanski (1990) and Glesjer (1990), and since then it has become one of the most commonly used methods for the adjustment of ME (Freedman, 2008). One of the principal advantages of RC is its simplicity. In a nutshell, the RC adjustment only involves replacing the mismeasured variable, $X^*$, by the best approximation, $\hat{X}$, that can be obtained using any informative data available.
Unlike other methods for the adjustment of ME such as SIMEX or any likelihood-based methods, RC does not require a specification of the distribution of the ME. Instead, RC relies on access to a replicated or validation subsample of the mismeasured variable. When replicated data is available the mean of the repeated observations for a particular variable $\bar{X}^*$ is taken as the true value $X$ under the assumption that $U$ is classical. When validation data is available the true variable $X$ can be observed for a subsample, which eliminates the need to approximate $X$ assuming classical ME. In particular, validation data allows adjustments for systematic ME.

Here I develop RC from the perspective of adjusting for ME in the response variable, $Y$. The mismeasured observations from the original sample are now defined as $Y_i^*$, for $i \in I = \{1, 2, \ldots, N\}$, the validated observations are defined by $Y_s$, where $s \in J = \{1, 2, \ldots, S\}$ for a subset $J \subseteq I$, that is, the validation subsample is contained within the original sample.

The RC process consists of three steps. First, for the subsample of observations $J$ that are validated a calibration model is specified where the true variable $Y_s$ is regressed on the mismeasured observations $Y_s^*$, for $s \in J$, and on a set of potentially informative auxiliary data, giving a predictive model $\hat{Y} = \hat{\zeta}_{RC,Y}Y^* + \hat{\zeta}_{RC,Z}Z$. By auxiliary data, I mean any variables available covering at least some elements of $i$ (i.e., the sample units considered in that outcome model) that can be used to inform a calibration - or imputation - model, and which I here denote by $Z$. Normally a standard linear model is used to specify the calibration model although this does not need to be so; in the following Section 5.3.1 different alternatives will be explored.

Second, the calibration model is used to predict the true values that were not covered by the validation subsample. These cases that are not validated are denoted as $Y_{s'}$, where $s' = 1, 2, \ldots, N - S$, and $s' \in i$. So, using the regression estimates of the calibration model, defined as $\hat{\zeta}_{RC}$, and extrapolating for cases in $s'$, we can obtain an estimation of $Y_{s'}$, denoted by $\hat{Y}_{s'}$, derived from the conditional expectation of $E(Y|Y^*,Z)$, which is used to replace the mismeasured observations $Y_{s'}^*$.

The adjusted parameter estimates for the model $f(Y; X, \beta)$ are obtained when the model is estimated using a new variable, $Y_i'$, consisting of validated observations, $Y_i$. 

156
for $i \in J$, i.e. when they are available, and of predicted cases, $\hat{\boldsymbol{Y}}_i$, for cases $i \in I \setminus J$ where validated observations are not accessible.

The process explained above focuses on the adjustment of the estimates of $\beta$. To adjust for any bias found in the standard errors of those estimates an additional procedure is needed to account for the fact that $\hat{\boldsymbol{Y}}_{si}$ has been predicted based on an estimated RC model in the previous phase.

One option is to use the bootstrap, which is computationally intensive. As an alternative, Carroll and Stefanski (1990) derived a general formula based on the M-estimator to obtain asymptotic standard errors.\footnote{This is now available in STATA.} Their formula can be applied for the case of generalized linear models adapting the specific link function that is used in the outcome model. However, to use this formula in my study would require a modification to accommodate censoring. More importantly, the formula relies on an estimate of the variance of the ME term, which is straightforward to estimate for replicated data but not for validation data. For these reasons, I use the bootstrap to estimate the standard errors from the RC adjustment.

**Multiple Imputation**

RC was developed in 1990 by biostatisticians but it has been only in the last few years that the method has been used in disciplines other than the biomedical sciences. Partly this is due to the popularity of an alternative - multiple imputation (MI) - a technique first developed by Rubin (1987) that is nowadays used extensively by survey statisticians to adjust for problems of missing data. The phenomenon of missing data shares many similarities with ME, and RC like MI relies on imputing data.

In spite of having being designed as a technique for the adjustment of missing data, MI can also be used to adjust for ME by treating it as a missing data problem. We only need to think of the ME problem as if “the information with more desirable measurement properties is not available for the entire sample” (Peytchev, 2012, p. 5). After all, the rationale of MI is essentially the same as in RC: obtaining estimates of the missing - or mismeasured - cases based on information obtained from auxiliary data. That is MI is based on the missing at random (MAR) assumption.
(Rubin, 1976) and it can be argued that RC implies a similar assumption. MAR indicates that the probability of a case being missing can be explained using auxiliary data. The ME equivalent might be expressed referring to failures of precision and accuracy in a variable prone to ME being accounted for using auxiliary data. Notice that MAR - unlike the assumption of missing not at random (MNAR) - assumes that the probability of an observation being missing does not depend on the value of the unobserved variable itself.

This similarity between the processes of missing data and ME has resulted in a growing number of studies using MI to adjust for ME problems in the last ten years (Cole et al., 2006; Durrant and Skinner, 2006; Freedman et al., 2008; Peytchev, 2012), and to assess its relative effectiveness compared to RC I have also carried out an adjustment using MI in my analysis.

Although the core idea of RC and imputation is similar, three important methodological differences set RC and MI apart: i) MI does not necessarily rely on the existence of a validation or replicated subsample; ii) the predicted variables - or more appropriately, the imputed variables - $\hat{Y}$ in MI, are not predicted as extrapolations from point predictions based on a linear model, but based on probabilistic imputation from conditional distributions; iii) MI does not rely on a second step such as the bootstrap or the M-estimator to appropriately reflect the uncertainty derived from using $\hat{Y}$ in the outcome model. Instead it generates $K$ different imputations for each missing case so the outcome model can be run $K$ times and the variance between results from each imputation can be used to reflect the uncertainty of using imputed cases.

In multiple imputation $K$ completed datasets are generated, where the imputed values, $\hat{Y}_k$, vary across the data sets due to the probabilistic imputation model. The variability reflects that there is uncertainty in the imputation. Because datasets with different imputed values are generated, the parameter estimates from each of the outcome models $f(\hat{Y};X,\beta)$ will differ, something which propagates the uncertainty derived from the imputation process. This additional form of error, known as the between imputation error, is the key element of MI, which can be calculated in the pooling stage, the last phase of the process.
In the pooling stage results from the $K$ outcome models are combined in order to yield a final result that captures both the original uncertainty reflected in the outcome model and that derived from using $\hat{Y}$. More formally, the parameter of interest, $\beta$, is estimated as the mean of the estimates obtained for each imputation, so

$$\bar{\hat{\beta}} = \frac{1}{K} \sum_{k=1}^{K} \hat{\beta}_k$$

(5.12)

while the variance of $\bar{\hat{\beta}}$ can be calculated as the sum of two variance components: i) the within imputation variance, calculated as the average standard error of $\hat{\beta}_k$ across the $K$ datasets, and ii) the between imputation variance, calculated as the variance of the values of $\hat{\beta}_k$ between the multiple imputed data sets. Formally, this procedure can be expressed as,

$$\text{var} \left( \bar{\hat{\beta}} \right) = U + B + \frac{B}{K}$$

(5.13)

where $U$ denotes the within imputation variance, or

$$U = \frac{1}{K} \sum_{k=1}^{K} \text{var} (\hat{\beta}_k)$$

(5.14)

and $B$ represents the between imputation variance,

$$B = \frac{1}{K - 1} \sum_{k=1}^{K} (\hat{\beta}_k - \bar{\hat{\beta}})^2$$

(5.15)

and $B/K$ is included to recognise that $\bar{\hat{\beta}}$ itself is estimated using a finite number of $K$ datasets.

### 5.3.1. Implementation of Regression Calibration and Multiple Imputation

Two authors have examined the adequacy of RC for the adjustment of the bias found in parameters from a PH Cox model – the outcome model of interest here. Wang et al. (1997) demonstrates its effectiveness in a problem of classical ME in explanatory variables as did Spiegelman et al. (2011) after applying RC to PH Cox and logit
models when the ME is not necessarily classical. In particular, Spiegelman et al. (2011) show that, unless the effect of ME is severe (reliability ratio < .5), standard RC approximations will typically be adequate, even with moderate heteroscedasticity in the measurement error model variance. However, I am unaware of researchers having used RC to adjust for ME affecting the response variable in the PH Cox model. This is probably due to the belief that classical ME in the response variable does not bias the parameter estimates of the outcome model. However, as we saw in Section 4.2, some features of event-occurrence ME affecting durations of unemployment, such as its systematic part and the presence of censoring can have an impact, making it a cause for concern.

In addition to the possibility of being adapted to different outcome models, standard RC can also be modified to adjust for ME terms that are not linearly related to the true values. For example, Peytchev (2012) uses a logit calibration model to adjust for a false negatives problem for survey participants who reported not to have had an abortion when they really had.

Given the distinctive features of the distributions of variables capturing durations (left-truncated at zero and right-skewed) compared to the more common Normally distributed variables, we can anticipate that the choice of the functional form for the calibration model will be a relevant factor in determining the quality of the adjustment. The other major factor is the precision with which the parameters of the calibration model are estimated, which in turn is determined by the relative size of the subsample of validation data and by how informative the rest of the available data is in predicting the ME process.

Here, I assume that the researcher has access to a random selection of the original sample of 381 observations where the durations from the PRESO register can be observed. To explore how the effectiveness of the method relies on the size of the validation subsample different proportions for $S$ will be used. In addition, I explore three different functional forms for the calibration model: a standard linear, an accelerated life (AL) exponential, and a logit calibration model. All of these calibration models use the observed durations, $Y^*$, and a dichotomous variable capturing whether those observed durations were censored, $C^*$, as explanatory
variables\textsuperscript{44}, where $Y^* = 395$ when censored. The latter is included to inform the imputation model that censored observations are different than non-censored ones – they are longer.

The first of the functional forms used to specify the calibration is the standard linear model, which in my analysis can be expressed as,

$$Y = \zeta_0 + \zeta_1 Y^* + \zeta_2 C^* + \delta$$  \hspace{1cm} (5.16)

with the error term, $\delta$, Normally distributed with mean zero and constant variance.

This model is then extended by an exponential model with the intention of better specifying the particularities of duration data: left truncated at zero and skewed to the right. These two features also define the distribution function of the ME (see figure 15 in Section 4.2) when derived as the probability density function of the difference between PRESO and LSA durations, or

$$V = Y - Y^*$$  \hspace{1cm} (5.17)

The exponential model can be expressed linearly as an accelerated life model by taking the logarithm of the durations of unemployment as follows,

$$\log(Y) = \vartheta_0 + \vartheta_1 Y^* + \vartheta_2 C^* + \omega$$  \hspace{1cm} (5.18)

with the error term $\omega$ now following an extreme value type-1 distribution. Predictions from this model used to replace the mismeasured durations will be transformed back to the original duration scale by using the exponential function, $\exp(\hat{Y})$. In addition, predictions obtained using both the linear and the AL exponential calibration model will be constrained to the interval $(0, 395)$, similar to the truncation in Section 5.2.1 for simulated durations in the SIMEX process.

The third approach that I propose involves the specification of a logit model. Here the aim is to explore the possibility of considering the ME problem affecting durations of unemployment as a discrete process of misclassification (MC) of durations as right-censored. As we saw in Figure 15 in Section 4.2, one of the most noticeable effects of ME on the distribution of true durations was the loss in the

\textsuperscript{44} All of the other variables available show very weak correlations with the ME term regardless of how it is defined. Here I include a list of the Pearson correlation estimates between the variables used in the model and the binary variable indicating whether the duration was not reported as censored: observed duration -.25, observed censoring .17, age .06, experience .02, age*experience .07.
number of right-censored cases. More formally, such a MC model entails the specification of two equations to define the probability of observing false positives (FP) and false negatives (FN); that is, the possibility of observing a duration being right-censored \((C^* = 1)\) when it really is not \((C = 0)\), and the possibility of observing a duration as not right-censored when it really is,

\[
\begin{align*}
P(C^* = 1|C = 0) &= \theta_{0|1} ; & \text{False positive} \\
\P(C^* = 0|C = 1) &= \theta_{1|0} ; & \text{False negative}
\end{align*}
\]

(5.19)

To complete the MC model, the two probabilities are complemented by the sensitivity and specificity; respectively, the probability of observing a true positive or a true negative,

\[
\begin{align*}
P(C^* = 1|C = 1) &= \theta_{1|1} ; & \text{True positive (Sensitivity)} \\
P(C^* = 0|C = 0) &= \theta_{0|0} ; & \text{True negative (Specificity)}
\end{align*}
\]

(5.20)

According to this definition of MC, and given that there is a much bigger proportion of FN durations in the sample under study (true right-censored cases that were reported not to be so) than there are FPs (30% of cases are FNs and only 1% are FPs), the RC adjustment is restricted to the former. That is, I don’t try to adjust for FP cases. Specifically, the calibration model used here estimates the probability of a true duration being right-censored conditional on both the observed censoring and duration, \(P(C = 1|C^*, Y^*)\), which here is denoted as just \(P(C)\). I use the logit transformation to approximate the probability scale from the binary variable \(C\) without having to rely on a latent variable specification, so

\[
\text{logit}(C) = \log \left( \frac{P(C)}{1 - P(C)} \right) = \zeta_0 + \zeta_1 Y^* + \zeta_2 C^*
\]

(5.21)

Predictions from this model were used to set cases as right censored, \(\hat{Y} > 395\), if the model could generate substantial evidence of that being the case, which I established by \(P(C) \geq .6\), with cases below that probability threshold left unaltered, or \(\hat{Y} = Y^*\). The reasoning behind the use of such a threshold instead of making a prediction of censoring being misclassified from the more common \(P(C) \geq .5\) was to avoid artificially creating additional false positive cases\(^{45}\).

\(^{45}\) As seen in the previous footnote, the correlation of the explanatory variables with the ME term, regardless of how it is defined- is quite poor. So, it is possible that in instances where the validation subsample is not very informative, predictions from the RC process using a probability threshold
The advantage of this discrete approach over the standard RC is that I can better discriminate between the two types of ME that reports from LSA are, arguably, prone to; namely 42% of durations were correctly reported or subject to classical additive ME, however, 30% of cases were much affected by more severe problem of underreported durations. Those 30% cases were right-censored in PRESO and appeared to take any value within the window of observation in LSA. Using the MC approach I can now focus on adjusting observations that are suspected to be prone to the more serious ME problem of underreported right-censored cases, whereas in the standard RC I would be adjusting for an average of both the underreporting and classical additive mechanisms at the same time, and the adjustments would be carried out on all the observed durations, regardless of the type of ME that they are prone to.

All these three functional forms of RC employ the bootstrap to calculate the SEs of the estimates of interest in the outcome model. In particular, I use 100 bootstrap samples instead of the 200 resamples recommended by Efron and Tibshirani (1993) to make the process less computationally intensive. To assess the adequacy of this method, an analysis for the same data using MI is also included. In addition, to constrain imputed durations to the (0, 395) interval I use the imputation method of predictive mean matching (PMM) - programmed in the R package “BaBooN” - which uses a chained equation algorithm. Predictive mean-matching subsamples from the observed data, which prevents the generation of nonsensical imputations. Specifically, the method calculates the predicted value of the target variable for each missing entry \( \hat{Y}_s \) according to the specified linear imputation model (equation 5.16), then a small set of candidate donor (here I used 3 donors) is taken from all complete cases \( Y_s \) that are close enough to the predicted value \( \hat{Y}_s \). A random draw is made among the candidates, and one of the donors \( Y_s \) is taken to replace the missing value \( Y_s' \).

Missing data theorists have often claimed that good inferences can be made with the number of imputed data sets \( K \) as few as 3 or 5. They have argued that the relative efficiency of the estimations is very high under these circumstances, compared to an infinite number of imputations. However, Graham et al. (2007) show that the effects close to \( P(C) \geq .5 \) could be generating additional MC cases. It is for this reason that I decided to use the more conservative threshold of \( P(C) \geq .6 \).
of $K$ on statistical power for detecting small effects (e.g. an F-test statistic for the equality of the variances of two variables that in the presence of complete data had a p-value equal to one) can be strikingly different from what is observed when assessing the relative efficiency as suggested above. The authors showed that if statistical power is an essential consideration, the number of imputations typically must be higher than previously thought. Graham et al. (2007) recommend that at least $K = 40$ imputations are needed with 50% missing information to guarantee less than a 1% power falloff compared to the comparable full information maximum likelihood approach. Here I decided to use $K = 10$ - which is higher than what is customary but not as high as Graham et al. (2007) recommend - to avoid making the process too computationally intensive.

Besides these four forms of RC and MI I also report the regression estimates obtained when the observed LSA durations are used but replacing the mismeasured durations with the durations obtained from the validation subsample. Formally, this would be equivalent to specifying the outcome model using $Y_s$ or $Y_s^*$ according to whenever the observation $s \in J$ or $s \notin J$, respectively; I call this approach “validation without calibration”. Results from this last model are used to assess the possibility of any of the three types of RC increasing the size of the bias in the regression estimates of the naïve model. In addition, to assess how the effectiveness of RC varies depending on the size of the validation subsample I carried out adjustments assuming validations subsamples that go from 10% to 50% of the original sample.

Finally, given the small size of the original sample, 381 observations, and the different ME patterns observed (see Section 4.2), I decided to include an iterative process for the selection of validation subsamples. The RC algorithm for each of the functional forms used is repeated 10,000 times, with the characteristic that in each of these iterations a different validation subsample of the same size is taken at random from the register data. This is a process that I use to better evaluate the effectiveness of the adjustment. Given the small size of the sample, the effectiveness of the method will depend on the composition of the validation subsample that is taken. However, by averaging the results over the 10,000 iterations I manage to eliminate the volatility induced by the choice of a specific subsample.
Results for the RC and MI Adjustments

I start by reporting results obtained for one validation subsample with size 25% of the original sample (that is, 95 out of 381 subjects). This is the percentage used in Cole et al. (2006), a similar study, where the effectiveness of RC and MI is compared. Figure 32 shows scatterplots of the reported durations in the survey against the durations used in each of the four adjustments studied. In these figures, durations from the validation subsample are represented by the black circles, while the red circles denote predicted cases.

*Figure 32. LSA and calibrated durations: logit, linear, exponential, and PMM*

The comparison of predictions from RC using a linear or an exponential calibration shows the relevance of the choice of an appropriate functional form. For example, we can see how the effect of LSA durations - $\zeta_1$ in equations 5.16 and 5.21, and $\vartheta_1$ in equation 5.18 - changed sign when an exponential calibration model is used.

---

46 Specifically Cole et al. (2006) studied the use of RC and MI for a problem of misclassification in one of the explanatory variables of a PH Cox model with 600 observations.

47 The PMM scatterplot shows durations obtained from the first of ten iterations used before the pooling step against the LSA durations.
instead of the standard linear one; although this is also related to the estimate for LSA not being statistically significant in both the linear and the exponential calibration model. In addition, for this particular validation subsample, the RC calibration using a logit model has identified LSA durations lower than 100 to be FN cases of right censored durations, with the rest of the predictions for the calibration taken as unchanged LSA durations. Finally, as could be expected from the configuration of the method, PMM predictions only take values that were present in the validation subsample, which facilitates the implementation of the method when using duration data.

To observe the extent to which the adjusted durations reflect the original shape of the PRESO durations four graphs have been included (in Figure 33) showing the kernel density functions of durations from LSA (red line), PRESO (black line) and adjustments from the different functional forms that we used (blue line). Notice that the scales for the y-axis change across some of the four plots presented in Figure 33 to encompass the distributions of calibrated durations that were particularly pointed.

Using these graphs we can determine the inadequacy of some of the adjustments visually. In particular, it is clear that the adjusted durations using both the standard linear and the exponential calibration models fail to reflect the shape of the true durations. The former generates a vast number of predictions of durations around 250 days; the latter also includes an exaggerated concentration of values for durations of 395 days. On the contrary, the MC approach using a logit calibration model manages to reflect roughly the density function of PRESO durations. In particular, the bump around 50 days observed in LSA is eliminated, while the bump for right-censored cases in PRESO that are omitted in LSA is now recreated.
Figure 33. Probability density functions for the different adjusted durations

The best solution at first sight seems to be the PMM. This MI approach generates predicted durations that map the distribution of true durations remarkably well. Specifically, as Table 35 shows, the PMM adjustment produced a practically unbiased estimate of the standard deviation of PRESO durations, only one day shorter while the rest of adjustments fell at least 28 days short. The PMM was also relatively accurate regarding the percentage of right-censored cases (ten percentage points than the true value) and the mean (20 days higher). However, regarding adjustments for measures of centrality such as the mean and the median, the linear model preformed better than any other adjustment.

Table 35. Effectiveness in R.R.BIAS and R.RMSE with 25% validation subsample

<table>
<thead>
<tr>
<th></th>
<th>Validation</th>
<th>PRESO</th>
<th>Linear</th>
<th>Exponential</th>
<th>Logit</th>
<th>PMM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>252</td>
<td>241</td>
<td>258</td>
<td>349</td>
<td>306</td>
<td>261</td>
</tr>
<tr>
<td>Median</td>
<td>256</td>
<td>253</td>
<td>246</td>
<td>395</td>
<td>395</td>
<td>330</td>
</tr>
<tr>
<td>Std. Dev.</td>
<td>143</td>
<td>145</td>
<td>79</td>
<td>94</td>
<td>117</td>
<td>144</td>
</tr>
<tr>
<td>% Censored</td>
<td>41%</td>
<td>35%</td>
<td>16%</td>
<td>60%</td>
<td>55%</td>
<td>45%</td>
</tr>
</tbody>
</table>
The capacity to recreate the distribution of true durations demonstrates the effectiveness of the standard RC (using a linear calibration model) and PMM to adjust for the ME problem when we are interested in calculating univariate statistics such as the mean or variance of durations of unemployment. However, it is not a sufficient condition to demonstrate the effectiveness of the method in multivariate analyses such as the PH Cox model. To assess and compare the effectiveness of the different methods explored in the PH Cox model of interest I calculate their mean R.R.BIAS and R.R.RMSE over 10,000 iterations, and for a validation subsample of size 25%.

Results are shown in Table 36, where each method is divided in three rows to differentiate between each regression estimate from the outcome model; with the exception of the first three rows under the name validation, which indicate adjustments using the benchmark of “validation without calibration”. This is used to assess whether the adjustment performs better than just a naive analysis that replaces the observed durations included in the validation subsample by the true ones.

Regarding effectiveness in terms of bias reduction, we can see that all the methods explored manage to reduce the R.BIAS found under the naive analysis. However, the use of RC with exponential calibration made things worse for every estimate than if I had just opted for rerunning the naive analysis substituting the LSA for PRESO durations included in the validation subsample. This is not the case for the standard calibration approach using a linear model, which in spite of its poor capacity to recreate the distribution of true durations, obtains R.R.BIASes around eight point smaller than the benchmark “validation without calibration”. The PMM approach using MI also performs slightly better than the benchmark, with an R.R.BIAS about four points smaller on average for the three regression estimates. This small effect contrasts with the more promising picture shown in Figure 33, and with the much better result obtained using a logit calibration form. The adjustment considering the MC of right censored cases, in spite of being an ad hoc approach, managed to reduce the size of the R.BIAS obtained using the naive analysis by almost a half.

The relative effectiveness of these four methods changes when considering the R.R.RMSE. In this case, the standard RC approach is the only solution that performs better than the simpler “validation without calibration” approach, while the use of an exponential calibration again shows the worst results. In addition, it is worth noting
that PMM—which is based on Rubin’s laws to estimate the SEs—performs roughly similarly to the rest of the RC procedures using paired bootstraps. In particular, PMM performs slightly better in terms of R.R.RMSE on average than the exponential or the logit RC solutions.

Table 36. Effectiveness in R.R.BIAS and R.R.MSE with 25% validation subsample

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>SE</th>
<th>R.R.BIAS</th>
<th>R.R.RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Age</strong></td>
<td>.026</td>
<td>.033</td>
<td>72.5</td>
<td>61.9</td>
</tr>
<tr>
<td><strong>Validation</strong></td>
<td>Experience</td>
<td>.475</td>
<td>.436</td>
<td>72.8</td>
</tr>
<tr>
<td>Age*Experience</td>
<td>-.012</td>
<td>.012</td>
<td>72.7</td>
<td>63.4</td>
</tr>
<tr>
<td><strong>Linear</strong></td>
<td>Age</td>
<td>.029</td>
<td>.033</td>
<td>64.2</td>
</tr>
<tr>
<td>Experience</td>
<td>.516</td>
<td>.434</td>
<td>63.6</td>
<td>50.6</td>
</tr>
<tr>
<td>Age*Experience</td>
<td>-.013</td>
<td>.012</td>
<td>64.4</td>
<td>54.4</td>
</tr>
<tr>
<td><strong>Exponential</strong></td>
<td>Age</td>
<td>.021</td>
<td>.037</td>
<td>79</td>
</tr>
<tr>
<td>Experience</td>
<td>.429</td>
<td>.493</td>
<td>77.3</td>
<td>71.1</td>
</tr>
<tr>
<td>Age*Experience</td>
<td>-.011</td>
<td>.013</td>
<td>77.1</td>
<td>74.6</td>
</tr>
<tr>
<td><strong>Logit</strong></td>
<td>Age</td>
<td>.066</td>
<td>.052</td>
<td>57.9</td>
</tr>
<tr>
<td>Experience</td>
<td>.796</td>
<td>.708</td>
<td>57.2</td>
<td>72.7</td>
</tr>
<tr>
<td>Age*Experience</td>
<td>-.023</td>
<td>.019</td>
<td>53.9</td>
<td>68.8</td>
</tr>
<tr>
<td><strong>PMM</strong></td>
<td>Age</td>
<td>.032</td>
<td>.038</td>
<td>64.8</td>
</tr>
<tr>
<td>Experience</td>
<td>.496</td>
<td>.503</td>
<td>70.6</td>
<td>63.5</td>
</tr>
<tr>
<td>Age*Experience</td>
<td>-.013</td>
<td>.014</td>
<td>68.5</td>
<td>67.1</td>
</tr>
</tbody>
</table>

*The R.R.BIAS for RC adjustments in age, age*experience, and experience, using the logit functional form and for the standard P(C) ≥ 0.5 are .66, .67 and .74, respectively; whereas for a higher cut-off point such as P(C) ≥ 0.7 these are .61, .68 and .72.

Figures shown in Table 36 represent averages over the 10,000 iterations that were used to make results representative of any possible 25%-size validation subsample. Such a high number of iterations was required because of the vast dispersion observed on the adjustment of the regression estimate of experience when different validation subsamples where used. The standard deviations of the sampling distributions empirically derived from the 10,000 iterations using PMM adjustments are .027, .009, and .349, for the estimates of age, the interaction, and experience, respectively. These sampling distributions are also shown graphically in Figure 34, where age is denoted by a blue line, the interaction in red, and experience in green.
Figure 34. PMM adjusted estimates across validation subsamples of size 25%

The green line in Figure 34 for the estimate of experience shows the risk of reaching an erroneous conclusion regarding the effectiveness of the methods tested here if an iterative mechanism had not been used. For example, there is a 2.5% that by taking a 25% validation subsample at random we would estimate the effect of experience equal or bigger than 1.22, concluding that the adjustment was successful. This risk would probably be lower if a bigger sample for the estimation of the outcome model would have been available, since in that case the sample size of the validation subsample would increase, hence, reducing its sampling variability. In addition, from the comparison of the different levels of uncertainty for the three estimates it seems that the stronger the original effect the more sensitive the adjustment is to the configuration of the validation subsample.

To complete the assessment of the different functional forms I also report their effectiveness in terms of the R.R.BIAS when using nine different validation subsamples with sizes going from 10% to 50%. These results are reported in Table 37 and Figure 35. The latter includes “RC adjustment curves” for each of the estimates and validation subsamples to facilitate the interpretation of the results. The black line represents the validation without calibration process, the blue line captures the standard linear calibration, the red line is for the calibrations using an AL exponential model, the green line for the calibration using a logit model to estimate MC of right censored cases, and the brown line denotes results for the PMM approach.
Table 37. Effectiveness in terms of R.R.BIAS for different validation subsamples and functional forms

<table>
<thead>
<tr>
<th>Validation</th>
<th>Linear</th>
<th>Exponential</th>
<th>Logit</th>
<th>PMM</th>
</tr>
</thead>
<tbody>
<tr>
<td>10%</td>
<td>87.9</td>
<td>88</td>
<td>88</td>
<td>86.5</td>
</tr>
<tr>
<td>15%</td>
<td>82.2</td>
<td>82.5</td>
<td>82.4</td>
<td>78.8</td>
</tr>
<tr>
<td>20%</td>
<td>76.4</td>
<td>76.7</td>
<td>76.7</td>
<td>72.4</td>
</tr>
<tr>
<td>25%</td>
<td>72.3</td>
<td>72.6</td>
<td>72.5</td>
<td>67.1</td>
</tr>
<tr>
<td>30%</td>
<td>66.3</td>
<td>66.6</td>
<td>66.6</td>
<td>61.8</td>
</tr>
<tr>
<td>35%</td>
<td>61.6</td>
<td>61.9</td>
<td>61.9</td>
<td>56.8</td>
</tr>
<tr>
<td>40%</td>
<td>56.9</td>
<td>57.1</td>
<td>57</td>
<td>52</td>
</tr>
<tr>
<td>45%</td>
<td>52.3</td>
<td>52.6</td>
<td>52.4</td>
<td>47.7</td>
</tr>
<tr>
<td>50%</td>
<td>43.1</td>
<td>43.2</td>
<td>43.0</td>
<td>43.9</td>
</tr>
</tbody>
</table>
Figure 35. Adjustment curves for different calibration models

**Age**

**Experience**

**Age*Experience**
The effectiveness of the different approaches can be observed graphically as the distance between the different lines with the black one. Using this comparison we can see that the use of an AL exponential calibration model was a bad solution, as it increases the size of the R.BIAS for most of the validation subsamples - the only exception being the case of experience and the interaction effect for a validation subsample of 45%.

The inadequacy of using such a functional form for the calibration model becomes even clearer when we compare it against the results obtained from the standard linear calibration. Here, RC performs positively across the three regression estimates and all the different sizes of validation data. However, the effectiveness of the adjustment is quite small. The average improvement for the estimate of age with respect to the validation without calibration is three R.R.BIAS percentage points, becoming almost non-existent for the extreme cases of a 10% and 50% validation subsample, and stretching the improvement up to five percentage points in the middle of the curve, when 25% and 30% validation subsamples are used.

In general, the PMM approach performs very similarly to the linear RC calibration, indicating that choosing between a standard RC and a MI approach using PMM is almost irrelevant regarding R.R.BIAS. The only slight difference detected refers to the specific estimate of age, which is four percentage points more effective using MI.

More satisfactory results are obtained for the logit calibration model. As was found for the linear case, the improvements at the extremes of the validations subsamples are negligible, however, the improvement for the middle cases of 25% and 30% validation in the interaction term reached a maximum distance of 29 percentage points lower than the validation without calibration benchmark, which resulted in an average R.R.BIAS adjustment across the validation subsamples of 55.7%. That is, on average, the RC solution using a logit calibration model managed to reduce the size of the bias found using the naïve analysis by almost a half.
5.3.2. Conclusion from Regression Calibration and Multiple Imputation

I have compared the effectiveness of RC - in various forms - and MI, a method commonly used to adjust for problems of missing data, but which has many similarities with RC.

In the exploratory part of the analysis we have seen that MI using PMM approximates the distribution of the true durations, which makes it a potentially useful technique when the interest lies on estimating adjusted univariate statistics. This contrasts with the results from RC using the same validation subsample and a linear model, where most of the predicted durations concentrated around durations of 250 days.

An exponential calibration model was also explored with the intention of better representing typical features of duration data (such as right skewness, lower bound at zero, and right censoring). However, this process did not generate the expected results, making things worse than the alternative of not carrying out a calibration adjustment at all.

The more standard linear RC and the PMM approaches managed to reduce the size of the bias in the estimated model coefficients, although never by more than 10%. However, the effectiveness of PMM in terms of R.R.RMSE showed worse results than both the standard RC and the benchmark of the validation without calibration. This higher statistical efficiency of RC than MI contrasts with results from Cole et al. (2006), and Messer and Natarajan (2008), where substantial similarities between the effectiveness of MI and RC were found both in terms of R.BIAS and R.RMSE, for different ME scenarios.

These differences could be due to different circumstances: the nature of the ME term, the type of variable being affected (i.e. a response variable prone to censoring here), or the type of MI approach used. Both Cole et al. (2006) and Messer and Natarajan (2008) used conditional simulations from a multivariate normal distribution to generate the imputations without relying on an additional PMM mechanism. Cole et al. (2006) used a sample of 600 subjects with different levels of MC affecting an explanatory variable of a PH Cox model using simulated and validation subsamples of 15% and 25%. Messer and Natarajan (2008) used a sample
of 500 simulated cases where one of the explanatory variables of a logistic model is subject to different levels of classical ME with a systematic component and the size of the validation subsample is 50%.

In Cole et al. (2006), MI adjustments in terms of R.RMSE performed as well as or better than RC when measurement properties were poor (sensitivity and specificity equal to .7) or when measurement properties were good (sensitivity and specificity equal to .9) with a larger validation substudy (25% validation data instead of 15%). In Messer and Natarajan (2008) the MI adjustment in terms of R.RMSE was slightly more effective than RC in all the scenarios explored, becoming substantially better (13% lower R.RMSE) when the ME is large and correlated with the rest of the variables in the model.

Arguably the limited effectiveness of the standard RC and MI processes in the analyses presented here is related to the complex ME generated by these data. As was explained in Section 4.2, we can identify two distinct error mechanisms operating simultaneously: one involving classical type of errors with a standard deviation of about 30 days, which could be related to the problem of misdating, and the other generating durations that are unrelated to their true values, which could relate to the problems of omission of spells or MC of work-status.

Standard RC and MI can overcome some of the restrictions that other methods impose when requiring classical ME, for example it can produce effective adjustments in the presence of systematic ME. However, because of the complexity of the ME under analysis, methods using a single linear calibration model are bound to perform poorly. Predictions made from such methods run the risk of altering durations that were unaffected – or very little affected - by ME.

On the other hand, we have seen that changing the perspective, so the event-occurrence ME problem is turned into a MC problem, has the potential to provide sensible adjustments. This is possible if, as I argue, the ME mechanism affects some cases and leaves others unaffected, but also because of the presence of the same right censoring date, 395 days, for all durations in the sample. This MC approach left unadjusted all those durations affected by the first of the postulated error mechanisms; a standard problem of classical ME affecting a continuous variable. However, by focusing on the second error mechanism the bigger problem is tackled.
First, this second mechanism, unlike the previous one, is systematic, and second, the size of the ME is much larger. In particular, the quality of this logit RC adjustment could have been much higher if I have had access to more informative variables to be used as explanatory variables in the calibration model: specifically, variables correlated with the second ME mechanism reflecting FN censored cases.

In the next section the scope of likelihood-based methods for the adjustment of ME will be reviewed. In particular I carry out Bayesian adjustments that will allow me to use more flexible solutions in my aim of adjusting for event-occurrence ME.

5.4. Bayesian Adjustments

In this section I present adjustments using a Bayesian approach. In comparison to the previous methods described in this chapter - i.e. SIMEX, RC and MI - likelihood methods such as Bayesian adjustments belong to the family of structural ME methods (see Section 5.1). In addition to the outcome model (or model of interest), structural methods require the specification of both the observed variable prone to ME (the measurement model) and the true but unobserved variable (the model of exposure\textsuperscript{48}), which adds complexity to the adjustment and increases the probability of incurring in problems of misspecification. On the other hand, likelihood-based methods are more flexible since they can in principle be applied to any type of ME process through the appropriate specification of the measurement model.

For an outcome variable \( Y \) that we observe with ME as \( Y^* \), a likelihood-based adjustment entails the specification of the following parts: i) the outcome model, which would be the one in use in the absence of ME, \( f(Y|X,\beta) \); and ii) the measurement model where the distribution of the observations with errors \( Y^* \) is defined, conditional on the other variables, as \( f(Y^*|Y,X,y) \). The likelihood function given only observables may be obtained as

\[
f(Y^*|X,Y,\beta) = \int f(Y^*|X,Y,y)f(y|X,\beta) \, dy
\]  

(5.22)

\textsuperscript{48} The term exposure model comes from epidemiology, where normally ME adjustments are carried out in the context of ME affecting an explanatory variable that captures exposure to a toxicant.
Note that in a case where an explanatory variable is affected by ME, \( X^* \), an exposure model, \( f(X|Z) \), representing the distribution of the true variable \( X \) given the observable explanatory variables \( Z \), also needs to be inserted in the likelihood function. In a case like the one described in equation 5.22, where only the response variable is affected by ME, the exposure model is already encompassed by the outcome model.

The possibility of specifying the ME freely to map the error generating mechanism adequately is the key element that gives likelihood-based methods an advantage over the previously discussed methods in terms of flexibility and applicability. Another feature that makes likelihood-based methods a useful approach in the presence of ME is their ability to propagate the uncertainty associated with each parameter of the model both automatically and appropriately. Methods that rely on second stage estimations (e.g. SIMEX and RC) do not facilitate such propagation, which makes them statistically less efficient in comparison (Gustafson, 2003).

However, this flexibility comes at a price. In particular, the likelihood function is obtained by integrating over the true but unobserved variable, and this step can be difficult due to its possible high dimensionality. In addition, quite often this integration cannot be done explicitly in closed-form. In such cases, one possibility is to evaluate the likelihood function using a numerical integration scheme, so that iterative algorithms to maximize the likelihood function can be employed. Specifically the expectation-maximization algorithm (Dempster et al., 1977) is the method often used in these situations, although other options such as adaptive quadrature (McKeeman, 1962) can also be used.

This additional flexibility and precision also comes at the cost of having to include extra assumptions. A common criticism of the parametric approach that likelihood-based methods involve is that one runs the risk of specifying a distributional form for the measurement model (and in the case of ME in explanatory variables the exposure model) that is incorrect. This criticism is valid for all parametric statistical modelling, though one might argue that the problem is particularly acute for exposure models and ME models. Particularly because the variables are not observable, making them hard to be checked empirically.
Bayesian inference gets its name from Bayes’ Theorem which provides us with a technique for updating our a priori uncertainty about something unknown to a posterior uncertainty through observations. The theorem states that we may use the conditional probability of an event, A, that has occurred to calculate the conditional probability that an event, B, is true

\[ P(B|A) = \frac{P(A|B)P(B)}{P(A|B^c)P(B^c) + P(A|B)P(B)} \]  

(5.23)

where \( P(B^c) = 1 - P(B) \). The event \( A \) is the outcome for some variable following a distribution \( P(A|B) \) that depends on some true but unknown state of the world \( B \) (usually represented through one or more parameters). The conditional probability \( P(B|A) \) thus tells us how probable it is that the true state of the world is \( B \) given the evidence provided by our observation \( A \). Note that this operation requires that we allocate an a priori probability to the state of the world \( B \) being true. We can use this theorem to derive the posterior distribution for any unknown quantity. The posterior distribution represents the uncertainty about a parameter, \( \theta \), given a sample of observed data, \( y \). In the general case the posterior distribution may be expressed in terms of two constituent parts: i) the prior distribution, \( p(\theta) \), which embodies our prior knowledge about the model parameters; and ii) a likelihood function, \( p(y|\theta) \), which is the probability distribution (mass) function for \( y \) given \( \theta \); so

\[ p(\theta|Y) = \frac{p(Y|\theta)p(\theta)}{\int p(Y|\theta)p(\theta) d\theta} \propto p(Y|\theta)p(\theta) \]  

(5.24)

The denominator in the posterior is a normalising constant assuring that the posterior is a proper probability distribution. The normalising constant is commonly referred to as the marginal likelihood as it is the likelihood marginalised with respect to the parameters. As such the marginal likelihood itself does not carry any information about the parameters of interest. Typically, the marginal likelihood is unavailable in closed form making the posterior density analytically intractable. Markov chain Monte Carlo (MCMC) algorithms are a flexible class of methods for drawing parameters from the posterior that do not require the evaluation of the marginal likelihood. These algorithms generate a sequence of parameter draws that can be shown to be serially correlated draws from the posterior distribution of the model parameters after a sufficiently long ‘burn-in’ time. Based on these samples from the posterior distribution, one can obtain direct estimates of the parameters of interest
and their respective uncertainty by using the MCMC equivalents of expected values and variances.

The requirement, in a Bayesian approach, to use prior distributions is a controversial issue. Priors are used to include in the model any previous knowledge about the phenomenon that is being studied, and as such are sometimes dismissed as unscientific. What critics seem to miss is that sensitivity of results may be investigated using a range of prior distributions, including 'diffuse priors', i.e. probability distributions that assign equal probability to the possible values that a parameter could take. Furthermore, prior distributions become a really useful tool in the presence of models that cannot be identified - as is often the case when dealing with ME problems. “One intuitive way of thinking about Bayesian inference in the absence of parameter identifiability is that the prior distribution plays more of a role than usual in determining the posterior belief about the parameters having seen the data”, (Gustafson, 2003, p. 64). All the same, it has to be kept in mind that the greater the reliance on the prior distributions the less the inferences will be based on the data, and in consequence, the worse the priors’ misspecification problem might be.

5.4.1. Implementation of the Bayesian Adjustments

In this section I make use of the flexibility of Bayesian methods to explore the effectiveness of three different adjustments. I start the analysis by re-estimating the true and naïve AL exponential models (based on PRESO and LSA durations) using a Bayesian approach. After that, I run the first of the adjustments assuming that the durations of unemployment are affected by classical additive ME. I use this method to investigate whether the inadequacy of methods assuming classical ME for the problem of event-occurrence ME can be generalised to methods other than SIMEX. Second, I carry out an adjustment for the same outcome model that does not rely on any specific knowledge of the distribution of the ME term, instead it relies on a validation subsample to inform the imputation of unobserved true durations under the assumption of a linear model and MAR. As such, this adjustment shares some similarities with the standard RC or MI implemented in the previous section. Finally, I make use of what we have postulated in Section 4.2 to specify a mixture model that could account for the different types of event-occurrence ME.
To run these models I use WinBUGS (Spiegelhalter et al., 2003), freely available software using a Bayesian approach that relies on Markov chain Monte Carlo (MCMC) (Geman and Geman, 1984). MCMC may be used when the full conditional posterior distributions of all unobservables are known up to normalising constants and, for the particular case when these reduce to known distributions that are simple to draw from ‘directly’, Gibbs sampling may be used. This is the case when the model has a simple factorisation in terms of the conditional distribution of each node given its parents. Such a model represents the assumption that, conditional on its parents’ nodes, each node is independent of all the other nodes except descendants, and chaining these relationships together, a full joint distribution over all unknown quantities can be expressed using their factorisation.

Graphical models (Lauritzen, 1996) can be used to represent complex models subject to child-parent conditioning, “the graphical representation allows us to reduce globally complex models into a set of fairly simple local components” (Lunn et al., 2013, p. 15). This is done using Directed Acyclic Graphs (DAG), which express the joint relationship between all known and unknown quantities in a model.

**The True and Naïve Outcome Models**

The true and naïve AL exponential models analysed in Section 4.2.2 using a frequentist approach (equation 1.36) are represented in Figure 36 as a DAG. The ellipses of the DAGs denote the parameters to be estimated, squares represent the exogenous data, arrows are used to indicate the direction of a dependency and the dashed line indicates the threshold of censored observations, $C_i$. The true model is represented in Figure 36 since the true durations are being used as the response variable, to characterise the naïve model we would only need to substitute $Y_i$ and $C_i$ in Figure 36 by $Y_i^*$ and $C_i^*$. 
The joint distribution captured by Figure 36 can be formally expressed as follows,

\[
p(Y, \beta | X) = p(\beta) \prod_{i} p(Y_i | \mu_i) \tag{5.25}
\]

where \( X \) represents the three explanatory variables, \( age, ex, \) and their interaction term \( ae, \) and \( \mu_i \) is a linear function of the explanatory variables and the parameters \( \beta. \)

For the outcome model, \( p(Y_i | \mu_i) \), I use a Weibull distribution with its scale parameter fixed at one, which de facto converts it in an exponential distribution\(^49\),

\[
Y_i | \mu \sim \text{Weibull}(\mu, \alpha = 1)
\]

where \( \mu \) is a function of the data and the parameters of the exponential model such as,

\[
\mu_i = g(x_1, ..., x_N; \beta) = \beta_0 + \beta_1 age_i + \beta_2 exp_i + \beta_3 ae_i \tag{5.26}
\]

I estimated an AL exponential model instead of the PH Cox model that I used for the SIMEX and RC adjustments because the PH Cox is more computationally demanding, which would make some of the models that I explore later harder to estimate. This choice should not, however, be problematic since interest lies in the estimates of the constant, \( age, experience, \) and their interaction which, as we saw in Sections 4.2.2 and 4.2.3, are very similar regardless of the EHA model used. To complete the specification of the model presented in Figure 36 I give diffuse priors to the regression estimates included in \( \mu. \) In particular, I assume that \( \beta \sim N(0, 100^2) \) independently for each of the regression estimates.

\(^49\) Such transformation was necessary instead of simply using an exponential distribution for the outcome model to prevent WinBUGS from crashing when implementing some of the more complex Bayesian adjustments that follow.
In all the models I use two chains with an equal number of iterations to assess convergence and sample from the posterior distribution. For the naïve model, I sampled 500000 iterations from which the first 2000 were discarded as burnt-in. For the true model I increased the number of iterations to 800000 while burning-in the first 10000 due to the slower convergence that it showed – probably due to the higher proportion of right censored cases. In the models that follow I always burn-in the first 10,000 iterations and, to estimate the posterior distribution, I use a range of iterations going from 400,000 to 800,000, depending on how well the two chains used to assess convergence mixed. The specific number of iterations that I use in each model can be found in Appendix G, where I include trace plots for the regression estimates of age to show that convergence is sufficient for all the models. In addition, I include the WinBUGS syntax that I used to specify each of the Bayesian models in Appendix H.

Results from both the true and naïve models are shown in Table 38.

<table>
<thead>
<tr>
<th></th>
<th>Register</th>
<th>Survey</th>
</tr>
</thead>
<tbody>
<tr>
<td>constant</td>
<td>9.10 (1.36)</td>
<td>5.12 (1.15)</td>
</tr>
<tr>
<td>age</td>
<td>-.088 (.038)</td>
<td>-.001 (.032)</td>
</tr>
<tr>
<td>experience</td>
<td>-1.39 (.50)</td>
<td>-.13 (.42)</td>
</tr>
<tr>
<td>age*exp</td>
<td>.038 (.014)</td>
<td>.002 (.012)</td>
</tr>
</tbody>
</table>

Compared to the standard errors obtained using a frequentist approach in Section 4.2.2 (Table 13) the posterior standard deviations are identical, while the biggest difference for the point estimates is for the estimate of experience, which is only .002 higher in both the true and naïve models. These similarities demonstrate that: a) the priors are diffuse enough so inferences from the posterior distribution are dominated by the likelihood function; and conditional on this being true, b) assuming convergence is plausible for both models, and c) the number of iterations used to form the posterior distribution was big enough to discount Monte Carlo error (i.e. the sampling error arising from the simulation process).

Adjustment Assuming Classical ME

The first adjustment that I present assumes very specific knowledge about the distribution of the error term - or the conditional distribution of the observed
durations on the true durations, \( f(Y^*|Y) \). In particular, I assume that \( f(Y^*|Y) \) can be characterised as a classical additive ME model (equations 1.1 and 1.2), Normally distributed with mean zero and known standard deviation of 31.6 (this is equivalent to .001 in terms of precision, which is defined as the inverse of the variance). This is an ad-hoc approach which may be conceived of as putting a point mass prior on the ME (that is also data-driven), but it facilitates the study of the additional uncertainty catered for by the Bayesian parameter inference.

The joint distribution for this adjustment assuming classical ME has the following form,

\[
p(Y, Y^*, \beta | X) = p(\beta) \prod_i p(Y_i | \mu_i) \prod_i p(Y_i^* | Y_i)
\]  

where \( i = 1,2,\ldots,395 \) is used to index the respondents captured in the sample.

The outcome model is specified as the Weibull model with \( \alpha = 1 \) defined previously. Note that the variable \( Y \) is now considered unobserved and the target posterior distribution that I aim to draw from is

\[
p(\beta|Y^*, X) = \frac{p(\beta) \prod_i \int p(Y_i | \mu_i)p(Y_i^* | Y_i) dY_i}{\int p(\beta) \prod_i \int p(Y_i | \mu_i)p(Y_i^* | Y_i) dY_i d\beta}
\]

The ME is classical additive with a standard deviation equal to 31.6, so

\[
Y^*|Y = Y + \nu, \quad \nu \sim N(0, 31.6^2)
\]

For the regression estimates of the outcome model I use similar a priori independent priors as in the previously case,

\[
\beta \sim N(0, 100^2)
\]

Figure 37 and Table 39 below show the graphical representation of this model and the results obtained.
Notice the differences from the DAG depicting the true model (Figure 36), which was extended here by connecting the unobserved true durations, $Y_i$, to the observed durations, $Y_i^*$, censored at $C_i^*$, and affected by the error term $v$.

Table 39. Results for the adjustment assuming classical ME

<table>
<thead>
<tr>
<th></th>
<th>Register</th>
<th>Survey</th>
<th>Adjustment</th>
</tr>
</thead>
<tbody>
<tr>
<td>constant</td>
<td>9.10 (1.36)</td>
<td>5.12 (1.15)</td>
<td>5.07 (1.11)</td>
</tr>
<tr>
<td>age</td>
<td>-.088 (.038)</td>
<td>-.001 (.032)</td>
<td>.001 (.031)</td>
</tr>
<tr>
<td>experience</td>
<td>-1.39 (.50)</td>
<td>-.13 (.42)</td>
<td>-.12 (.40)</td>
</tr>
<tr>
<td>age*exp</td>
<td>.038 (.014)</td>
<td>.002 (.012)</td>
<td>.002 (.011)</td>
</tr>
</tbody>
</table>

Results assuming classical ME are very similar to what was obtained for the naïve model, rendering the adjustment entirely ineffective. This is perhaps not surprising as the full conditional distribution of $Y_i$ given $\beta$ and $X$ has the kernel of a normal distribution with mean $Y_i^* = 31.6^2 / \mu_i$ and variance $31.6^2$. In fact, the bias\(^{50}\) in the estimates of $age$ and the constant increased with the implementation of the adjustment. This result reinforces my previous conclusion on the inadequacy of using adjustments assuming classical ME for problems of event-occurrence ME, regardless of the method implemented - SIMEX or likelihood-based methods.

\(^{50}\) I will keep using the term ‘bias’ in spite of not being entirely adequate in a Bayesian context to facilitate the presentation of the results from the different adjustments presented here. In particular, bias implies the difference between a parameter and the expectancy of its sampling distribution. However, sampling distributions are not used in the Bayesian framework.
Adjustment Assuming MAR

I now turn the ME problem into one of missing data using LSA durations as auxiliary data and assuming that 25% of the true durations are observable. Given the complexity of the model it was not possible to carry out an iterative process like the one used in Section 5.3 to assess the effectiveness of RC and MI independent of the configuration of the validation subsample\(^{31}\). Instead I took a validation subsample at random and checked whether it could be considered representative of the original sample. This assessment is shown in Table 40, where we can see that the means of age and experience and the medians of PRESO and LSA durations in both the original and the validation subsample that I took at random are relatively similar.

\[\text{Table 40. Means (and medians for durations) of variables included in the two samples}\]

<table>
<thead>
<tr>
<th></th>
<th>Original</th>
<th>Validation</th>
</tr>
</thead>
<tbody>
<tr>
<td>age</td>
<td>37</td>
<td>36.5</td>
</tr>
<tr>
<td>experience</td>
<td>2.6</td>
<td>2.5</td>
</tr>
<tr>
<td>PRESO duration</td>
<td>253</td>
<td>242</td>
</tr>
<tr>
<td>LSA duration</td>
<td>92</td>
<td>84</td>
</tr>
</tbody>
</table>

The joint distribution for the adjustment that I propose here can be defined as follows,

\[
p(Y, Y^*, \beta, \gamma, \sigma|X) = p(\beta)p(\gamma)p(\sigma) \prod_i p(Y_i|\mu_i) \prod_i p(y_i^*|\omega_i) \tag{5.30}
\]

where, as previously, the outcome model is defined as a Weibull model with shape \(\mu\) and scale fixed at one. The last factor of the joint distribution, \(p(Y^*|\omega)\), represents the ME model, which I specify as a Normal distribution with mean \(\omega\) and standard deviation \(\sigma\),

\[
Y^*|\omega \sim N(\omega, \sigma^2)
\]

where \(\omega\) is a linear function of PRESO durations and \(\gamma\) regression estimates, with binary \(c_i\) representing censored durations,

\(^{31}\) Bayesian models with large proportions of missing cases in WinBUGS tend to be unstable and can crash while the Gibbs sampler is updating. In addition, there is also a practical reason of time to reject the implementation of an iterative process, as the model that I present here took about four hours to produce a robust posterior distribution. Finally, in an iterative process it would be complicated to assess whether each of the models to be run reached convergence.
\[ \omega = g(y_1, \ldots, y_N; \gamma) = \gamma_0 + \gamma_1 Y_i + \gamma_2 C_i \] (5.31)

The rationale is to link \( Y \) with the fully observed \( Y^* \) so, akin to the RC and MI adjustments, the missing true durations can be predicted by extrapolating from the linear relationship between \( Y \) and \( Y^* \) included in the validation subsample. Here, however, the model is specified in terms of the conditional distribution of \( Y^* \) given \( Y \). That is, I assume that the unobserved \( y \) are missing at random conditioning on \( Y^* \).

Finally, for the parameters of this model I use a mix of a prior independent semi-informative and diffuse priors to facilitate convergence without imposing severe restrictions,

\[ \begin{align*}
\gamma_0 & \sim N(0, 100^2) \\
\gamma_1 & \sim N(0, 10^2) \\
\gamma_2 & \sim N(0, 100^2) \\
\beta & \sim N(0, 100^2) \\
\frac{1}{\sigma^2} & \sim \text{gamma}(0.001, 0.001)
\end{align*} \]

The last term represents the prior used for the standard deviation of the imputation model. Here it has been expressed in terms of the precision, which is how it needs to be expressed in WinBUGS. Setting priors for variances is notoriously difficult. To reflect lack of knowledge with respect to the variance term, Gelman (2006) recommends using a uniform prior on the scale of the standard deviation over a large (semi-infinite) range, or a half-normal distribution with large variance. However, here I use a gamma distribution with .001 for the shape and scale parameter as it is the choice of almost all the examples presented in Lunn et al. (2013) the reference book for the package WinBUGS. The expected value of a gamma distribution under such a parameterisation is one and the standard deviation 31.6, which assures that the vast majority of observations are concentrated in the (0,1) interval, but with a right tail that is wide enough to accommodate the possibility of the precision being much higher. The model is represented in DAG format in Figure 38, and results are given in Table 41.
Figure 38. Adjustment assuming MAR

The left-hand side of the graph is identical to that of the assumed true model, with the difference that 75% of the true durations are now missing. The right-hand side represents what could be understood as the imputation model that is used to link Y with the fully observed $Y^*$ using a linear model with means $\omega_i$ and standard deviation $\sigma$. Note that in the absence of this right-hand side part of the model 75% missing Y durations would have been estimated based on the relation of age, experience and their interaction with the Y durations included in the validation subsample. Note as well that Y is a parent node of $Y^*$, this structure is necessary since Y is already the response variable of the outcome model.

In Table 41, we see that the regression estimates for both the PRESO durations and the variable indicating whether they are censored have posteriors far removed from zero. In particular, the estimate for durations in PRESO is lower than one while that for PRESO-censored cases is negative, which reflects the general underreporting observed in LSA durations. However, and in spite of having used 25% of the true durations, we can see that the adjustment is relatively ineffective. The posterior standard deviations of the parameters of the outcome model are identical to those in the naïve model - with the exception of the constant term - whereas none of the ‘biases’ of these estimates were reduced by more than 25%.
The problem here seems to be the same one that was observed for the adjustments carried out in Section 5.3, namely, that these imputation processes cannot reflect the two seemingly distinct types of event-occurrence ME produced by, on the one hand minor random deviations and, on the other, strong underreports of certain durations. To account for this distinction I now implement an approach based on a mixture model.

Adjustments Using a Mixture Model

The final adjustments that I implement aim to account for the postulated twofold pattern of event-occurrence ME by controlling for the different ME mechanisms simultaneously. To do that, I use a mixture model that can differentiate between cases affected by classical ME and a second mechanism involving heavily underreported durations unrelated to the true values. The intuition is that as was shown in the scatter plot in Figure 15, the twofold ME pattern is represented by some observations being tightly clustered around the unit reference line while other observations are scattered around in a seemingly random pattern. In order to inform the mixture model about the allocation to mixture components and their respective variances for the ME process I explore two approaches: one relies on having access to a validation subsample while the other resorts to informative priors.

Specifically, for the first of these adjustments, the ability to discriminate between the two mixtures depends on what proportion of the two types of ME are evidenced by the validation sample. I use a validation subsample of just 10% of the true durations, which is a relatively low percentage, but large enough to make the model identifiable regardless of the configuration of the validation subsample. This was not the case for smaller validation subsamples of 5%.

### Table 41. Results for the adjustment assuming MAR

<table>
<thead>
<tr>
<th></th>
<th>Register</th>
<th>Survey</th>
<th>Adjustment</th>
</tr>
</thead>
<tbody>
<tr>
<td>constant</td>
<td>9.10 (1.36)</td>
<td>5.12 (1.15)</td>
<td>6.10 (1.44)</td>
</tr>
<tr>
<td>age</td>
<td>-0.088 (.038)</td>
<td>-.01 (.032)</td>
<td>-0.014 (.032)</td>
</tr>
<tr>
<td>experience</td>
<td>-1.39 (.50)</td>
<td>-.13 (.42)</td>
<td>-0.40 (.42)</td>
</tr>
<tr>
<td>age*exp</td>
<td>0.038 (.014)</td>
<td>.002 (.012)</td>
<td>-.010 (.12)</td>
</tr>
<tr>
<td>γ-constant</td>
<td></td>
<td></td>
<td>27.7 (5.94)</td>
</tr>
<tr>
<td>γ-PRESO</td>
<td></td>
<td></td>
<td>.59 (.05)</td>
</tr>
<tr>
<td>γ-Censored</td>
<td>-248.0 (24.1)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>sigma</td>
<td></td>
<td></td>
<td>55.0 (3.75)</td>
</tr>
</tbody>
</table>
The characteristics of the validation subsample used here are shown in Table 42. As in the previous adjustment, the means for *age* and *experience* for the original and validation samples are practically identical, whereas the medians of PRESO and LSA durations are now 24 and 20 days longer than in the subsample.

**Table 42. Means (and medians for durations) of variables included in the two samples**

<table>
<thead>
<tr>
<th></th>
<th>Original</th>
<th>Validation</th>
</tr>
</thead>
<tbody>
<tr>
<td>age</td>
<td>37</td>
<td>36</td>
</tr>
<tr>
<td>experience</td>
<td>2.6</td>
<td>2.6</td>
</tr>
<tr>
<td>PRESO duration</td>
<td>253</td>
<td>277</td>
</tr>
<tr>
<td>LSA duration</td>
<td>92</td>
<td>72</td>
</tr>
</tbody>
</table>

These differences are not unexpected given that the validation subsample contains only 38 cases. A small Monte Carlo study considering the sample size of the validation subsample and the standard deviations of the two types of durations (145.4 and 113.3) indicated that, in terms of durations of PRESO, these results are more representative of the original sample than 32% of the most extreme subsamples that could have been taken and the durations of LSA lie within the 70% of more representative samples.

The joint distribution for the adjustment based on a mixture model is defined as follows,

$$ p(Y, Y^*, \beta, \gamma, \sigma, T, \pi | X) = p(\beta \gamma) \prod p(\sigma) p(\pi) \prod_{i} p(Y_i \mid \mu_i) \prod_{i} p(Y^*_i \mid \omega_i, T_i) p(T_i \mid \pi) $$

where the outcome is specified as a Weibull as before, but where the ME is modelled as a mixture of two MEs. For each respondent *i* I assume that they belong to one of two unobserved categories indicated by the latent variable $T_i \in \{0,1\}$. If $T_i = 1$, I assume that

$$ Y_i^* | T_i = 1, Y_i, \sigma_1^2 \sim N(Y_i, \sigma_1^2) $$

i.e. the observed variable is affected only by classical additive ME. If $T_i = 0$, $Y_i^*$ is completely unrelated to the true value $Y_i$. I model this as:

$$ Y_i^* | T_i = 0, Y_i, \sigma_2^2 \sim N(\theta, \sigma_2^2) $$

We may express this mixture in various alternative ways, for example,
\[ Y_i' = T_i \omega_i' + (1 - T_i) \omega_i'' \]

(5.35)

where \( \omega_i' = y_i + \sigma_1 Z_i \) and \( \omega_i'' = \theta + \sigma_2 Z_i \), and where the \( Z_i \)'s have independent standard Normal distributions. The second element of the model corresponding to \( T_i = 0 \) has been constructed to account for those durations that are simply unrelated to LSA or any of the explanatory variables included in the outcome model. To reflect that nothing is known about these cases \( \theta \) has been allocated the following prior,

\[ \theta \sim N(187, 316^2) \]

For the mean of the prior I used 187, the mean duration for the window of observation used in the sample design (0.395). However, to reflect ignorance about \( \theta \) a very high standard deviation of 316 was used. The number 316 might seem arbitrary but it actually represents the standard deviation equivalent to a level of precision of 0.00001. Such high level of dispersion reduces the pointedness of the distribution of \( \theta \), and spreads the probability of observing values across the window of observation relatively evenly. Note as well that, with this prior, the range of values that \( \theta \) can take goes beyond 395, as is necessary to account for cases that were strongly underreported. On the other hand, to prevent the drawing of negative values from the prior - which could generate an undefined real result in the outcome model - values for the outcome model were truncated to be positive.\(^{52}\)

The latent variable \( T \) denoting the part of the mixture model to which cases are allocated is set to follow a Bernouilli distribution,

\[ \Pr(T_i = 1|Y, \beta, \gamma, \sigma, \pi, X) = \Pr(T_i = 1|\pi) = \pi \]

(5.36)

independently for all respondents conditional on the mixture proportion \( \pi \). I chose to set the prior for the proportion to a uniform distribution independently of other model parameters. This may be expressed in terms of the following diffuse prior,

\[ \pi \sim \text{beta}(1,1) \]

To reflect a lack of previous knowledge about the variances of the two error mechanisms and the rest of parameters defining the outcome model, I used the following diffuse priors,

\[ \beta \sim N(0, 10^2) \]

---

\(^{52}\) See Appendix H to find how that truncation is done in WinBUGS.
Here I have reduced the standard deviation of the $\beta$ from 100 to 10 to make the model converge faster. In addition, I have not used the gamma(.001,.001) diffuse prior for the precision terms as this may cause numerical instabilities. In the event that all cases $i$ are allocated to class $T_i = 1$, none of the $Y_i^*$’s would carry any information about $\sigma_2^2$ as $Y_i^* = \omega_i'$ for all $i$. Consequently the posterior and prior for $\sigma_2^2$ would be the same. The analogous scenario applies for $\sigma_1^2$. As there is a non-zero probability that all cases are allocated to the same class, a proper prior is needed. For the first part of the mixture model I used a gamma(2,100) parameterisation associated with a narrower distribution that could still be considered diffuse enough to account for the variability of the error term amongst cases affected by classical ME. In particular, such a gamma distribution has a mean of .02 and standard deviation of .014. Note that these two figures are expressed in terms of precision; in terms of standard deviations these are 7.1 and 8.4, respectively. For the second part of the model a more diffuse prior was assigned by changing the scale parameter to 500, since the variance of the second part of the model could be expected to be higher. The model is represented graphically in Figure 39.

**Figure 39. Adjustment using a validation subsample and a mixture model**

\[
\frac{1}{\sigma_1^2} \sim \text{gamma}(2,100)
\]

\[
\frac{1}{\sigma_2^2} \sim \text{gamma}(2,500)
\]
In order to assess the effectiveness of the method I have also included in Table 43 the results obtained for the benchmark model (described in Section 5.3.1 as the naïve model where the observed durations are substituted by the true ones available in the validation dataset). As we can see in Table 43, the inclusion of 10% of the validation cases managed to reduce the bias observed in every estimate, although all of those that were found non-significant in the naïve model remained as such in the benchmark model. The adjustment using a mixture model and the same 10% validation cases could not reverse this situation either. However, the reduction of the ‘bias’ affecting each regression estimate is more pronounced, which points to a relatively successful adjustment.

<table>
<thead>
<tr>
<th></th>
<th>Register</th>
<th>Survey</th>
<th>Benchmark</th>
<th>Adjustment</th>
</tr>
</thead>
<tbody>
<tr>
<td>constant</td>
<td>9.10 (1.36)</td>
<td>5.12 (1.15)</td>
<td>6.30 (1.05)</td>
<td>7.59 (1.70)</td>
</tr>
<tr>
<td>age</td>
<td>-0.088 (.038)</td>
<td>-0.001 (.032)</td>
<td>-0.028 (.029)</td>
<td>-0.053 (.047)</td>
</tr>
<tr>
<td>experience</td>
<td>-1.39 (.50)</td>
<td>-1.13 (.42)</td>
<td>-0.50 (.39)</td>
<td>-0.94 (.63)</td>
</tr>
<tr>
<td>age*exp</td>
<td>0.038 (.014)</td>
<td>0.002 (.012)</td>
<td>0.012 (.011)</td>
<td>0.023 (.017)</td>
</tr>
<tr>
<td>$\pi$</td>
<td></td>
<td></td>
<td></td>
<td>.52 (.06)</td>
</tr>
<tr>
<td>$\sigma_1$</td>
<td></td>
<td></td>
<td></td>
<td>4.29 (.94)</td>
</tr>
<tr>
<td>$\sigma_2$</td>
<td></td>
<td></td>
<td></td>
<td>69.2 (5.1)</td>
</tr>
<tr>
<td>$\theta$</td>
<td></td>
<td></td>
<td></td>
<td>99.7 (7.6)</td>
</tr>
</tbody>
</table>

In Figure 40 results for both the adjustment and the benchmark in terms of R.R.BIAS and R.R.RMSE (equation 5.1) have been plotted. Here we can see that, on average, the adjustment manages to reduce the bias found in the naïve model to 38.9% of its size whereas the benchmark analysis still retains 70.5% of the bias. This is quite a remarkable adjustment considering the precedents from previous methods explored and taking into account that only 10% of the PRESO observations have been used.

Not even the most successful of the RC and MI adjustments using a logit model achieved adjustments of the magnitude found here. In fact, had I chosen to use the standard linear RC approach, a validation subsample of 50% would have been necessary to achieve similar bias reductions to what we have seen here. Furthermore, the effectiveness of the adjustment in terms of R.R.RMSE is even higher, with average reductions of 65% the size of the RMSE in the naïve model, while the benchmark could only achieve reductions of 27.1% and less.
To better understand the success of this adjustment we can also inspect the estimates that define the mixture model. For example, it is interesting to note that the first part of the mixture model makes a relatively irrelevant contribution to the adjustment. It only increases the variability of the observed durations by 4.29 days on average, thus accounting for durations correctly reported or those affected by minor misdates. However, the fact that the model can differentiate between these cases and those much more seriously affected by ME is critical for the success of the adjustment.

In Section 3.4 we saw that, in the original sample, 43% of the cases could be considered to be correctly reported or affected by classical ME. Using the 10% validation subsample, the mixture model estimates that the proportion (i.e. \( \hat{\pi} \)) is 52%. The other 48% of unobserved PRESO cases have been drawn from what could be considered as a mixture of two normal distributions with mean 99.7 (\( \theta \) in Table 43) and standard deviation 69.2 (\( \sigma_2 \) in Table 43). The distribution of mean predictions of the unobserved PRESO durations (those not included in the validation subsample) from these two mechanisms is shown in Figure 41, together with the PRESO and LSA distributions for the same cases.
We can see that the majority of cases are predicted to have durations ranging from 100 to 300, although some others are also predicted to the right of the original window of observation (>395 days). This latter bump can also be observed in the posterior distributions of a majority of the predicted durations. Figure 42 shows a sample of four of them.

To assess the robustness of the adjustment, two sensitivity analyses were carried out. First, in spite of having used a considerably diffuse prior for $\theta$, I replicated the model by changing the mean of its prior from 187 to 600, a plausible value for durations that were right-censored. Results from this replication were practically identical with estimates from the outcome model changing in no more than a
hundredth part; the only noticeable difference was the value of $\theta$, which went from 98 to 105. Second, to assess the extent to which the quality of the adjustment is sample dependent, I replicate the analysis using a different 10% validation subsample taken at random.

**Table 44. Means (and medians for durations) of variables included in the second subsample**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Original</th>
<th>Validation.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>age</td>
<td>37</td>
<td>33.2</td>
</tr>
<tr>
<td>experience</td>
<td>2.6</td>
<td>2.8</td>
</tr>
<tr>
<td>PRESO duration</td>
<td>253</td>
<td>288</td>
</tr>
<tr>
<td>LSA duration</td>
<td>92</td>
<td>85</td>
</tr>
</tbody>
</table>

From Table 44 we can see that the alternative validation subsample that I used is also substantially similar to the original sample. The median of PRESO durations is now 35 days longer in the 10% validation subsample than in the original sample. However, the median for LSA durations is only 7 days shorter. Results for this adjustment using an alternative validation subsample are shown in Table 45 while its effectiveness in terms of R.R.BIAS and R.R.RMSE is shown in Figure 43.

**Table 45. Adjustment using a mixture model and a different validation subsample**

<table>
<thead>
<tr>
<th></th>
<th>Register</th>
<th>Survey</th>
<th>Benchmark.2</th>
<th>Adjustment.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>constant</td>
<td>9.10 (1.36)</td>
<td>5.12 (1.15)</td>
<td>6.59 (1.13)</td>
<td>7.027 (1.42)</td>
</tr>
<tr>
<td>age</td>
<td>-.088 (.038)</td>
<td>-.001 (.032)</td>
<td>-.037 (.032)</td>
<td>-.042 (.040)</td>
</tr>
<tr>
<td>experience</td>
<td>-1.39 (.50)</td>
<td>-.13 (.42)</td>
<td>-.53 (.42)</td>
<td>-.63 (.53)</td>
</tr>
<tr>
<td>age*exp</td>
<td>.038 (.014)</td>
<td>.002 (.012)</td>
<td>.013 (.017)</td>
<td>.016 (.015)</td>
</tr>
<tr>
<td>$\pi$</td>
<td></td>
<td>.56 (.06)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\sigma_1$</td>
<td></td>
<td>4.7 (1.0)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\sigma_2$</td>
<td></td>
<td>58.7 (5.1)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\theta$</td>
<td></td>
<td>89.5 (7.6)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The effectiveness of the adjustment is lower both in terms of R.R.BIAS and R.R.RMSE. However, the adjustment is still positive since it manages to reduce the ‘bias’ and to show more accurate standard errors than what was observed in the naïve and the benchmark analysis.
In spite of not having been able to generate an iterative process capable of drawing different validation subsamples (such as the one used for the RC and MI adjustments in Section 5.3.1), results presented here for two subsamples picked at random and representative of the original sample, show that using a mixture model is a promising method to adjust for event-occurrence ME in EHA models. The main limitation is the necessary access to validated data. Here a 10% validation subsample was required, 38 cases in total. Although small, this figure is still more than what most researchers could access.

The last of the adjustments implemented aims to eliminate the requirement of having access to any validation data. This can be achieved using more informative priors. In particular, here I compare results that could be obtained if the researcher assumed that the probability of cases being affected by each of the two parts of the mixture model was fixed. Table 46 and Figure 44 show results of the adjustments and their effectiveness, when that probability is assumed to be .3, .5 and .7.

### Table 46. Adjustment using mixture models with strong priors

<table>
<thead>
<tr>
<th></th>
<th>Register</th>
<th>Survey</th>
<th>( \pi = .3 )</th>
<th>( \pi = .5 )</th>
<th>( \pi = .7 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>constant</td>
<td>9.10 (1.36)</td>
<td>5.12 (1.15)</td>
<td>6.84 (2.46)</td>
<td>6.83 (2.16)</td>
<td>6.60 (1.93)</td>
</tr>
<tr>
<td>age</td>
<td>-.088 (.038)</td>
<td>-.001 (.032)</td>
<td>-.085 (.069)</td>
<td>-.078 (.062)</td>
<td>-.065 (.058)</td>
</tr>
<tr>
<td>experience</td>
<td>-1.39 (.50)</td>
<td>-.13 (.42)</td>
<td>-.93 (.89)</td>
<td>-.86 (.79)</td>
<td>-.73 (.69)</td>
</tr>
<tr>
<td>age*exp</td>
<td>.038 (.014)</td>
<td>.002 (.012)</td>
<td>.029 (.025)</td>
<td>.027 (.022)</td>
<td>.023 (.020)</td>
</tr>
<tr>
<td>( \sigma_1 )</td>
<td>1.65 (.61)</td>
<td>1.63 (.60)</td>
<td>1.63 (.61)</td>
<td>1.63 (.61)</td>
<td>1.63 (.61)</td>
</tr>
<tr>
<td>( \sigma_2 )</td>
<td>125.8 (6.5)</td>
<td>127.5 (8.4)</td>
<td>126.3 (12.3)</td>
<td>126.3 (12.3)</td>
<td>126.3 (12.3)</td>
</tr>
<tr>
<td>( \theta )</td>
<td>186.6 (9.8)</td>
<td>218.5 (13.1)</td>
<td>254.4 (20.4)</td>
<td>254.4 (20.4)</td>
<td>254.4 (20.4)</td>
</tr>
</tbody>
</table>
The first thing to notice is that the three models succeeded in reducing the ‘bias’ seen in the naïve analysis, although the standard errors were much higher than in the true model. Given that the proportion of cases identified as being affected by classical ME in the sample used here was 43% it could have been expected that the best adjustment of the three presented here would have been that with $\pi = .5$. This is not the case in terms of R.R.BIAS since the best adjustment was achieved using $\pi = .3$ – with a 70% reduction of the average bias seen in the naïve analysis. Neither is it the best adjustment in terms of R.R.RMSE, where the adjustment assuming $\pi = .7$ obtained the standard errors that are closest to those from the true model.

**Figure 44. Effectiveness of the adjustment using a mixture model with fixed $\pi$**

![Bar charts showing R.R.BIAS and R.R.RMSE for different models and values of $\pi$.]

Compared to the previous adjustments using a 10% validation subsample we can see how these adjustments assuming fixed $\pi$ can be more effective in terms of R.R.BIAS. The average reduction of the R.BIAS is higher in each of the three models used with $\pi$ fixed than in the model using the second validation subsample; for $\pi = .3$ and $\pi = .5$ it is more effective than for the model using the first validation subsample. On the other hand, in terms of R.R.RMSE, none of the models explored with $\pi$ fixed have been able to perform better than any of the two models specified using a 10% validation subsample. Since this last measure of the impact of ME covers the effect on standard errors it could be considered a more encompassing estimate of the quality of the adjustment. Hence, when validation data is available it would be advisable to carry out the previous adjustments using diffuse priors.
5.4.2. Conclusions from the Bayesian Adjustments

Thanks to the flexibility of Bayesian models, I have been able to carry out a variety of adjustments based on different requirements of access to validation data or assumptions about the distribution of the ME. For the first of the Bayesian adjustments presented here, I assumed that LSA durations were affected by a classical additive ME process with known variance. For the second adjustment, I assumed that the unobserved PRESO durations were MAR and that they could be predicted using a linear model, the fully observed LSA durations, and a validation subsample 25% the size of the original sample.

After implementing these two adjustments I found roughly similar results to those obtained when I implemented SIMEX, RC and MI in Sections 5.2 and 5.3: adjustments relying on the assumption of classical ME do not account for event-occurrence ME, whereas those relying on a linear imputation model can help although they are relatively ineffective and require substantial amounts of validated data. These results reinforce the arguments set out in the previous sections regarding the inadequacy of using standard methods for the adjustment of event-occurrence ME. On the other hand, it is important to note that unlike the SIMEX, RC, and MI methods, the adjustments presented here have not had to rely on any second-stage process such as bootstrap to obtain measures of uncertainty as they can be directly calculated from the posterior distribution, making the analysis simpler.

The value of the flexibility of Bayesian methods was more apparent when I implemented an alternative adjustment aiming to account for the different patterns of ME seen in event-occurrence questions on unemployment using a mixture model. One part of the model accounts for the less serious classical ME mechanism affecting some of the durations, while the second part accounts for reported durations that are not related to their true PRESO values, and which are responsible for the biggest part of the ME impact in the regression estimates. To inform the mixture model about the proportion of cases that needed to be predicted according to each of the mechanisms, I used 10% validation subsamples or informative priors.

This approach performed remarkably well. It captured the small differences between LSA and PRESO for some cases and remained completely agnostic about the rest, although the quality of the adjustment varied according to the configuration of the
validation subsample. For the first validation subsample that was used the adjustment managed to reduce the R.BIAS and R.RMSE for the regression estimates of the outcome model by 61% and 70% on average. A replication of the analysis using a different validation subsample reduced the R.BIAS and R.RMSE in 43% and 54% on average.

Here I have been able to carry out the adjustment thanks to a small validation subsample of true durations of unemployment provided by the Swedish register of unemployment. However, such access to administrative data is normally not possible, which makes the adjustment not generalizable as it is.

Nonetheless, in the last part of this section, I showed that the adjustment using a mixture model could be replicated when no validation data is available by using strong priors for the probability of cases being predicted according to the first or the second part of the mixture model. In particular, by fixing that probability to .3, .5, or .7, I found that the R.BIAS could be reduced by an average of 70%, 65%, and 54%, respectively. However, the posterior standard deviations from these models were substantially affected, showing an inverse relationship between the value of the mixing proportion used and the spread of the posterior distributions. This issue contributed to obtain less promising adjustments in terms of R.RMSE; for probabilities fixed at .3, .5, and .7 the R.RMSE was reduced by 25%, 41%, and 47%, respectively.

In conclusion, due to the complexity of the ME observed in retrospectively collected work-histories standard adjustments assuming classical ME are entirely inadequate, whereas those based on predicting unobserved cases using a validation subsample such as RC or MI are relatively ineffective. Instead models that take into account the complexity of the ME processes arising in event-occurrence questions on unemployment should be used. The combination of mixture models and Bayesian statistics appear to be a potentially powerful tool to do this. In the final chapter I discuss some other alternative solutions that could be explored in the future using Bayesian statistics.
CHAPTER 6. CONCLUSION

The assumption of variables being perfectly measured is one that a vast majority of quantitative researchers take for granted. This position is often difficult to maintain but in practice it is hardly questioned; probably as a result of lack of awareness about both the potential consequences for the validity of the analysis and the methods that can be implemented to attenuate the problem. The assumption of perfectly measured variables is especially difficult to justify by researchers using survey data. Due to problems of interviewee fatigue, misunderstanding, social desirability, memory failures, lack of cooperation, or plain fraud, the presence of ME in survey data is undeniable. The question that researchers using survey data should ask what the extent of ME is, rather than to question its presence.

As was theoretically and empirically demonstrated in Section 1.1.2 and Chapter 4, the consequences of using variables measured with errors are often unpredictable. Hence, covering any findings using such data with a halo of scepticism; or as Nugent et al. (2000) put it “Measurement error is, to borrow a metaphor, a gremlin hiding in the details of our research that can contaminate the entire set of estimated regression parameters.” (Nugent, et al. 2000, p. 60). Nonetheless, in spite of such potential dangers, the reliance of surveys as a research tool is immense.

Survey data is used as the main research input in fields as diverse as immigration, medicine, or marketing; and even if we can observe a transition towards the use of alternative sources of data (take the example of studies exploiting the possibilities of Big Data53), in many disciplines survey research is both necessary and unavoidable (such as research on attitudes, or psychological disorders, for example). This ubiquity of survey data, its propensity to be affected by ME, and the previously mentioned dire consequences derived from its naïve use, entails a research problem of great importance; a problem that could be endangering the validity of a great proportion of current quantitative research production.

Further research on the topic of ME in surveys can contribute to tackling this problem on two fronts. First, the occurrence of ME could be minimised through the implementation of better questionnaires. Second, for data that has already been

53 For example, Ansolabehere and Hersh (2012) used data from sources different from surveys to demonstrate how voting turnout tends to be overreported.
generated, a better understanding of the ME mechanisms in place could help to mitigate its consequences through the implementation of better statistical adjustments. “Although it is impossible to eliminate all errors, it is possible to use a fuller understanding of measurement error in designing research, analyzing and interpreting data, and acknowledging limitations.” (Viswanathan, 2005, preface).

Such has been the aim of my research. Specifically, I have studied the ME generated when work-histories are collected using retrospective questions. These types of questions have been heavily used by sociologists (e.g. Gash, 2008) and economists (e.g. Katz and Meyer, 1990) to investigate the transitions occurring between different work-statuses and the factors that cause these transitions. As they only require contacting respondents once – unlike prospective designs where respondents are contacted repeatedly over a period of time -, they have a distinct advantage in that they are cheaper to administer, do not suffer from attrition, and are capable of detecting short transitions. Their main downside is also a result of this single contact, they require the recall of events that took place in the past, hence, making retrospective questions notoriously prone to ME due to memory failures.

In spite of the important implications of this problem, just as for the general case of ME in survey data, ME in retrospective questions has not been sufficiently studied. So far only two studies, Mathiowetz and Duncan (1988), and Pyy-Martikainen and Rendtel (2009), have looked at the prevalence of ME in retrospective questions on work histories using validation data from official records. From these two studies, only the latter extended its analysis to investigate the consequences of using work histories retrospectively reported in EHA models, while the few studies that have explored the effectiveness of statistical adjustments have had to rely on simulated data where the ME process is hypothesised. The uniqueness of my thesis stems from having integrated the study of the three dimensions (prevalence, impact, and adjustment) of the problem of ME in retrospectively reported work histories using validation data from the Swedish register of unemployment.
6.1. Findings

The use of a research design based on a validation sample has revealed to be key to understand some major aspects of the ME processes that take place in retrospective questions about past work histories using an event-occurrence design. That is, where respondents need to recall work status experienced before the interview, and date them in chronological order. Unlike research designs relying on replicated data – which are predominant in the literature - I have been able to detect some systematic aspects of the ME process such as the general underreporting of days spent in unemployment by the long-term unemployed. Furthermore, thanks to the use of validation data I have been able to assess both the impact of ME for the sample studied, and on the effectiveness of different adjustments.

6.1.1. Large Prevalence and Impact

Probably the main conclusion from this thesis is the considerable unreliability of work histories collected from retrospective questions using a one-year event-occurrence design. For example, only 54% of the subjects reported the correct number of spells of unemployment that they had experienced the previous year. Interestingly, the comparison of this finding to what is found in another retrospective question where an eleven year time frame was considered illustrates the prominence of memory failures as a cause of ME in retrospective questions. Specifically, the analysis of a question with such an extended recall period showed that only 7.5% of the respondents reported the correct number of spells of unemployment.

The extent to which these questions fail to provide accurate information is even starker when we consider not just the ME in the number of reported spells of a particular status, but in the duration of those spells, which ultimately is the information that will be used in the specification of many EHA models. The biggest cause of concern here is not just the ME arising as a consequence of the beginning or the end of those spells being misdated, but the more complex forms of ME that take shape when spells are omitted or misclassified. In particular, the omission of even the shortest spell could induce the elimination of a transition, and the artificial link of two different spells, while the inclusion of false positive cases implies the use of durations that are only formed of noise. Regrettably, the presence of these complex
forms of ME is quite widespread; for the main question of analysis where the recall time period was not much longer than a year, 42% of the sample reported durations within ±15 days of what was captured in the register data, whereas for the remaining 58% no apparent relationship could be elucidated between true and observed values.

Ultimately, we have seen that the consequences of using spells of unemployment captured by such a retrospective question are quite harmful. Specifically, the regression estimates for different variables included in EHA models were found to be biased towards the null by around 90% of their true size. This attenuation effect rendered statistically significant effects obtained for each of the variables included into non-significant relationships, hence, inducing type I errors. Accordingly, the validity of other studies having used these types of questions should be questioned, and the implementation of statistical adjustments becomes extremely relevant, while future survey designers should opt for alternative ways to retrieve data on individuals’ work histories.

6.1.2. Different Questions Need to Be Used

The analysis presented in this thesis should be replicated using different samples to be able to assess the extent to which the prevalence and impact of ME found here is due to the type of question used or to certain peculiarities of the sample used. However, given the remarkable extent of ME detected, it is clear that the survey question under study has failed to capture the true work histories.

One problem derived from the use of event-occurrence schemes is the possibility of respondents reporting two spells during the same timespan when they are unsure of the state that best define them. This was not a problem in LSA-2001, but in LSA-1993 24 spells were found to be duplicated. A multi-state framework is not affected by these problems as it forces respondents to choose only one state for each of the given time periods. Furthermore, the a priori main advantage of event-occurrence over multi-state designs, namely the possibility to date spells using days instead of months, turns out not to be useful if we take into account that about one third of the spell dates were rounded in LSA-1993.

But perhaps the differences between the results obtained from questions using event-occurrence and multi-state designs can be shown more clearly through the
comparison of the findings obtained here for LSA-1993 against those from Pyy-Martikainen and Rendtel (2009). The authors analysed work histories retrospectively collected under a similar time frame but using a multi-state question, and found biases no bigger than 30% the size of the true estimates when using the reported spells of unemployment as the response variable of EHA models. This much milder impact found by Pyy-Martikainen and Rendtel (2009) suggests that multi-state questions are more effective in capturing the true work histories. Arguably, this is due to the lower cognitive burden placed on the respondent. For this reason, even if multi-state questions are incapable of observing transitions taking place within months, they should be taken as the better of the two designs. However, more research comparing these two types of question is necessary since the observed differences in the levels of ME might be due to differences in the composition or the size of the sample used in Pyy-Martikainen and Rendtel (2009), or to its wider window of observations, which reduced the share of cases that were right-censored.

Ultimately, although milder, the impact of the ME observed by Pyy-Martikainen and Rendtel (2009) in multi-state questions is still substantial and should be a cause of concern. Given the serious consequences of the MC of spells on the quality of reported work histories, perhaps the use of simpler questions restricting the choice of work status should be recommended. Much like multi-state questions seeking to trade off accuracy in the spells’ start dates to reduce the overall presence of ME, perhaps it would be helpful to use questions where the number of categories from which the respondent can choose is reduced. For example, from the ten used in the European Community Household Panel (eight in LSA) to just three: employed, unemployed, and out of the labour force. Otherwise, if accuracy in both the reported dates and the different work statuses is necessary, recall periods shorter than a year should be considered. Furthermore, for the specific case of LSA, the addition of a question asking about the current employment to the questionnaire would be very useful. First, it could help respondents to rethink about the last spell reported, since as we saw in Section 3.5.3 there is still a substantial prevalence of ME even for spells that took place near the day of the interview. Second, the comparison of reports to the current state with the last spell reported in the retrospective question could offer very interesting insights regarding the reliability of these measures and facilitate the implementation of adequate adjustments.
6.1.3. The Data Needs to Be Treated Differently

In spite of being plagued with ME, the data generated by these retrospective questions still contains valuable information, which could be enhanced through the implementation of adjustment methods. However, in Chapter 5 we have seen that due to the complexity of the ME mechanisms in place most of the standard adjustment methods are ineffective. In particular, I demonstrated how methods based on the assumption that the ME is classical, such as SIMEX, can make things worse by increasing the size of the bias observed in the naïve analysis.

This demonstration is particularly relevant for two reasons: i) the scarcity of research on the adjustment of ME affecting longitudinal data; and ii) the fact that previous studies on the topic have relied on simulated data to assess the effectiveness of the adjustment methods (Holt et al., 1991, Augustin, 1999, Skinner and Humphreys, 1999, Cole et al., 2006, and Dumangane, 2007). All of these studies improved their external validity by considering different types of ME processes. However, most of them were elaborations of classical additive or classical multiplicative ME that could be related to the misreporting of start or end dates of a spell. Hence, they did not consider the possibility of more complex forms of ME arising in event-occurrence questions as a consequence of misclassified or omitted spells.

In Chapter 5 we observed that disregarding the possibility of more complex forms of ME can distort critically the effectiveness of the methods used. In particular, we have seen that methods assuming classical additive or multiplicative ME such as SIMEX or likelihood-based methods were entirely ineffective.

The implementation of methods that do not rely on simplistic assumptions about the distribution of the ME, but on subsamples of validation data for the variable that is prone to ME, such as RC and MI, show the potential to reduce the impact of ME. However, the substantial amount of validation data that is required, together with the limited effectiveness of the adjustment, - which could not reduce the bias in the regression estimates much more than if the study had been carried out with a combination of observed durations plus the true ones available from the validation subsample - makes these methods relatively ineffective. Arguably, this is due to their inadequacy in dealing with the complexity of the ME studied here, namely the
presence of multiple ME patterns caused by different forms of misreports (misdates, omissions, MCs) of work histories.

In order to attempt to account for the most prevalent ME mechanisms, I implemented a Bayesian mixture model. This solution allows differentiating between relatively well reported spells of unemployment and those that show no relationship between the ME and the true values. To inform the mixture model about the proportion of cases that needed to be predicted according to each of the mechanisms I first carried out the model using a 10% validation subsample. This approach performed remarkably well as, for the average of the regression estimates of the outcome model, the adjustment managed to reduce the size of the bias by 43% to 61%, according to the configuration of the validation subsample being used.

However, although quite promising, this adjustment will not be easily reproducible by other researchers as very few of them will be able to have access to administrative data from which to derive the required validation subsample. Thanks to the flexibility of Bayesian methods, I was also able to explore adjustments that substitute this dependence on validation data by using informative priors for some of the parameters of the model. Specifically, I implemented a similar mixture model where the probability of different cases to be adjusted by each of the two mechanisms was assumed to be known. The quality of the adjustment was not as high as for the model relying on validation data, and it varied according to the value at which that probability was set. However, for all of the values used the different adjustments manage to reduce the impact of ME considerably.

The possibility of using strong priors as a substitute of validation data illustrates the importance of the study of ME. The more we know about the ME mechanisms arising in different settings the more effective will be the adjustments. For example, thanks to what I have learnt about the ME arising in work histories captured using an event-occurrence scheme in Chapter 3, I was able to carry out the tailored adjustment using a mixture model. Interestingly, this adjustment was relatively effective when other standard methods used in the literature relying on standard assumptions about the way the ME is distributed fail.

When using information obtained from studies on ME that are not based on the same sample for which the adjustments will be implemented there is always the risk that
the assumption of transportability will not hold. That is, the ME processes although similar will probably never be exactly the same due to issues such as inconsistencies on the wording of the question or sampling variability. The best way to be protected against breaches of the assumption of transportability is to carry out sensitivity analyses, and to rely on as wide a pool of evidence as possible, which once again highlights the importance of directing additional research efforts to this field. Alternatively, another way around the problem of register data being unavailable could be the use of auxiliary data, which nowadays is much more ubiquitous. In Section 6.3, a possible upgrade of the mixture model using Bayesian statistics is explained.

6.2. Generalisations and Caveats

The worrying results found here beg the question of how much can they be generalised to other retrospective questions on different life-course events. It is likely that reports on any past events will always be prone to memory failures and consequently to ME. However, there are reasons to believe that work histories are probably one of the most difficult life-course events to report, and as such we should not expect the levels of ME seen here in retrospective reports on different topics.

In particular, we have seen that one of the major reasons why retrospectively reported work histories are so problematic is due to the consequences of misclassifying work status, which is a very common problem. For example, the definition of unemployment requires that the subject must meet certain conditions, such as being actively looking for work. However, what different people consider “actively looking” varies, hence it is quite normal to find subjects who are out of the labour force defining themselves as unemployed.

Extended recall periods will probably increase the chance of misclassifying work status, but we would still find much MC even in questions referring to the current work status. For example, Poterba and Summers (1984) found that after repeating a set of questions on the same subjects 10.4% of those initially reporting to be unemployed defined themselves to be out of the labour force in a second interview. This fact makes much of the ME observed here a product of the topic being reported.
and not just a consequence of memory failures stemming from the retrospective design.

Two factors that are clearly responsible for the memory failures observed in retrospective questions are the levels of saliency and interference of the events reported. The perception and composition of work histories varies substantially across subjects. However, on average, we could expect the report of work histories to be defined by lower levels of saliency and higher levels of interference than other life-course events that are normally captured using retrospective questions such as housing history or marital life.

In conclusion, it is difficult to think of other life-course events that are more commonly misunderstood, more subject to short transitions, and less salient than work histories, which leads me to think that the extent of ME in the retrospective report of different life-course events will be lower than what I have detected here for the case of work-histories. Obviously this deduction is subject to the use of a sensible recall period. Reports on marital lives might not be too much affected by the recall period used due to the high saliency of their occurrence. However, reports of events roughly as salient as work histories, as could be the case for housing histories, might find their quality affected when recall periods are expanded beyond one year.

6.2.1. Caveats

So far, in this concluding chapter, I have summarised the most important findings of my research and discussed their external validity, i.e. the extent to which the findings are generalizable to other settings. I will now end this section by discussing the internal validity of those findings, i.e. the extent to which the results are robust.

I have argued that the prevalence of ME could be expected to be lower both for retrospective reports of unemployment captured under a different type of question and for different types of life-course events. In addition to that, there are reasons to believe that the presence of ME detected in the analysis presented here may have been exaggerated. This suspicion is mainly related to the validation research design and its reliance on the assumption of data from the PRESO register being a gold standard. This assumption implies considering work histories from PRESO as the true values, and differences with the reported work histories as evidence of ME in
the survey reports. However, the assumption of PRESO being a gold standard can be challenged from different directions.

In Chapter 2 I presented evidence on the PRESO dataset being affected by what could be understood as coding errors. In particular, during the exploratory analysis of PRESO I detected 72 work histories containing at least one spell that had been dated to start before the previously recorded spell. These cases were dropped from the study. In comparison with the 13,478 episodes recorded in PRESO for the 1990 to 2001 period they only represented .5% of the spells registered by the remaining sample of 719 subjects. However, they serve as evidence to prove the existence of recoding errors in PRESO, which could be even more prevalent if we take into account that other types of coding errors might have gone undetected. In addition, we could also expect registers of unemployment to contain false positive spells of unemployment as a result of fraudulent practices. For example, some subjects may have sought to become registered as unemployed in order to claim benefits when in reality they have been employed elsewhere or not actively looking for work. Hence, at least some of the ME observed in the report of unemployment might be due to the incapacity of unemployment registers to capture genuine spells of unemployment.

The reliability of the findings presented here might also be affected by the relatively low sample sizes that I have used. Across the different analyses that I carried out the sample size varied around 400 subjects, for which we could expect considerable sampling variability. In addition, it is important to highlight that although the sample was randomly picked, it was composed of subjects who were unemployed at a particular date and who gave permission for retrieving their work histories from the register. Hence, the sample is not perfectly representative of the Swedish work-force, which could have implications on the extent of ME that was detected. For example, in Chapter 3 we saw that the long-term unemployed tend to underreport their reports of time spent unemployed.

Lastly, the way I have treated both the LSA and PRESO datasets may have influenced the results presented here. Some decisions that I have taken in my analyses may have had an effect on the findings presented here. For example, the choice made in Chapter 4 to restrict the window of observation so it could start from 28th February probably exaggerated the impact of ME that would have been observed if the whole window of observation had been used. As we saw in Figure
13, where the prevalence of ME in person-day observations was plotted, it was the first spells that were more accurately reported. However, to comply with my intention of analysing subjects who were known to start from a state of unemployment I did not use the first 59 days reported.

6.3. Further Analyses

In this thesis I have aimed to study the different aspects of ME in survey data in order to describe the problem and offer solutions in the most complete way possible. To do so I have carried out separate analyses on the prevalence, implications, and adjustments of ME in retrospectively collected work histories. However, as it is clear from what has been discussed previously in this concluding chapter, more research is necessary to assess the extent to which my findings are comparable to other studies on the ME stemming from reports of work histories, or generalizable to other retrospective questions. Moreover, in addition to the replication of this study on similar questions, there are still some key areas that are in need of further research. In particular, to assess the validity of studies based on validation designs more precisely a better understanding of the ME present in register datasets is essential, whereas to improve the validity of studies relying on survey data better adjustment methods are necessary.

6.3.1. The Study of Measurement Error in Register Data

The violation of the gold standard assumption in validation studies implies that a share of the observed differences between the register and survey datasets are wrongly attributed to ME in the latter. Hence, results from these studies that do not question the gold standard assumption will systematically overestimate the presence of ME in survey data. Here I have acknowledged the violation of that assumption; however, it would be more informative to accompany that acknowledgement with an estimate of the prevalence of ME in the register data.

Estimates of the prevalence of ME in register data could be used to convey a measure of uncertainty for the findings from validation studies. We could take the findings obtained under the assumption of the register being a gold standard as the
upper bound of a confidence interval for the prevalence of ME in survey reports. The lower bound could be estimated by rerunning the validation analysis but taking a number of cases from the survey to be the true values, with the specific number of cases being proportional to the estimated levels of ME affecting the register\textsuperscript{54}.

The improvement of our understanding on the validity of administrative data is an area of paramount importance towards which more research efforts should be directed. At the beginning of this chapter I indicated how, given the ubiquity of survey data in research, it is crucial to understand and adjust for the ME that stems from it. However, that will not be completely possible unless we can obtain a clearer picture of the extent to which datasets taken as gold standards are affected by ME.

### 6.3.2. Additional Adjustments

The mixture model presented in Section 5.4 has shown promising results, although it is still a rather rough adjustment. It is a great proof of concept - as it acknowledges the need to adjust for different ME patterns simultaneously - upon which better adjustments could be built. In particular, the probability of cases being considered by each of the parts of the mixture model could be made conditional on certain factors known to be associated with the omission or MC of spells of unemployment. For example, in Chapter 3 we observed that the probability of omitting spells of unemployment increased when the interview was taken on the phone rather than face-to-face. The DAG of such a model is shown in Appendix I (Figure A10).

The use of paradata for the adjustment of ME has already been explored in the literature. For example, see Da Silva and Skinner (2014), where the authors used a binary variable capturing whether participants in a survey looked at their payslip to remember their current earnings in order to adjust for heteroscedastic ME in those reports. Given the ever increasing presence of paradata in surveys, similar types of variables related to the work history process could also be used to inform the probability of omission or MC of spells.

\textsuperscript{54} The research design that I propose here resembles that used by Kapteyn and Ypma (2007) where register data on earnings, pensions and taxes was allowed to be prone to ME problems in the form of mismatched records. Here the authors traced out algebraically the impact of such ME when the register data is used as the dependent variable in a linear model.
An alternative approach involves the consideration of a proportional odds logit model as the outcome model. This way, the spells of unemployment will be specified as person-day observations and the ME problem becomes one of MC. As was explained in Section 1.2.3, such a model assumes that the conditional probability of a person-day case being unemployed can be expressed by a linear combination of variables of interest (in this case age, experience, and their interaction effect), and a sequence of fixed effects accounting for the different time periods (given how computationally intensive the model is as a result of the more than 245,606 person-day observations contained in the sample I will only include fixed effects at the month level rather than at the day level).

This would reproduce the PO logit naïve analysis (defined in equation 1.40). In order to adjust for the MC affecting the response variable I will specify a new conditional probability as a combination of true but unobserved conditional probability of a case being unemployed and the probabilities of sensitivity and specificity. Sensitivity and specificity could be introduced as informative beta priors with mean equal to the values estimated from the validation study. A DAG depicting how the model could be structured is shown in Appendix I (Figure A11).

A similar approach has been implemented by Cheng et al. (2010) for a simpler logit model where the response variable was misclassified. Here the novelty stems from the control of the longitudinal component using fixed effects. In addition, once this adjustment is consolidated, it would be possible to add lags for the probabilities of sensitivity and specificity. This way, I could account for quadratic effects on the probability of MC across the window of observation, just as it was found in Mathiowetz and Duncan (1988) and in Table 10 in Section 3.5.3.

This approach has the potential to be more widely applicable for the adjustment of different life-course events than the one using a mixture model, as the latter has been specifically tailored to account for cases where due to omissions or MCs of spells the observed values are unrelated to the true ones. However, for more salient life-course events such as housing or marital histories, it could be expected that the ME problem will be one of the misdating spells, and as such this last adjustment might offer better results.
References


Appendix A: Descriptive Statistics for the Samples Used

Across all the tables presented in this appendix I note with an asterisk (*) variables capturing durations that will be represented by their median instead of their mean.

**Table A1. Descriptive statistics for sample a (Section 3.4)**

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>SD</th>
<th>Min.</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>36.8</td>
<td>8.5</td>
<td>26</td>
<td>55</td>
</tr>
<tr>
<td>Female</td>
<td>0.3</td>
<td>0.4</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Phone interview</td>
<td>0.1</td>
<td>0.3</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Experience</td>
<td>2.6</td>
<td>0.6</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>Spells of unemp.</td>
<td>1.7</td>
<td>0.7</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>Cumulative unemp.*</td>
<td>384</td>
<td>205</td>
<td>25</td>
<td>1237</td>
</tr>
<tr>
<td>Spell length*</td>
<td>213</td>
<td>168</td>
<td>9</td>
<td>1237</td>
</tr>
</tbody>
</table>

**Table A2. Descriptive statistics for sample b (Section 3.4)**

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>SD</th>
<th>Min.</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>37.3</td>
<td>8.8</td>
<td>26</td>
<td>55</td>
</tr>
<tr>
<td>Female</td>
<td>0.2</td>
<td>0.4</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Phone interview</td>
<td>0.1</td>
<td>0.3</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Experience</td>
<td>2.6</td>
<td>0.5</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>Spells of unemp.</td>
<td>1.8</td>
<td>0.7</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>Cumulative unemp.*</td>
<td>407</td>
<td>203</td>
<td>25</td>
<td>1237</td>
</tr>
<tr>
<td>Spell length*</td>
<td>247</td>
<td>155</td>
<td>9</td>
<td>1215</td>
</tr>
</tbody>
</table>

**Table A3. Descriptive statistics for sample c (Section 3.5.1)**

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>SD</th>
<th>Min.</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>36.5</td>
<td>8.4</td>
<td>26</td>
<td>55</td>
</tr>
<tr>
<td>Female</td>
<td>0.3</td>
<td>0.5</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Phone interview</td>
<td>0.1</td>
<td>0.4</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Experience</td>
<td>2.6</td>
<td>0.7</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>Spells of unemp.</td>
<td>1.7</td>
<td>.08</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>Cumulative unemp.*</td>
<td>379</td>
<td>207</td>
<td>25</td>
<td>1237</td>
</tr>
</tbody>
</table>

**Table A4. Descriptive statistics for sample d (Sections 3.5.2 and 3.5.3)**

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>SD</th>
<th>Min.</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>36.6</td>
<td>8.5</td>
<td>26</td>
<td>55</td>
</tr>
<tr>
<td>Female</td>
<td>0.3</td>
<td>0.5</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Phone interview</td>
<td>0.1</td>
<td>0.4</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Experience</td>
<td>2.6</td>
<td>0.6</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>Spells of unemp.</td>
<td>1.7</td>
<td>0.7</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Cumulative unemp.*</td>
<td>359</td>
<td>205</td>
<td>25</td>
<td>1237</td>
</tr>
<tr>
<td>Timespan*</td>
<td>304</td>
<td>143</td>
<td>0</td>
<td>546</td>
</tr>
</tbody>
</table>
Table A5. Descriptive statistics for sample e (Sections 3.5.1 and 3.5.3)

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>SD</th>
<th>Min.</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>43.7</td>
<td>8.8</td>
<td>33</td>
<td>63</td>
</tr>
<tr>
<td>Female</td>
<td>0.3</td>
<td>0.4</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Phone interview</td>
<td>0.1</td>
<td>0.3</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Experience</td>
<td>2.6</td>
<td>0.7</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>Spells of unemp.</td>
<td>4.3</td>
<td>3.4</td>
<td>1</td>
<td>18</td>
</tr>
<tr>
<td>Cumulative unemp.*</td>
<td>15</td>
<td>16</td>
<td>0</td>
<td>95</td>
</tr>
<tr>
<td>Timespan*</td>
<td>74</td>
<td>42</td>
<td>0</td>
<td>149</td>
</tr>
</tbody>
</table>

Table A6. Descriptive statistics for sample f (Sections 4.2, 5.2, 5.3, and 5.4)

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>SD</th>
<th>Min.</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>37</td>
<td>8.7</td>
<td>26</td>
<td>55</td>
</tr>
<tr>
<td>Experience</td>
<td>2.6</td>
<td>0.6</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>PRESO durations*</td>
<td>253</td>
<td>143</td>
<td>1</td>
<td>395</td>
</tr>
<tr>
<td>LSA durations*</td>
<td>92</td>
<td>113</td>
<td>1</td>
<td>395</td>
</tr>
</tbody>
</table>

Table A7. Descriptive statistics for sample g (Section 4.3)

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>SD</th>
<th>Min.</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>37</td>
<td>8.7</td>
<td>26</td>
<td>55</td>
</tr>
<tr>
<td>Experience</td>
<td>2.6</td>
<td>0.6</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>PRESO durations*</td>
<td>180</td>
<td>140</td>
<td>1</td>
<td>395</td>
</tr>
<tr>
<td>LSA durations*</td>
<td>89</td>
<td>97</td>
<td>1</td>
<td>395</td>
</tr>
</tbody>
</table>
Appendix B: Normality of the Transformed Durations of Unemployment

Figure A1. Density functions of the difference in the registered and reported time spent in unemployment (left) and the squared root of that difference (right)
### Appendix C: Person-Day Mismatches

**Table A8. Crosstab of person-day cases for the complete categories of LSA-1993 and PRESO**

<table>
<thead>
<tr>
<th>PRESO LSA</th>
<th>Replacement scheme</th>
<th>Without job</th>
<th>Part-time employed</th>
<th>Temporary job</th>
<th>Permanent job</th>
<th>Public temporary job</th>
<th>Employability rehabilitation programme</th>
<th>Labour market training</th>
<th>Other</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Jobseeker</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>659 (21%)</td>
<td>124,562 (64%)</td>
<td>3,004 (31%)</td>
<td>2,032 (23%)</td>
<td>304 (19%)</td>
<td>852 (9%)</td>
<td>455 (24%)</td>
<td>1,921 (12%)</td>
<td>6 (1%)</td>
<td>133,795 (54%)</td>
<td></td>
</tr>
<tr>
<td><strong>Employee</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1,709 (54%)</td>
<td>43,253 (22%)</td>
<td>5,657 (58%)</td>
<td>5,287 (61%)</td>
<td>1,019 (63%)</td>
<td>8,164 (87%)</td>
<td>664 (34%)</td>
<td>1,126 (12%)</td>
<td>176 (20%)</td>
<td>67,055 (27%)</td>
<td></td>
</tr>
<tr>
<td><strong>Job training</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0 (0%)</td>
<td>6,646 (3%)</td>
<td>404 (4%)</td>
<td>168 (2%)</td>
<td>190 (12%)</td>
<td>330 (3%)</td>
<td>625 (32%)</td>
<td>12,424 (78%)</td>
<td>362 (41%)</td>
<td>21,149 (9%)</td>
<td></td>
</tr>
<tr>
<td><strong>Entrepreneur</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0 (0%)</td>
<td>7,504 (4%)</td>
<td>181 (2%)</td>
<td>260 (3%)</td>
<td>96 (6%)</td>
<td>0 (0%)</td>
<td>0 (0%)</td>
<td>0 (0%)</td>
<td>0 (0%)</td>
<td>8,041 (3%)</td>
<td></td>
</tr>
<tr>
<td><strong>Homeworker</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0 (0%)</td>
<td>1,542 (1%)</td>
<td>0 (0%)</td>
<td>53 (1%)</td>
<td>0 (0%)</td>
<td>17 (0%)</td>
<td>0 (0%)</td>
<td>2 (0%)</td>
<td>11 (1%)</td>
<td>1,625 (1%)</td>
<td></td>
</tr>
<tr>
<td><strong>Parental leave</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0 (0%)</td>
<td>4,379 (2%)</td>
<td>342 (3%)</td>
<td>175 (2%)</td>
<td>0 (0%)</td>
<td>0 (0%)</td>
<td>0 (0%)</td>
<td>113 (1%)</td>
<td>59 (7%)</td>
<td>5,068 (2%)</td>
<td></td>
</tr>
<tr>
<td><strong>Employment development</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>781 (25%)</td>
<td>53 (0%)</td>
<td>0 (0%)</td>
<td>1 (0%)</td>
<td>0 (0%)</td>
<td>18 (0%)</td>
<td>0 (0%)</td>
<td>0 (0%)</td>
<td>0 (0%)</td>
<td>853 (0%)</td>
<td></td>
</tr>
<tr>
<td><strong>Other</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0 (0%)</td>
<td>6,432 (3%)</td>
<td>94 (1%)</td>
<td>680 (8%)</td>
<td>0 (0%)</td>
<td>13 (0%)</td>
<td>182 (9%)</td>
<td>285 (2%)</td>
<td>261 (30%)</td>
<td>7,175 (3%)</td>
<td></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>3,149 (100%)</td>
<td>194,371 (100%)</td>
<td>9,682 (100%)</td>
<td>8,656 (100%)</td>
<td>1,609 (100%)</td>
<td>9,394 (100%)</td>
<td>1,926 (100%)</td>
<td>15,871 (100%)</td>
<td>875 (100%)</td>
<td>245,533 (100%)</td>
</tr>
</tbody>
</table>

*Each cell captures the absolute number of cases and between brackets the percentages of those cases over the column total (PRESO total for that category).*
Appendix D: Bootstrap

Bootstrap is a widely used technique to estimate the sampling variability from complex statistical models. It is normally used in two circumstances, when the regression assumption of normality of the residuals is violated, and for two-step processes such as SIMEX, RC, or other single imputation methods.

The logic of bootstrap is to empirically generate the sampling distribution of some estimates of interest ($\hat{\beta}_{adj}$) using a resampling mechanism. This is done in a process of three stages: i) a total of $M$ simulated dataset are generated through sampling with replacement from the original data; ii) the model of interest is re-estimated using each of the $M$ simulated datasets, so $\hat{\beta}_{adj}^m$ are obtained, where $m = 1, \ldots, M$; and iii) $\bar{\beta}_{adj}$, the average of $\hat{\beta}_{adj}^m$, is calculated so the empirical covariance matrix of $\hat{\beta}_{adj}$ can be estimated as follows,

$$\text{var}(\hat{\beta}_{SIMEX}) = (M - 1)^{-1} \sum_{m=1}^{M} (\hat{\beta}_{SIMEX}^m - \bar{\beta}_{SIMEX})^2$$

This process was slightly modified when I proceeded to apply bootstrap for the SIMEX and RC methods in Sections 5.2 and 5.3. In both applications I used $M = 100$ and the bootstrapping pairs algorithm\(^{55}\), where entire cases covering the response and explanatory variables are resampled with replacement. Arguably, $M = 100$ is a higher number than what is normally used in SIMEX (e.g. Nolte, 2007, used 50 bootstrap samples), but lower than what is customary in RC (the STATA help-file recommends using 200 to 1000 samples). Due to the iterative process carried out in the RC approach to explore different validation subsamples, a relatively low $M$ was necessary to keep such a computationally intensive process manageable.

Lastly, for the RC applications, steps 1 and 2 described above were further modified so the reliance on a validation subsamples could be covered by the resampling mechanism. A new sample, $b$, is taken from the validation subsample, $s$, to specify the calibration model and estimate its parameters, $\xi_{RC}$, then another sample, $t$, is taken from the main sample, $i$, to estimate $\hat{Y}$ using $\xi_{RC}$ and the data in $t$.

---

\(^{55}\) See Keele (2008) for a description of the differences between bootstrapping algorithms.
Appendix E: R Code for the SIMEX Process

#Comparison of the true and naïve outcome models.
library("survival")
surv.LSA = Surv(time=data$duration_LSA, event=data$censored_LSA)
surv.PRESO = Surv(time=data$duration_PRESO, event=data$censored_PRESO)
cox_LSA = coxph(surv.LSA ~ data$age+data$ageexp+data$experience)
cox_LSA
cox_PRESO = coxph(surv.PRESO ~ data$age+data$ageexp+data$experience)
cox_PRESO

#The classical additive process    #I run it for sd(u) equal to 20, 40 and 80.
#noise.5
noise.5 = matrix(c(0),nrow=10000,ncol=3,byrow=TRUE)
for(i in 1:10000)
  data$noise.5 = sqrt(.5)*rnorm(381,0,80)
data$duration_LSA.5 = data$duration_LSA + data$noise.5
data$duration_LSA.5 = ifelse(data$duration_LSA.5>395,395,data$duration_LSA.5)
data$censored = ifelse(data$duration_LSA.5>=395,0,1)
surv_noise.5 = Surv(time=data$duration_LSA.5, event=data$censored)
cox_noise.5 = coxph(surv_noise.5 ~ data$age+data$ageexp+data$experience)
noise.5[i,1] = coef(cox_noise.5)[1]
noise.5[i,2] = coef(cox_noise.5)[2]
noise.5[i,3] = coef(cox_noise.5)[3]
}

avg_noise.5_age = mean(noise.5[1,])
avg_noise.5_ageexp = mean(noise.5[2,])
avg_noise.5_exp = mean(noise.5[3,])
#noise1
noise1 = matrix(c(0),nrow=10000,ncol=3,byrow=TRUE)
for(i in 1:10000)
  data$noise1 = sqrt(1)*rnorm(381,0,80)
data$duration_LSA1 = data$duration_LSA + data$noise1
data$duration_LSA1 = ifelse(data$duration_LSA1>395,395,data$duration_LSA1)
data$duration_LSA1 = ifelse(data$duration_LSA1<0,0,data$duration_LSA1)
data$censored = ifelse(data$duration_LSA1>=395,0,1)
surv_noise1 = Surv(time=data$duration_LSA1, event=data$censored)
cox_noise1 = coxph(surv_noise1 ~ data$age+data$ageexp+data$experience)
noise1[i,1] = coef(cox_noise1)[1]
noise1[i,2] = coef(cox_noise1)[2]
noise1[i,3] = coef(cox_noise1)[3]
}

avg_noise1_age = mean(noise1[1,])
avg_noise1_ageexp = mean(noise1[2,])
avg_noise1_exp = mean(noise1[3,])
#noise1.5
noise1.5 = matrix(c(0),nrow=10000,ncol=3,byrow=TRUE)
for(i in 1:10000)
  data$noise1.5 = sqrt(1.5)*rnorm(381,0,80)
data$duration_LSA1.5 = data$duration_LSA + data$noise1.5
data$duration_LSA1.5 = ifelse(data$duration_LSA1.5>395,395,data$duration_LSA1.5)
data$duration_LSA1.5 = ifelse(data$duration_LSA1.5<0,0,data$duration_LSA1.5)
data$censored = ifelse(data$duration_LSA1.5>=395,0,1)
surv_noise1.5 = Surv(time=data$duration_LSA1.5, event=data$censored)
cox_noise1.5 = coxph(surv_noise1.5 ~ data$age+data$ageexp+data$experience)
noise1.5[i,1] = coef(cox_noise1.5)[1]
noise1.5[i,2] = coef(cox_noise1.5)[2]
noise1.5[i,3] = coef(cox_noise1.5)[3]

avg_noise1.5_age = mean(noise1.5[1,])
avg_noise1.5_ageexp = mean(noise1.5[2,])
avg_noise1.5_exp = mean(noise1.5[3,])
data$censored = ifelse(data$duration_LSA1.5>=395,0,1)
surv_noise1.5 = Surv(time=data$duration_LSA1.5, event=data$censored)
cox_noise1.5 = coxph(surv_noise1.5 ~ data$age+data$ageexp+data$experience)
noise1.5[1,1] = coef(cox_noise1.5)[1]
noise1.5[1,2] = coef(cox_noise1.5)[2]
noise1.5[1,3] = coef(cox_noise1.5)[3]
}

avg_noise1.5_age = mean(noise1.5[1,])
avg_noise1.5_ageexp = mean(noise1.5[2,])
avg_noise1.5_exp = mean(noise1.5[3,])

#noise2
noise2 = matrix(c(0),nrow=10000,ncol=3,byrow=TRUE)
for(i in 1:10000){
data$noise2 = sqrt(2)*rnorm(381,0,80)
data$duration_LSA2 = data$duration_LSA + data$noise2
data$duration_LSA2 = ifelse(data$duration_LSA2>395,395,data$duration_LSA2)
data$duration_LSA2 = ifelse(data$duration_LSA2<0,0,data$duration_LSA2)
data$censored = ifelse(data$duration_LSA2>=395,0,1)
surv_noise2 = Surv(time=data$duration_LSA2, event=data$censored)
cox_noise2 = coxph(surv_noise2 ~ data$age+data$ageexp+data$experience)
noise2[i,1] = coef(cox_noise2)[1]
noise2[i,2] = coef(cox_noise2)[2]
noise2[i,3] = coef(cox_noise2)[3]
}

avg_noise2_age = mean(noise2[1,])
avg_noise2_ageexp = mean(noise2[2,])
avg_noise2_exp = mean(noise2[3,])

#I put all the average coefficients in a same dataset.
avg_noiseADJ_age = NA
avg_noiseADJ_ageexp = NA
avg_noiseADJ_exp = NA
lambda = c(-1, 0, .5, 1, 1.5, 2)
addi1 = c(avg_noiseADJ_age, coef(cox_LSA)[1], avg_noise.5_age, avg_noise1_age, avg_noise1.5_age, avg_noise2_age)
addi2 = c(avg_noiseADJ_ageexp, coef(cox_LSA)[2], avg_noise.5_ageexp, avg_noise1_ageexp, avg_noise1.5_ageexp, avg_noise2_ageexp)
addi3 = c(avg_noiseADJ_exp, coef(cox_LSA)[3], avg_noise.5_exp, avg_noise1_exp, avg_noise1.5_exp, avg_noise2_exp)
SIMEX = data.frame(lambda, lambda2, addi1, addi2, addi3)
names(SIMEX) = c("lambda","lambda2","age","ageexp","exp")

#I obtain the adjusted estimate using a linear extrapolation function.
SIMEX_age = lm(SIMEX$age ~ SIMEX$lambda)
SIMEX[1,3] = coef(SIMEX_age)[1] - coef(SIMEX_age)[2]
SIMEX_ageexp = lm(SIMEX$ageexp ~ SIMEX$lambda)
SIMEX[1,4] = coef(SIMEX_ageexp)[1] - coef(SIMEX_ageexp)[2]
SIMEX_exp = lm(SIMEX$exp ~ SIMEX$lambda)
SIMEX[1,5] = coef(SIMEX_exp)[1] - coef(SIMEX_exp)[2]
Appendix F: Extrapolation Functions

Figure A2. Extrapolation functions for simulated classical additive ME
Figure A3. Extrapolation functions for simulated classical multiplicative ME
Appendix G: Visual Test of Convergence

Figure A4. The naïve AL exponential model

Figure A5. The true AL exponential model

Figure A6. Adjustment assuming classical ME
Figure A7. Adjustment assuming MAR and using a validation subsample

Figure A8. Adjustment using a mixture model and a validation subsample

Figure A9. Adjustment using a mixture model and strong priors
Appendix H: WinBUGS Code for the Bayesian Adjustments

The true outcome model

model{
  for(i in 1:381) {
    durPRESO[i] ~ dweib(1, mu[i]) I(cenPRESO[i],)
    log(mu[i]) <- -(bcons + bage*age[i] + bageexp*ageexp[i] + bexp*experience[i])
  }
  bcons ~ dnorm(0,.001)
  bage ~ dnorm(0,.0001)
  bexp ~ dnorm(0,.0001)
  bageexp ~ dnorm(0,.0001)
}
list(list(bcons=3, bage=0, bageexp=2, bexp=-1),list(bcons=5, bage=1, bageexp=1, bexp=1))

Adjustment assuming classical ME

model{
  for(i in 1:381) {
    durPRESO[i] ~ dweib(1, mu[i])
    log(mu[i]) <- -(bcons + bage*age[i] + bageexp*ageexp[i] + bexp*experience[i])
    durLSA[i] ~ dnorm(durPRESO[i],tau) I(cenLSA[i],)
  }
  bcons ~ dnorm(0,.0001)
  bage ~ dnorm(0,.0001)
  bexp ~ dnorm(0,.0001)
  bageexp ~ dnorm(0,.0001)
  tau <- .001
}
list(list(bcons=2, bage=.03, bageexp=-.03, bexp=1), list(bcons=2, bage=.01, bageexp=-.01, bexp=1))

Adjustment assuming MAR and using a validation subsample

model{
  for(i in 1:381) {
    durPRESO[i] ~ dexp(mu[i]) I(cenPRESO[i],)
    log(mu[i]) <- -(bcons + bage*age[i] + bageexp*ageexp[i] + bexp*experience[i])
    durLSA[i] ~ dnorm(omega[i], tau)
    omega[i] <- acons + adurPRESO*durPRESO[i]
  }
  bcons ~ dnorm(0,.0001)
  bage ~ dnorm(0,.0001)
  bexp ~ dnorm(0,.0001)
  bageexp ~ dnorm(0,.0001)
  acons ~ dnorm(0,.0001)
  adurPRESO ~ dnorm(0,.01)
  acenPRESO ~ dnorm(0,.0001)
}
tau ~ dgamma(.001,.001)
sigma <- 1/sqrt(tau)
}
list(list(bcons=5, bage=-1, bageexp=1, bexp=-1, acons=4, adurPRESO=.5, tau=.001),
list(bcons=3, bage=0, bageexp=2, bexp=-1,acons=3, adurPRESO=.9, tau=.05))

Adjustment using a mixture model and a validation subsample

model{
  for(i in 1:381) {
    durPRESO[i] ~ dexp(mu[i])I(cenPRESO[i],)
    log(mu[i]) <- -(bcons + bage*age[i] + bageexp*ageexp[i] + bexp*experience[i])
    durLSA[i] ~ dnorm(omega[i], tau)(cenLSA[i],)
    omega[i] <- T[i]*(durPRESO[i]) + (1-T[i])*theta
    T[i] ~ dbern(p)
  }
  bcons ~ dnorm(0,.01)
  bage ~ dnorm(0,.01)
  bexp ~ dnorm(0,.01)
  bageexp ~ dnorm(0,.01)
  tau ~ dgamma(2.0001,100)
  tau ~ dgamma(1.0001,500)
  sigma1 <- 1/sqrt(tau1)
  sigma2 <- 1/sqrt(tau2)
  theta ~ dnorm(187,.00001)
  p ~ dbeta(1,1)
  T[1] <- p
  T[2] <- 1-p
}
list(list(bcons=5, bage=1, bageexp=1, bexp=1, acons=4, adurLSA=.5, tau=.001),
list(bcons=3, bage=0, bageexp=2, bexp=-1,acons=3, adurLSA=.9, tau=.05))

Adjustment using a mixture model and strong priors

model{
  for(i in 1:381) {
    durPRESO[i] ~ dexp(mu[i])I(cenPRESO[i],)
    log(mu[i]) <- -(bcons + bage*age[i] + bageexp*ageexp[i] + bexp*experience[i])
    durLSA[i] ~ dnorm(omega[i], tau)(cenLSA[i],)
    omega[i] <- T[i]*(durPRESO[i]) + (1-T[i])*theta
    T[i] ~ dbern(p)
  }
  bcons ~ dnorm(0,.01)
  bage ~ dnorm(0,.01)
  bexp ~ dnorm(0,.01)
  bageexp ~ dnorm(0,.01)
  tau ~ dgamma(2.0001,100)
  tau ~ dgamma(1.0001,500)
  sigma1 <- 1/sqrt(tau1)
  sigma2 <- 1/sqrt(tau2)
  theta ~ dnorm(187,.00001)
  p ~ dbeta(1,1)
  T[1] <- p
  T[2] <- 1-p
  p <- .5
T[1] <- p
T[2] <- 1-p
}
list(list(bcons=5, bage=1, bageexp=1, bexp=1, acons=4, adurLSA=.5, tau=.001),
    list(bcons=3, bage=0, bageexp=2, bexp=-1,acons=3, adurLSA=.9, tau=.05))
Appendix I: DAGs for Future Adjustments

The first adjustment proposed in Section 6.3 recommends the use of auxiliary data and can be represented by the DAG in Figure 39 representing the simpler adjustment using a mixture model. The difference resides on including the variable for phone interviews, $p_{hi}$, and its regression estimate, $\gamma_{ph}$, as parents of the node for the probability of being part of the second ME mechanism, $\pi$.

Figure A10. Adjustment using a mixture model and paradata

The second adjustment suggests the use of known sensitivity (SN) and specificity (SP). As was explained in Section 1.2.3, we could formulate the outcome model using an EHA model for discrete data. Such a model assumes that the conditional probability $\pi_{ij}$ of a person-day case being unemployed, $Y_{ij}^* = 1$, can be expressed by a linear combination of variables of interest and a sequence of fixed effects. In order to adjust for the MC affecting the response variable, $Y_{ij}^*$, I will specify a new conditional probability, $\pi_{ij}^*$, as a combination of $\pi_{ij}$ and the probabilities of SN and SP as follows, $\pi_{ij}^* = \pi_{ij}SN + (1 - \pi_{ij})(1 - SP)$. 

The second adjustment suggests the use of known sensitivity (SN) and specificity (SP). As was explained in Section 1.2.3, we could formulate the outcome model using an EHA model for discrete data. Such a model assumes that the conditional probability $\pi_{ij}$ of a person-day case being unemployed, $Y_{ij}^* = 1$, can be expressed by a linear combination of variables of interest and a sequence of fixed effects. In order to adjust for the MC affecting the response variable, $Y_{ij}^*$, I will specify a new conditional probability, $\pi_{ij}^*$, as a combination of $\pi_{ij}$ and the probabilities of SN and SP as follows, $\pi_{ij}^* = \pi_{ij}SN + (1 - \pi_{ij})(1 - SP)$.
Figure A11. Adjustment using person-day cases and known SN and SP