The Cathie Marsh Centre for Census and Survey Research

Measurement Error in Retrospective Reports of Unemployment

CCSR Working Paper 2012-02

Jose Pina-Sánchez, (jose.pinasanchez@postgrad.manchester.ac.uk)

Johan Koskinen, (johan.koskinen@manchester.ac.uk)

Ian Plewis, (ian.plewis@manchester.ac.uk)

www.ccsr.ac.uk

Contents

Measurement Error in Retrospective Reports of Unemployment

Jose Pina-Sánchez, (jose.pinasanchez@postgrad.manchester.ac.uk)

Johan Koskinen, (johan.koskinen@manchester.ac.uk)

Ian Plewis, (ian.plewis@manchester.ac.uk)

Cathie Marsh Centre for Census and Survey Research, School of Social Sciences,
University of Manchester, Manchester, M13 9PL, UK

Abstract

In this paper the presence of measurement error in two retrospective questions on work status is analysed. Measurement error in retrospective reports of work status has been difficult to quantify in the past. Issues of confidentiality have made access to datasets linking survey responses to a valid administrative source very problematic. This study uses a Swedish register of unemployment as a benchmark against which responses from two survey questions are compared and hence the presence of measurement error elucidated. We carry out separate analyses for the different forms that measurement error in retrospective reports of unemployment can take: miscounting the number of spells of unemployment, mismeasuring duration in unemployment, misdating starts of spells, and misclassification of status. The prevalence of measurement error for different social categories and interview formats is also examined, leading to a better understanding of the error-generating mechanisms that interact when interviewees are asked to produce retrospective reports of past work status.

Key words

Measurement error, survey, retrospective design, work history, unemployment, administrative data.

1. Introduction

A widely used tool in surveys when there is an interest in capturing changes over time is to rely on retrospective questions. These types of questions ask respondents for information about events from the past. They can obtain information on a particular span of time at a single occasion, and are therefore cheaper to carry out than the alternative approach of repeatedly contacting respondents during that span of time, known as longitudinal or prospective designs.

Because the interviewee is contacted only once, there is no risk of attrition (that is subjects dropping out of the study) or lack of consistency derived from changes in the wording or design of questions over time. Moreover, retrospective questions capture information on the full history of an event for a particular period of time, whereas prospective questions are only able to offer a snapshot of the state of the event in repeated occasions (waves) of the same period of time; hence being unable to capture within wave transitions[1]. Although this can be corrected if retrospective questions are included into each wave.

The major problem for retrospective questions stems from the higher propensity of finding measurement error (ME from here on) in the responses. In particular, interviewees answering retrospective questions are faced with a higher cognitive challenge since not only do they need to interpret the question correctly but they also need to recall it. Furthermore, the memory failures that generate ME in retrospective questions are often interrelated with the nature of the topic and with the relative difficulty of reporting it (low saliency, social desirability, etc.), resulting in complex error-generating mechanisms. These issues need to be examined in order to generate more accurate assessments of the quality of the data obtained from retrospective questions. A better understanding of how ME affects these questions would help to improve survey designs, to acknowledge the limitations of studies using this type of data, and to enhance the use and effectiveness of statistical methods for the adjustment of ME.

In this paper we study the nature and extent of ME found in the answers to two retrospective questions. We use data from the Swedish register of unemployment that has been linked to the participants in a survey. We assume the register data are error-free and by comparing it with

---

[1] See Solga 2001 for a comparison of data quality derived from prospective and retrospective questions.

responses from the survey we are able to ascertain the amount of ME found in those questions. By fitting different regression models we show how the effect of recall time varies for different groups of the population and different survey formats. Moreover, evidence on these effects is also used to shed some light on the ME-generating mechanisms operating in retrospective questions on work histories.

The paper continues with a theoretical analysis of these ME-generating mechanisms, followed by a review of the empirical findings in the literature. Section 3 presents the details of the datasets that are used. The analysis in Section 4 is divided into four parts; each one examining a different form of ME. At the end of Section 4 we summarize the findings and in Section 5 we review a series of caveats regarding the validity of the study.

2. ME Generating Mechanisms Affecting Work Histories

First we present the hypotheses that have been proposed to explain the appearance of ME in retrospective questions in general and also those that relate to the reporting of work histories in particular.

2.1. Theoretical Issues

ME in retrospective questions is mainly related to the saliency of the event and recall time (Bound et al. 2001). The former refers to the extent to which the event of interest left an imprint in the respondent's memory, the latter measures the distance in time between the occurrence of the event and the date of the interview. The lower the saliency and the longer the recall time, the higher the levels of ME are expected to be. In turn, saliency is determined by how often an event occurs, or its level of interference. Interference as a source of ME relates to the difficulty of discerning the occurrence of specific events when several of them have taken place during the reference period. *"Classical interference and information-processing theories suggest that as the number of similar or related events occurring to an individual increases, the probability of recalling any one of those events declines"* (Mathiowetz et al., 2001, p.163).

Other sources of ME that affect the report of work histories in particular are: 'misunderstanding' and social desirability. The term 'misunderstanding' refers to the first step identified by Tourangeau (1984) 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 interviewee. 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. This complexity of the behavioural experience is a significant factor affecting the quality of retrospective reports. In this respect, being more embedded into the labour market can be expected to be associated with being more familiar with its functioning and therefore to favour accurate reports (Morgenstern and Barrett (1974), Horvath (1982), Levine (1993), Bound (2001), and Paull (2002)).

Social desirability bias appears in value-laden topics. Socially undesirable events tend to go unreported whereas socially desirable events are often over-reported (Pyy-Martikainen and Rendtel, 2009). For work status, employment and unemployment are respectively the most and least desirable categories. Hence, the state of being unemployed might be more prone to be wrongly reported. *"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 order to predict the prevalence of ME these error generating mechanisms need to be accurately assessed; however this is not always possible. Whereas recall time and interference can be easily spotted by the length of the time-span and the number of events reported, saliency, misunderstanding, and social desirability are more elusive. In order to determine their effects a set of hypotheses about how ME affects different groups in the population is needed.

Paull (2002) argues that the overall saliency of employment and unemployment can be expected to be greater for men than women because of the financial importance of being the prime household earner. Likewise, regarding misunderstanding, Bound (2001) argues that population subgroups with lower labour force participation such as women or teenagers are more likely to generate ME because of their lower engagement with the labour market. With respect to social desirability, the long-term unemployed suffer a stronger social stigma than people unemployed for a short period and are then more likely to generate ME, much more so if the interview uses a face-to-face format (Mathiowetz and Duncan, 1988).

One last cause of ME that we consider is discretization, which is derived from the question's format and hence might be expected to affect every respondent equally. This mechanism is at work when the question forces respondents to use too a coarse time-unit leading them to omit short spells.

For example, when responses are constrained to months instead of days, spells of unemployment shorter than a month cannot be reported. 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. These mechanisms are not considered here because they can be assumed not to affect retrospective questions or subgroups of the population in a systematic way. Figure 1 below summarizes the different sources of ME that affect the collection of work histories using a retrospective design.

Figure 1: Sources of ME in retrospective reports of work histories

ME common in survey data	*ME affecting work-status*	*ME in retrospective designs*
Self-acquiescence; *Interviewer effects;* *Coding errors; etc*	*Social desirability;* *Misunderstanding*	*Recall errors*: determined by saliency of the event and length of the time frame; *Discretization*

2.2. Empirical Findings

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.

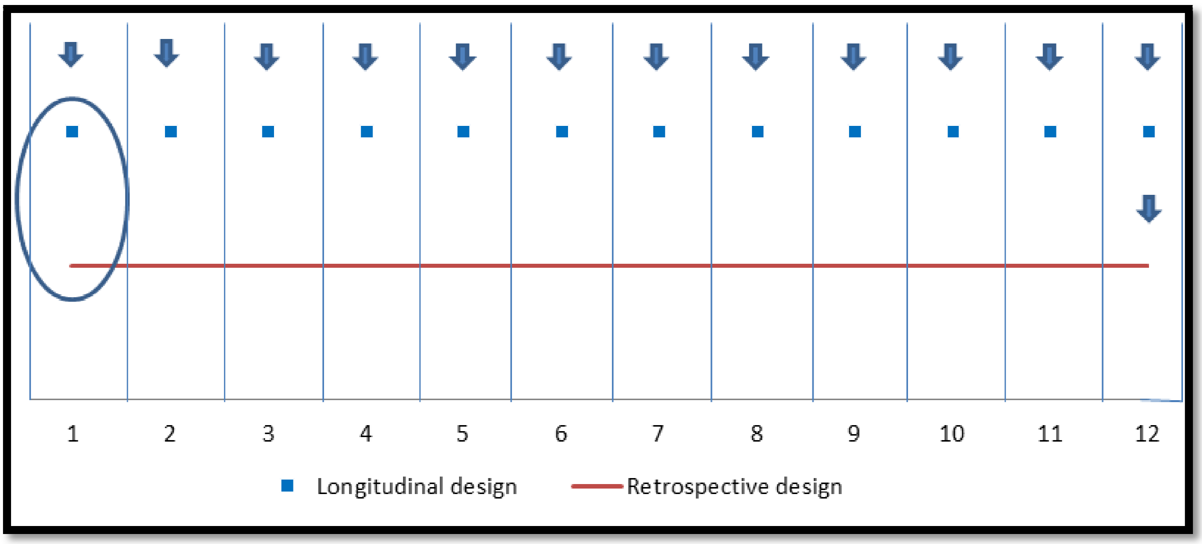
Two main research designs have been used to ascertain the presence of ME in surveys at the individual level: replication and validation studies. Replication designs identify ME from the variability in responses to identical questions taken from the same respondents. However, because none of the responses are free of ME, it is not possible to estimate a systematic component of the error. For that, a gold standard (a dataset where the true measures for the same subjects are available) is needed. Comparing responses against their true values, a full analysis of the presence of ME can be performed, and this is precisely the design that validation studies use.

Therefore, the only fully informative studies in the literature regarding ME in retrospectively reported work histories are those that look at retrospective questions on work status, and that use a validation design in the study of ME. Unfortunately, researchers' access to official data on people's work status has been restricted by concerns about confidentiality and this has limited the number of validation studies. Consequently, despite their limitations the bigger family of replicated studies generate the majority of useful insights into the effect of ME in surveys.

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 (retrospective question). Discrepancies between the two measures are seen as evidence of ME derived from memory failures. This research design is illustrated in Figure 2. Arrows are pointing at the specific time points when the interviews took place. The periods captured by the circle indicate the data that is matched and where retrospective and current reports are contrasted.

Most of the studies with this design have used the Current Population Survey[2], 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 (2001). Here, the author 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.

Figure 2. Current vs retrospective status design



From reviewing the studies of Morgenstern and Barrett (1974), Horvath (1982), and Levine (1990), Bound concluded that unemployment rates were being underestimated by retrospective questions. Furthermore, the rate of underreporting varied notably across different subgroups of

---

[2] http://www.census.gov/cps/

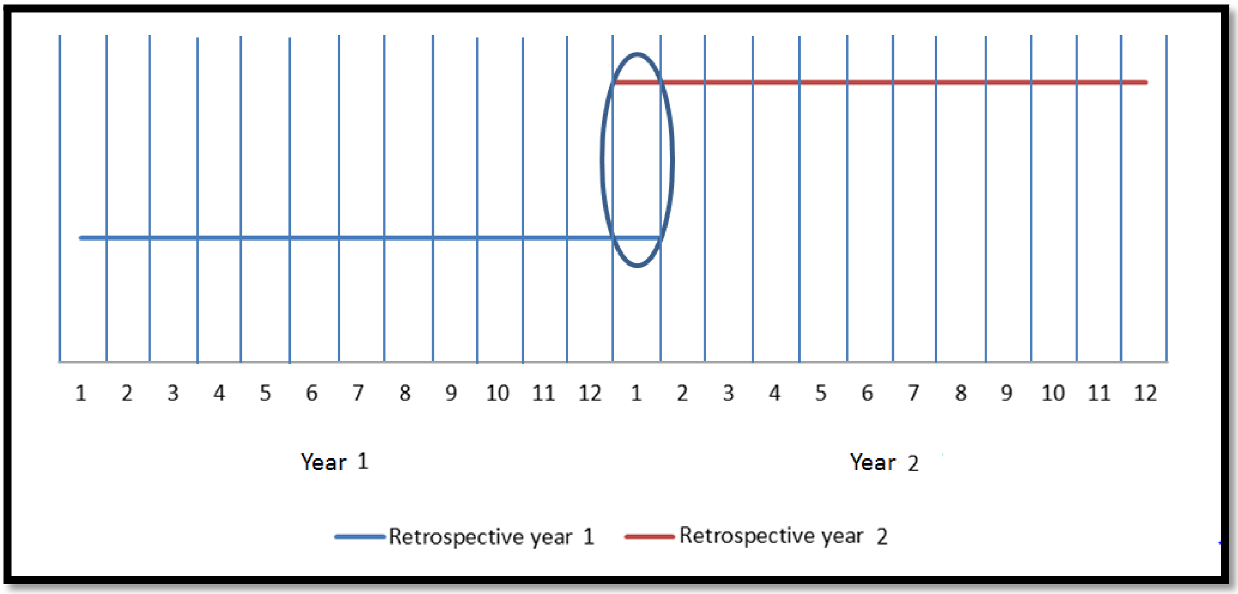
the population. Morgenstern and Barrett (1974) estimated that groups whose part-time employment and movement into and out of the labour force is much greater than the average (women and young people), are more likely to omit spells of unemployment. In most cases they declared themselves out of the labour 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 WES 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 (2005) compares current reports with retrospective work histories for the last year from the German Socio-Economic Panel (SOEP) 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 obtained 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 time when considering his/her activity in the present. Figure 3 below illustrates the period of study, which is encompassed by the circle.

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 on 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[3]. Jacob found that 86% of men who defined themselves as unemployed were also considered unemployed using the ILO approximation but only 44.5% of the women's self-definition agreed.

Figure 3. Comparing two retrospective designs



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 (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*

[3] Unemployment occurs when people are without jobs and they have actively sought work within the past four weeks.

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. These studies compare the data collected from a survey with a gold standard. However, it seems that only three studies on the topic have been published: 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 a US 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. The authors looked at ME for different topics not just unemployment. In general, little evidence on ME was found when reporting unemployment spells one year after, for example *"90% of the sample were able to report the year of hire accurately within one year"* (Duncan and Hill, 1985, p. 515). However, since the sample in this study is only made up of workers from the same firm, it is not entirely representative of the US population. It probably underrepresents those groups of the population less embedded in the labour market who have been found to be more prone to ME.

Mathiowetz and Duncan (1988) used the same dataset as Duncan and Hill (1985) but only studied ME in reports of unemployment. The authors find a very different picture of ME and its implications depending on what forms of ME are looked at. Respondents offered very accurate answers when asked to report the total time spent unemployed in the last year. However, when required to designate and time each spell of unemployment, results are far worse: 66% of spells were omitted.

An interesting development that Mathiowetz and Duncan (1988) included was to disentangle some confounding effects by comparing the probability of committing an error in different periods for different socio-demographic groups. This showed that demographic variables such as ethnicity, education, age and gender were not found to be statistically significant if other variables capturing saliency and difficulty of the task were included. Saliency was measured by the length of the spell and difficulty of the task by the number of events of unemployment that the respondent experienced during the period of analysis. From here, the authors hypothesized that it is not the condition of being younger or a woman that is associated with inaccurate reports, but their more complex work histories. However this finding could be questioned since the whole sample is made up of workers and therefore gender and age cannot be used to distinguish between different levels of engagement in the labour market.

Another of Mathiowetz and Duncan's findings challenges the hypothesis that accuracy deteriorates as the distance between the period to be recalled and the date of the interview grows. The authors found that the effect of time length 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, around the last five months, where it 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 Panel Study of Income Dynamics, is not composed only of workers, which improves the external validity of Pyy-Martikainen and Rendtel's study over Duncan and Mathiowetz's. In addition, every form of ME (omission of spells, underreported durations, misdated starts and ends of spells and misclassified status) can be 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 from the survey ended in the respondent reportedly 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 on 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 in the total time reported to be unemployed and the one registered as such. With respect to omission their results point in the same direction as that found in replication studies: being female increases the odds of omission by 23.7%, short spells are harder to remember, having less than one month of cumulative unemployment time multiplies the odds of omission by almost five, but for every additional month of unemployment it decreases by 27.6%.

An original finding from this study relates to the effect of age which was demonstrated to have a quadratic effect, with the probability of omission decreasing until age 37 and increasing after this age.

Respondents with longer cumulative unemployment, people with higher levels of education and those receiving unemployment benefits were more likely to underreport the time spent unemployed, whereas females were more likely to over-report.

The analysis that we carry out in the following chapters also uses a validation design. We introduce the specifics of the dataset that we use in the next section.

3. Data

The data we use has been obtained from the “Longitudinal Study of the Unemployed” (its acronym is LSA in Swedish), a research project designed by the Swedish Institute for Social Research (SOFI) at Stockholm University, directed by Sten-Ake Stenberg, and with the collaboration of the register of unemployment (PRESO[4]). This register provided individual-level data on the work status of the participants of three surveys, run in 1992, 1993 and 2001, which we introduce next.

3.1. Survey data from the LSA

The three surveys are relatively similar with respect to the composition of both the sample of participants and the questionnaire. The sample was designed to capture 830 jobseekers randomly selected from those who were registered as such in the PRESO files in February 1992.

In addition participants were selected from those meeting the following criteria: applicant category “not in employment”, a tag indicating that he or she can work immediately, applicant desire to work full-time, applicant age between 25-54, Nordic nationality, and no occupational disabilities. Because of unit non-response, the percentage of subjects responding to the survey in 1992 was 64.7%, dropping to 59.4% and 50% for the surveys run in 1993 and 2001, leaving a final sample size of 594 and 500 respondents respectively. The 1992 and the 1993 surveys have almost

[4] PRESO is a register from the Swedish employment office (Arbetsmarknadsstyrelsen).

an identical design, while the latter is much better documented with a richer codebook, hence we decided to discard the 1992 survey and focus the analysis on data from the 1993 and 2001 surveys.

Given the different sample sizes due to attrition, there is a concern that the composition of the sample might change from one survey to the other affecting comparisons between the two. However, the composition of the sample with respect to age and gender did not change over time. Men comprised 69% of the 1993 sample and 67% of the 2001 sample. The mean age went from 36.3 years old in 1993 to 45.4 in 2001[5], but taking into consideration the eight year gap between surveys it appears that the age distribution has not changed either.

The design limitation to participants aged from 25 to 54 (in February 1992) makes the sample not fully representative of the Swedish unemployed population. The reason behind this choice was to make sure that the cohort of people with a higher probability of being economically active was captured. In the codebook "PRESO 1989-1993"[6] it is argued that the age range 18-24 might include many students, and the risk for those over 55 years is that they would have left the workforce. For the 2001 survey, the age range was pushed eight years forward, now capturing respondents between the ages 33 and 63 and thus not being able to represent young people at all.

With respect to the questions in the survey capturing work status, LSA-1993 uses an event occurrence framework, rather similar to the one from the Panel Study of Income Dynamics used in Mathiowetz and Duncan (1988). Lawless (2003) coined the term 'event occurrence framework' to define questions where events are asked to be reported in order of occurrence indicating the particular status, and their start and end dates. The author differentiates these types of questions from a 'multi-state framework', where the status of the event is reported for each interval in which the time frame is divided. LSA-2001, much like the European Community Household Panel used by Pyy-Martikainen and Rendtel (2009), uses a design that is more similar to this multi-state framework.

For the 1993 survey those questions 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?*

[5] The standard deviations for age in 1993 and 2001 were 8.5 and 8.6 respectively.
[6]This document can be provided by the Swedish National Data Service under permission from Pr. Stenberg.

Which was the subsequent main activity? When did this activity start? When did it end?[7]

In addition, the last line asking for ensuing activities was repeated twelve times, so, a total of thirteen[8] "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".

Therefore, two main differences between the 1993 and 2001 surveys can be noted. Firstly, the recall time is vastly expanded in LSA-2001; from the common question using a time frame of little more than one year it changes to an eleven years' time frame, which could be expected to come at a much higher price in terms of recall failures and ME. Secondly, observations in LSA-2001 are dated on a monthly scale instead of the implied daily one used in the 1993 survey. On the other hand, the different states that can be chosen remained constant for the two surveys: *"w*orking", "studying", "jobseeker", "unpaid parental leave", "homeworker" (not employed), "pensioner", "AMS-training[9]", and "other".

The work histories 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 one transition to a specific state occurs. In addition to work status, two other variables retrieved from the surveys are used in the analysis: gender and interview format. 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%

---

[7]This and the following quote are translations from the original in Swedish.
[8]The number of time-periods allowed seems to be enough for respondents to reports their total activities experienced in one year since only three participants had ten or more activities to report.
[9]This category encompasses any training provided at the jobcentre.

were conducted by phone. In 2001 these figures remained almost identical 85.4% and 14.6% respectively.

3.2. Register Data from the PRESO

PRESO is the Swedish register of unemployment. It collects information from jobseekers on the last week-day 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[10] while at work, the PRESO register not only captures spells of unemployment, but it also offers validation data on other work statuses such as employed, self-employed 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 (Levin and Wright, 1996). 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, it is assumed that this proportion is 100%.

A variable that codes the work status can be retrieved from PRESO on a monthly basis and the work histories of an individual can be generated. 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 registered as unemployed in February 1992.

[10] 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.

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 status 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 the 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 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 of unemployed in February 1992.

Another weakness stems from a relative lack of consistency. The PRESO file has been through four periods (1990-1994, 1995-1997/June, 1997/July-2000/October, 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 would not be accurate 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, but for reasons of simplicity in the analysis we assume that the register captures spells of unemployment perfectly.

Other variables from this dataset that we use in the analysis are: age, experience, spells of unemployment, and cumulative unemployment. Experience is an ordinal variable that goes from 1 (minimum) to 4 (maximum) capturing the amount of experience that the subject reported to have in the category of work that he or she applied for. “Spells of unemployment” records the number of spells of unemployment experienced by the subject over the window of observation. Cumulative unemployment captures the number of days (or months) subject has spent registered as unemployed during that same window of observation.

Experience was reported by the subject and it is therefore 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.

3.3. Matched Dataset

The analysis of ME requires person-period unemployment data to be simultaneously available in the survey and the register. The first consequence of this is the loss of subjects included in PRESO but dropped due to attrition in the LSA. In addition, 152 subjects from LSA-1993 and 83 from LSA-2001 reported incomplete dates; they included the month but not the day (LSA-1993) or the year but not the month (LSA-2001). Other subjects reported impossible dates such as the 31st April. Moreover, PRESO contained 72 subjects who had at least one spell dated to have occurred after the start of the following spell, which generates incorrect overlaps between past and future spells.

While this situation doesn’t affect the analysis of ME when it takes the form of over and under-reporting of spells because that only takes into account the number of spells reported, analyses of misdating and misclassification require specific dates of the event if they are to be compared reliably. Hence, cases with inaccurate or incoherent dates were dropped, leaving a sample size for these sections of the analysis of 395 in 1993 and 387 in 2001.

Finally, in order to be able to match spells’ dates they need to be measured in the same time unit. This is not a problem for the 1993 study since both LSA-1993 and PRESO use days as the time unit. However, LSA-2001 records events in months, and therefore the PRESO data need to be discretized into months to make the match possible. For this we 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 unit does not allow short spells to be captured. As a result, 13.7% of the spells of unemployment were lost because they were shorter than 28 days.

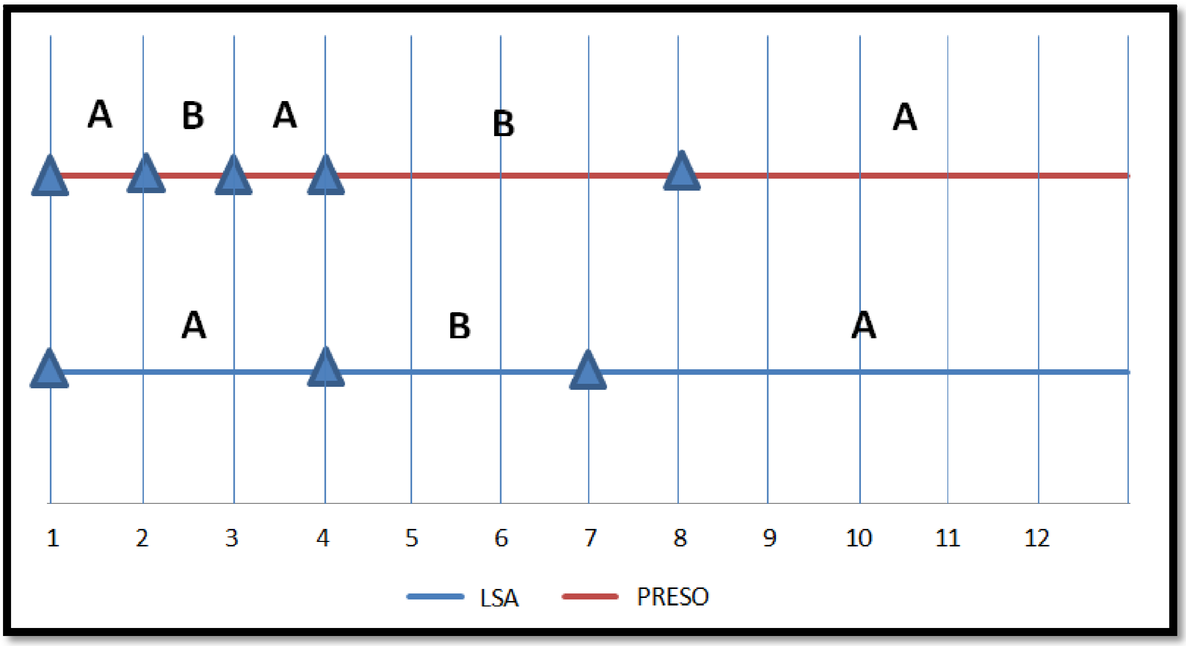
Every analysis presented in the following section is based on the comparison of survey (error-prone data) with register data which is assumed to be error-free. However, the problems seen above (incapacity to detect fraudulent behaviours, mistakes found at dating the event and omission of short spells due to discretization) damage the validity of this assumption. In Section 5 the assumption of PRESO being a gold standard is discussed in detail but in what follows we maintain this assumption and consider discrepancies between the two datasets to be evidence of ME in the responses to the survey.

4. Analysis

The analyses presented in this section focus on the presence and the correlates of ME in retrospective reports of unemployment. Four different forms of ME are considered. By "forms of ME" we refer to the ways that ME can be quantified, which for a categorical variable with a time component are: miscounting of the number of spells, mismeasuring the length of spells, misdating starts of spells, and misclassification of status. Figure 4 illustrates each of these different types of ME by plotting the registered and reported work histories during 12 months for a hypothetical subject. A represents spells of employment, B spells of unemployment, and the triangles their starts.

ME in the form of misclassification requires the data from a work-history to be represented as person-period observations. Thus, in Figure 4 there is misclassification at months two and seven. A miscount of the number of spells can be spotted when work histories are treated as count data (number of spells of unemployment in our analysis). In Figure 4 there is a case of omission of a spell of unemployment in the second month. Misdating and mismeasuring of spells' duration can be observed when work histories are interpreted as duration data. These two forms of ME appear when starts and ends of spells are not accurately reported. Figure 4 shows an example of underrepresentation of duration for the spell of unemployment that encompasses month four to eight but which was reported to finish in month seven.

Figure 4. Registered and reported work histories for the same subject and time-span



Other typical forms of misdating are heaping effects. Torelli and Trivellato (1993) defined them as *"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)"; or less formally as "a rounding-off by the respondent at a scale unit coarser than the one formally adopted in the survey instrument, due to fuzzy recall"* (Torelli and Trivellato, 1993, pp.189-190).

In what follows, the four types of ME are examined for the two surveys (LSA-1993 and LSA-2001), using descriptive and inferential analyses. In doing so we 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 disregarded the model based analysis of misclassification.

Two criteria determined the set of explanatory variables used in the analysis. First, variables used in Mathiowetz and Duncan (1988), and Pyy-Martikainen and Rendtel (2009), were prioritized in order to better replicate their analyses. Second, the variables that identify categories of the population and features of the interviews were selected. The evidence found for each type of ME in the two surveys plus the effects of the selected explanatory variables is used to test some of the hypotheses suggested in Section 2 regarding the functioning of ME-generating mechanisms:

1. Recall time. The probability of misreporting is greater the longer the distance between the interview and the event reported. In particular, we expect recall time to be associated with miscounted (omitted) spells.
2. Misunderstanding. Categories of the population that are relatively more embedded in the labour market (middle age men) provide more accurate reports than less engaged groups (young people and women). Misunderstanding is expected to be reflected in misclassified spells.
3. Social desirability. Groups of the population more susceptible to feel the social stigma derived from being unemployed (e.g. long-term unemployed), and interviews conducted face-to-face (as opposed to phone interviews) have a higher chance of misreporting. A social desirability effect is probably related to mismeasured (shortened) spells.
4. Interference. The greater the number of spells of unemployment the bigger the extent of ME. Interference is expected to be associated with miscounted (omitted) spells.

We now look at ME in the form of over/underreported spells. We start by comparing the survey with the register data in an aggregated way, using sample estimates from each dataset, to be followed by person-level comparisons. The other forms of ME follow the same structure with the exception of misdating where we only perform a descriptive analysis. We conclude this section summarizing the evidence found regarding the presence of ME in its different forms and what can be inferred with respect to its correlates.

4.1. Miscounting the Number of Spells

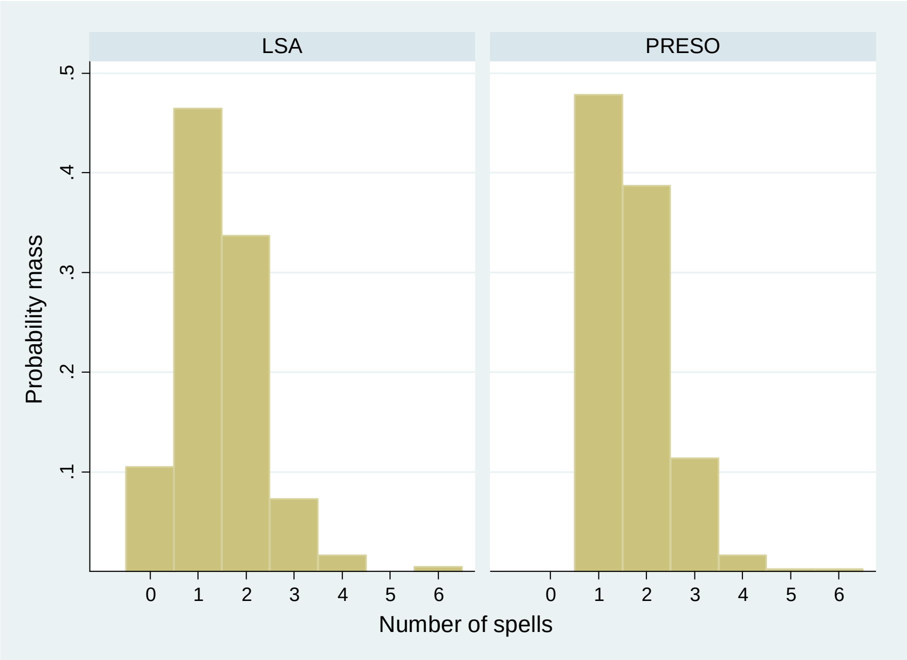
We start the analysis by comparing the number of spells of unemployment that were reported in the two surveys against what was registered in PRESO for the same time period and the same subjects. In order to exclude any variability in sample composition and to facilitate comparisons of the two surveys, we restrict the analysis to the 439 subjects who were jointly captured in LSA-1993 and LSA-2001. Lastly, for the study of ME in LSA-1993 only spells of unemployment with reported start dates in 1992 were counted, whereas in the study of LSA-2001 these were limited to starts dated from 1990 to 2000 inclusive.

LSA-1993

The distribution of spells of unemployment over the 15 to 16 month period covered in LSA-1993 is presented in Figure 5 below. The estimated sample mean was 1.4 spells per subject, and the standard deviation .9. Ten percent of the respondents reported not to have experienced a single spell of unemployment. The maximum number of spells of unemployment reported is six and the mode and median is one.

In the PRESO dataset the mean number of spells of unemployment registered for the same subjects and time frame is 1.7, and the standard deviation .8. The bigger mean in PRESO than in LSA is evidence of omission of spells of unemployment in the survey reports. Some other differences can be appreciated, there are no subjects that didn't experience unemployment (that was one of the criteria used to select the sample), and the median is 2.

Figure 5. Probability mass function of the number of spells of unemployment in LSA and PRESO -1993



The difference between the PRESO and LSA means is of .3 spells, that is the mean of reported spells is 86.3% of the registered one. Finally, the proportion of subjects reporting the correct number of spells in LSA-1993 is 54%.

LSA-2001

For LSA-2001 the mean of spells per subject and the standard deviation are both 3.2, and the maximum is now 13. Higher figures than in LSA-1993 are to be expected given the extension of the time frame from over one year to eleven years. However the percentage of subjects who reported not having experienced a single spell of unemployment increases to 16% (Figure 6). This might be evidence of exacerbated memory failure due to the wider time frame and/or omission of spells of unemployment shorter than a month due to the use of coarser time-units for the record of spells' starts and ends.

The comparison of the reported spells with the registered ones for this time frame shows an even larger discrepancy. The sample mean for the number of registered spells of unemployment is 8.2, the standard deviation 5.2 and the maximum 25. The mean of reported spells is 39% of the registered one. Furthermore, the proportion of subjects reporting the correct number of spells in LSA-2001 is just 7.5%.

Figure 6. Probability mass function of the number of spells of unemployment in LSA and PRESO 2001

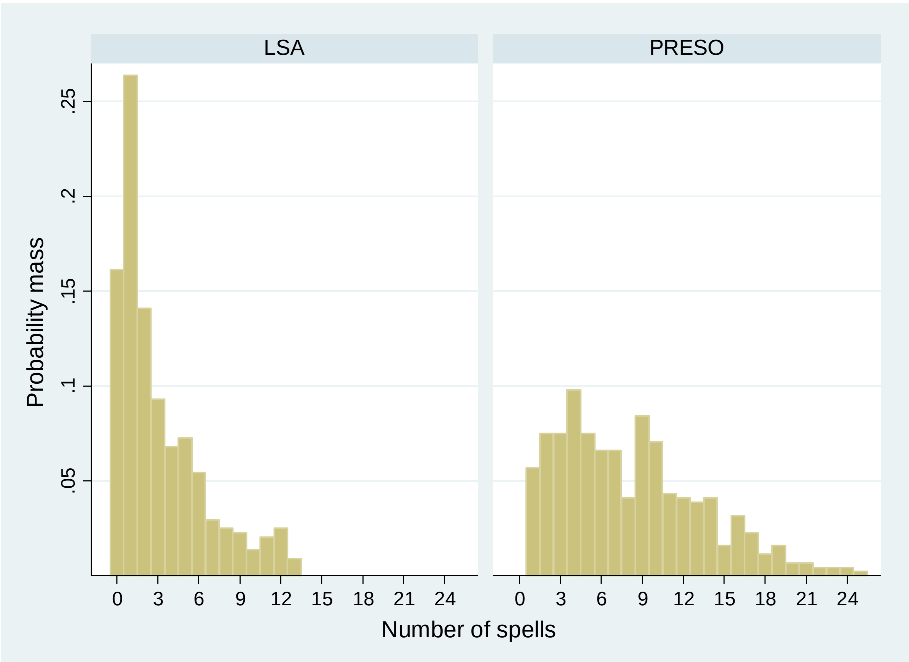


Table 1 below summarizes the results from the four datasets. The lower means for the two surveys compared to the registered ones for the same period shows evidence of omission. In addition omission is seriously aggravated in LSA-2001 as the difference between means with respect to the 1993 study grows, and the percentage of subjects reporting the right number of spells falls. Differences between the two studies can be accounted for by the two characteristics that differentiate them, i.e. the longer time frame and the use of months as time-units. In Section 3.2 and 3.3 it was seen that the use of a coarser time unit produces omission of short spells. The higher levels of omission in LSA-2001 could be corroborating hypothesis 1 on the impact of extended recall time, although longer time frames also generate more complicated work histories, hence the increased omission might also be due to the impact of interference (hypothesis 4).

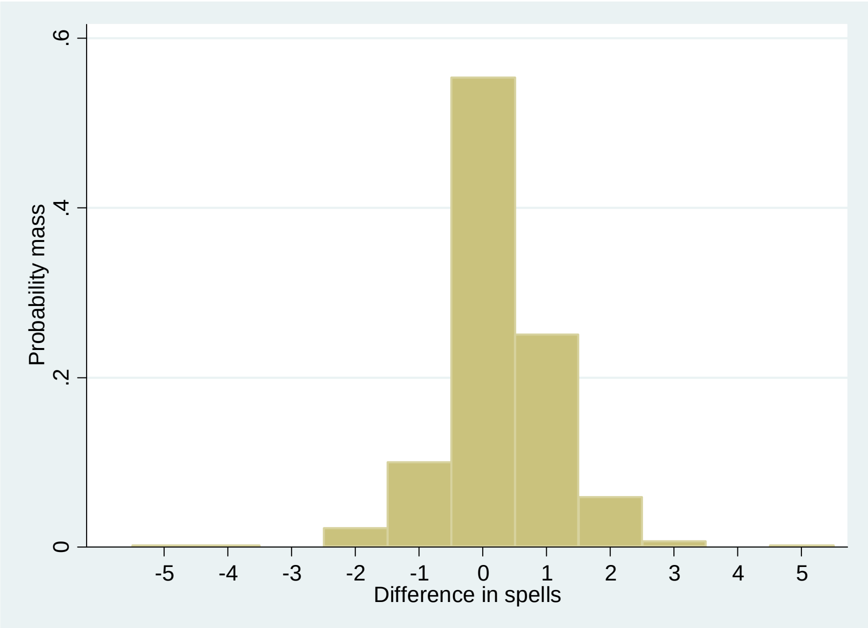
Table 1. Results from the descriptive analysis on omission of spells

Dataset	Number of spells			Spells reported as a percentage of spells registered	Percentage of subjects reporting spells correctly
	Mean	Standard deviation	Maximum		
LSA-1993	1.4	.9	6	**82.3%**	
PRESO-1993	1.7	.8	6		**54%**
LSA-2001	3.2	3.2	13	**38.8%**	
PRESO-2001	8.1	5.2	25		**7.5%**

Person Level Analysis

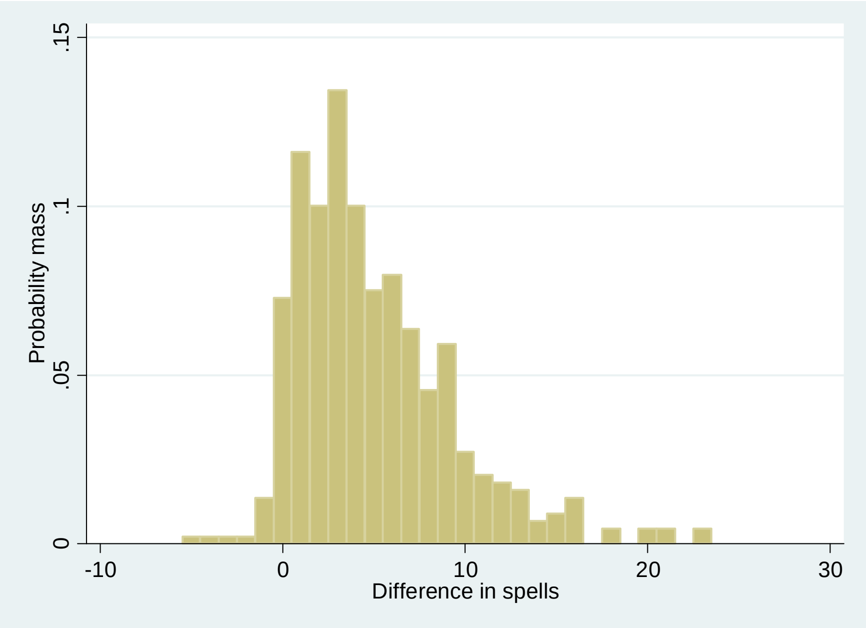
Here we continue the study of omission but now using person-level measures. Figure 7 depicts the probability mass function for the difference between the number of events of unemployment reported in LSA-1993 and registered in PRESO. The previously mentioned 54% of subjects reporting the same number of spells of unemployment can be seen here as the median of the histogram. Figure 7 also shows that 12% of cases have actually over-reported episodes of unemployment.

Figure 7. Probability mass function of the difference between the number of spells registered and reported in LSA-1993



The probability mass function for the difference in number of spells of unemployment between the LSA-2001 and the PRESO file for that period shows larger variability (Figure 7 and 8). Again, this is to be expected given the wider time frame (from 1990 to 2001). The median changes from 0 to 5 and there is a great reduction of cases over-reporting spells, 56 in 1993 and 10 in 2001.

Figure 8. Probability mass function of the difference between the number of spells registered and reported in LSA-2001



In order to model the variable capturing whether the number of spells is correctly reported or at least one spell was omitted we adopt a logit model, just like Pyy-Martikainen and Rendtel (2009) did. We assume that there is a latent variable y_i^* describing the propensity of a person i to omit reporting unemployment spells, and this latent variable is also assumed to follow the model

$$y_i^* = x_i\beta + \varepsilon_i \tag{1}$$

where $i = 1, 2, .., n$ indexes subjects, x_i is a (1 x p) vector of covariates (including a constant), β is a (p x 1) vector of the parameters to be estimated, and the error term ε_i is assumed to be independent and to follow a logistic distribution with mean zero and variance $\pi^2/3$.

The model can be alternatively expressed as

$$logit\ [P(y_i = 1|x_i)] = x_i\beta$$

where

$$y_i = \begin{cases} 1\ if\ y_i^* > 0 \\ 0\ if\ y_i^* \leq 0 \end{cases}$$

Variable y_i (omission) is thus a binary variable indicating: “omit” or “does not” omit spells. The explanatory variables are: age, gender, interview format, experience, cumulative unemployment and spells of unemployment. Age was centred around the mean, 37, here and in the subsequent models; the mean cumulative time spent in unemployment and the mean number of spells of unemployment for the window of observation considered in this model is 429 days and 1.7 spells, while the mean for experience was 2.6. A list with some descriptive statistics of the explanatory variables used in this paper is included in Appendix IV.

Results for this model for the two datasets are presented in Table 2 below/ In addition, statistically significant results ($p < .05$) appear in bold. Subjects who over-reported their number of spells of unemployment were discarded (13% in LSA-1993 and 2% in LSA-2001), which leaves a sample size of 383 subjects in LSA-1993 and 429 in LSA-2001.

Table 2. Estimates for the logit regression on omission in the 1993 and 2001 datasets*,**

	1993	2001
Age	-.02 (.02)	-.01 (.03)
Age^2	-.001 (.002)	<.001 (.003)
Female	.37 (.27)	1.04 (.60)
Phone interview	.48 (.36)	-.06 (.53)
Experience	-.22 (.19)	.01 (.35)
Cumulative unemployment	**-.002** (.001)	**-.05** (.02)
Spells of unemployment	**1.41** (.18)	**.83** (.18)
LR chi^2(7)	**89.78**	**71.43**
Sample size	383	429

*Here, and in the rest of Tables presenting model results, estimated regression coefficients appear at the top of every cell and standard errors at the bottom between brackets.
**The regression estimates represent the untransformed effect on the log-odds of omission.

Number of spells of unemployment (the variable created to account for the impact of interference) is positive and statistically significant in both datasets. Cumulative unemployment (the variable measuring the amount of time spent unemployed) is also significant, but in this case negative, in 1993 and 2001. Therefore the more spells of unemployment and the shorter they are the higher the probability of omitting them. This is an interesting result since from a social desirability standpoint (hypothesis 3) the opposite could be expected.

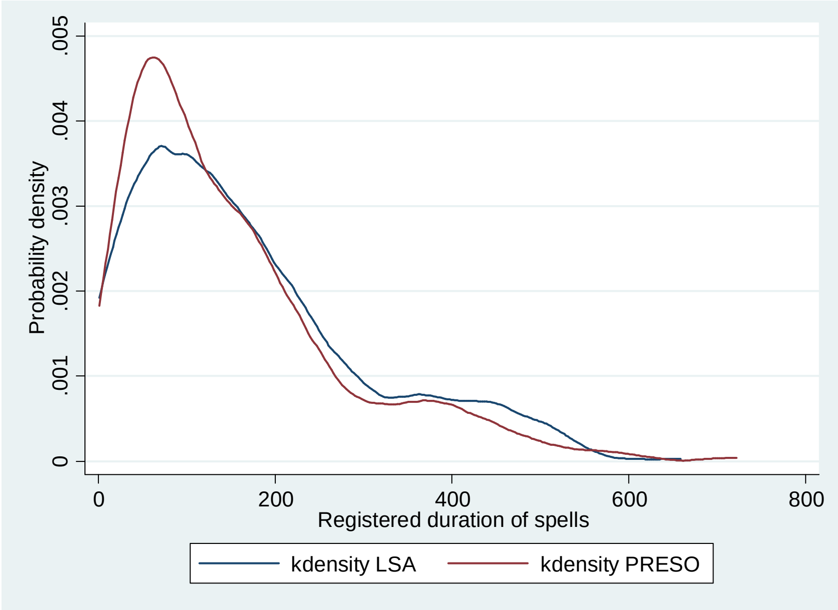
4.2. Mismeasuring the Length of Spells

In this part of the analysis we look at ME in the form of misreported durations of spells of unemployment. Again the subjects and time frames compared in LSA and PRESO are identical. As mentioned in Section 2, some spells were incompletely dated in the two surveys. Hence in this subsection and for the following analyses of misdating, and misclassification, we discard subjects who had reported or registered spells in such a way. In addition, for the descriptive study we restrict the analysis to those subjects who were jointly included in the four datasets, which reduces the sample size from 439 to 218. In order to ensure that the comparison of spell lengths between survey and register is derived from the same period we only include spells starting from 01/07/1989 and ending no later than 31/12/1993 in the LSA-1993 and 01/01/1990 to 31/12/2001 for the LSA-2001.

LSA-1993

These 218 interviewees reported 374 spells of unemployment. The probability density function for the length of spells of unemployment in LSA (Figure 9) shows positive skewness; the mean duration is 167 days while the median is 132. During the same time-period and for these same 218 interviewees PRESO registered 668 spells of unemployment. The probability density function for the length of spells in PRESO is even more skewed. The mean and median are now lower, 153 and 120 days respectively, which might be related to the omission of short spells in the survey. The higher level of skewness is also denoted by the lower standard deviation (132 in LSA and 125 in PRESO) and higher maximum (659 in LSA and 722 in PRESO). Other than this the two probability density functions are quite similar.

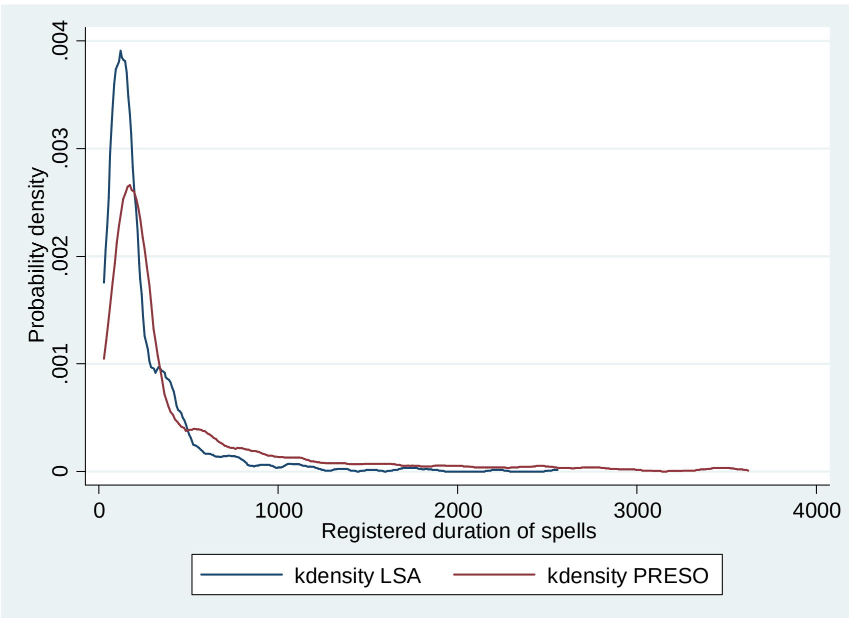
Figure 9. Probability density function of the length of spells of unemployment in LSA and PRESO -1993



LSA-2001

Unemployment spells from LSA-2001 reported in months were converted to days. The 218 subjects contemplated reported a total of 843 spells. The probability density function of the length of spells reported in LSA-2001 (Figure 10) is similar to the one from LSA-1993, the main difference being more skewness to the right. The mean is 238 and the standard deviation 259. In PRESO-2001 the number of spells for the same subjects and period is 1,144, the mean length for those spells is 456 days, with a standard deviation of 601.

Figure 10. Probability density function of the length of spells of unemployment in LSA and PRESO -2001



The main difference between LSA and PRESO-2001 is the considerable number of long spells that have been shortened or omitted in LSA-2001. While PRESO-2001 registered 12% of the spells to be longer than 1000 days, the corresponding figure in LSA-2001 is only 3.5%. This pattern was not observed in the comparison of LSA and PRESO-1993, which suggests that the extended time frame in the 2001 survey is the cause of this difference. However, because it seems to particularly affect long spells, it could be argued that the problem is derived from a social desirability component (Hypothesis 3) as much as from the extended recall time (Hypothesis 1).

Table 3. Results from the descriptive analysis on over/underreport of spells' length

Dataset	Length of unemployment spells in days			Reported duration time for spells of unemployment as the percent of the registered time
	Mean	Standard deviation	Maximum	
LSA-1993	167	132	659	**109%**
PRESO-1993	153	125	722	
LSA-2001	238	259	2557	**52%**
PRESO-2001	456	601	3623	

Person Level Analysis

The analysis now turns from examining duration at the spell level to duration at the person level. That is, we aggregate durations of spells by subject and look at misreports in total time spent in unemployment during the period of interest. Here we remove the restrictions on the sample from the previous sub-section. Interest now focuses on the investigation of the correlates of ME in each survey. For this new setting we no longer consider only subjects who were jointly included in LSA-1993 and 2001 which leaves a sample size of 394 subjects in LSA-1993, and 321 in LSA-2001.

A comparison of the difference between the reported and registered cumulative time in unemployment in the 1993 study (Figure 11) now shows that the majority of participants underreport their total time spent unemployed; 60 cases (15% of the total sample) still over-reported their time in unemployment, but the overall mean was to underreport it by 186 days. For the 2001 study the mean was to underreport the total duration in unemployment by 295 days, although again 76 subjects (23.7% of the sample) over-reported their duration.

Figure 11. Probability density function of the difference between the registered cumulative time in unemployment and the reported one in 1993

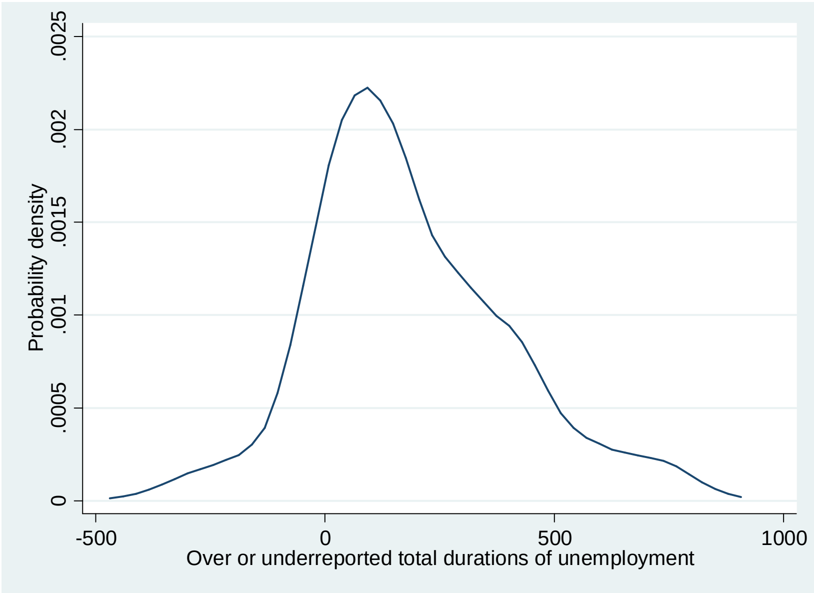
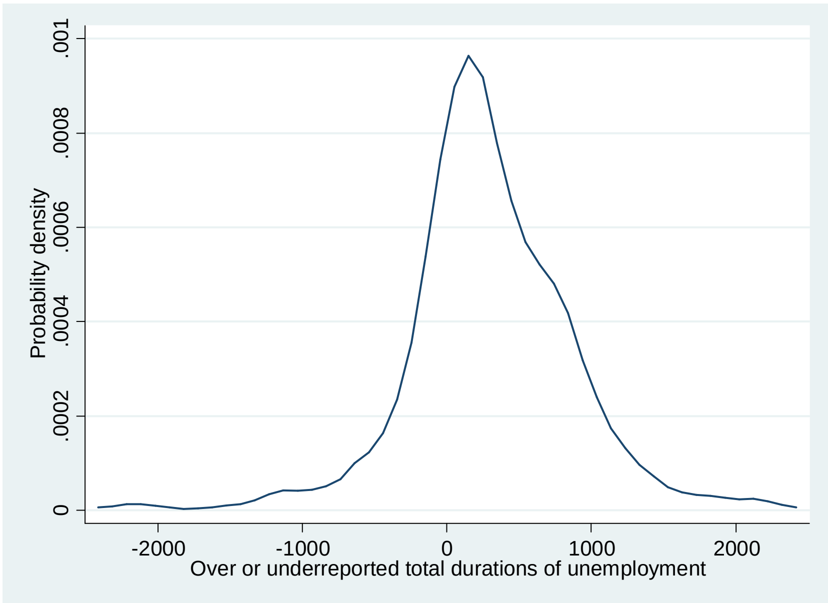


Figure 12 for the 2001 study shows a very similar distribution with the major difference being more extreme cases at the tails, which is understandable given the wider time frame used.

Figure 12. Probability density function of the difference between the registered cumulative time in unemployment and the reported one in 2001



In the following part we drop subjects that over-reported their durations and focus on modelling subjects that underreported. Over and underreporting of duration seem to be two different data-generating mechanisms and it is problematic to model them jointly[11]. We also produced a square

[11] Problems of heteroskedasticity and lack of normality in the residuals pointed at the inadequacy of such a model.

root transformation to avoid the skewness in the distribution of the response after dropping positive cases. This model choice replicates Pyy-Martikainen and Rendtel's (2009) analysis.

Hence, the response variable for the following model is characterized by y_i, which captures the square root of the difference in the cumulative time[12] reported and registered as unemployed for every subject:

$$y_i = \sqrt{\sum_{s=1}^{S_i} T_{si} - \sum_{r=1}^{R_i} T_{ri}} \tag{2}$$

where T is the duration of a particular spell for subject i; s and r are the subscripts used to index the spells reported and registered respectively by a particular subject, ($s = 1, 2, .., S$ and $r = 1, 2, .., R$) and S_i and R_i are the total number of spells reported and registered for person i.

The model used is a standard multiple regression, and the estimation method is ordinary least squares. The final model can be expressed in matrix form as follows

$$y_i = x_i\beta + \varepsilon_i \tag{3}$$

Where, just like in the logit model for omission (equation 1) x_i is a (1 x p) vector of covariates (including a constant), β is a (p x 1) vector of the parameters to be estimated, but the error term ε_i is now assumed to be independent and to follow a normal distribution. The set of explanatory variables is the same as for the previous model.

Results from these models are presented in Table 4[13] and they are not entirely consistent for the two time frames. The amount of underreported time in unemployment is higher by three days for phone interviews in 1993 while it was found not significant in 2001[14]. Similarly number of spells of unemployment significantly increases underreporting only in 1993. Unlike what was found in the study of omission, cumulative unemployment is now positive and significant in both studies.

---

[12] Days in 1993 and months in 2001.
[13] Appendix II shows Qplots and scatter plots testing the residuals' assumptions of normality and homoscedasticity.
[14] However it needs to be borne in mind that the allocation of modes was not done at random, so the mode effect could be confounded by the criteria used to contact participants.

Table 4. Results for the OLS regression on the square ratio of the underreported time in unemployment in the 1993 and 2001 datasets

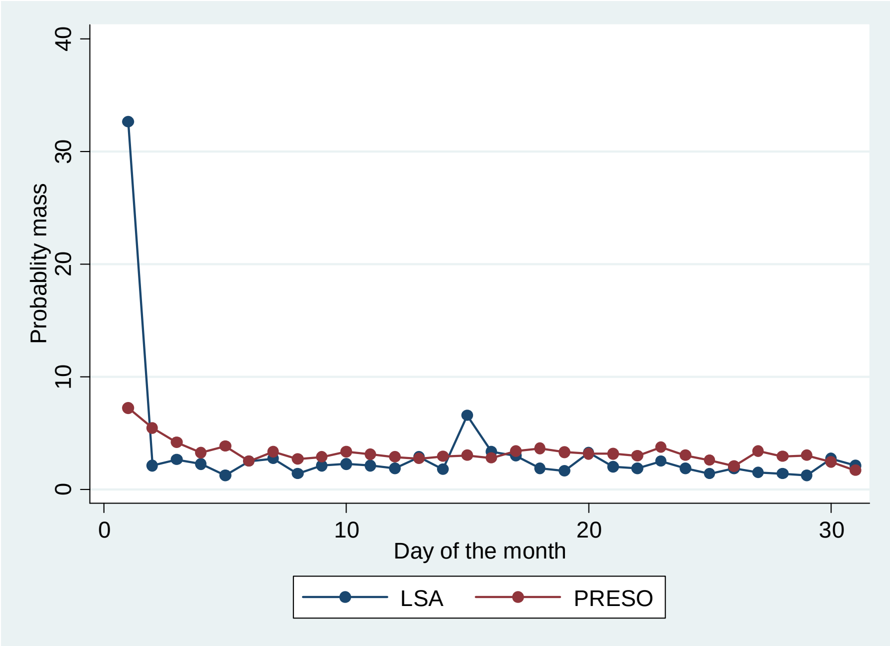
	1993	2001
Age	-.03 (.03)	-.10 (.07)
Age^2	.001 (.004)	.005 (.008)
Female	.18 (.54)	2.16 (1.12)
Phone interview	**1.81** (.71)	-.60 (1.67)
Experience	-.14 (.41)	.08 (.86)
Cumulative unemployment	**.016** (.001)	**.006** (.001)
Spells of unemployment	**.80** (.17)	.24 (.17)
Constant	3.68	18.97
Sample size	334	243
R^2	.51	.31

The effect of cumulative unemployment time is to be expected since the distribution of the response has been left-truncated so only positive differences between reported and registered times are included. Appendix III shows the same model but where positive and negative differences have been included in absolute value. Results differ for gender, which is positive and significant in 2001, and for cumulative unemployment that is now negative in 1993 and 2001. This outcome indicates that the longer the time spent in unemployment the more accurate (less disperse) the report of that total time. Just like in the model for omission (Section 4.1) these results contradict hypothesis 3 of social desirability as a source of ME.

4.3. Misdating the Start of Spells

For the case of misdating we limit the study to a graphical analysis. In Figure 13 we summarize the proportion of starts of spells of unemployment reported at each day of the month in LSA-1993 (blue line), and the ones that were registered in the PRESO for the same period and the same subjects (red line). The graph shows that survey participants have a propensity to choose the first day of the month as the starting day for their spells of unemployment: 33% of all spells. A second day that stands out from the trend is the 15th with 7%, while the PRESO only registered 7% the first day of the month while the rest of days were evenly distributed.

Figure 13. Probability mass function of the starts of spells of unemployment by day of the month for LSA and PRESO 1993



However, the fact that the day with more starts registered in PRESO is also the first indicates that not all the accumulation of spells reported to start at the beginning of the month is an effect entirely due to ME, that is in the real world it can also be observed a small concentration of spells start on the first of the month. That said, the presence of heaping effects is undeniable. In fact, the size of heaping effects found here exceeds what was found in Pyy-Martikainen and Rendtel (2009). In LSA-1993 the number of spells reported to start the first day of a month are four times those registered that same day, whereas in Pyy-Martikainen and Rendtel (2009) the number of spells reported to start in January was about two times bigger than the registered ones. These differences between the two studies are due to the use of days as time-units in our study while Pyy-Martikainen and Rendtel (2009) use months, which reduces the need to round up dates. Although perhaps the use of an event-occurrence framework in LSA-1993 instead of the multi-state framework used in Pyy-Martikainen and Rendtel (2009) might be having an effect too, since the time-periods are already defined in the questionnaire and the interviewee only needs to report the status and not the start date.

LSA-2001 shares similar features toPyy-Martikainen and Rendtel (2009) in that it uses months as time units and a multi-state framework; however, no particular evidence of heaping effects was found here. This might be due to the relative lower sample size across the 12 years’ time frame.

4.4. Misclassification

Finally, we examine ME by matching the survey and register status in the same dataset using a person-day format for the LSA-1993 and a person-month format for the LSA-2001. ME, now in the form of misclassification, is defined as a discrepancy between the reported and registered statuses at a particular time point (day or month).

Whereas the previous analyses have relied on the assumption that unemployed people kept their account activated in the register for as long as they remained unemployed[15] (as was explained in Section 2), this is not a required assumption for the assessment of misclassification because both the LSA and the PRESO are merged linking subject-time cases, excluding cases where there were missing values for any of the datasets. In addition, for this first descriptive part, we do not limit the study to cases of unemployment; instead we consider every work status that could be used in the LSA and PRESO questionnaires.

A strong caveat needs to be made here though. Due to the time component that characterizes the retrospective question and the additional problems of omission and misdating of spells derived from it (Section 4.1 and 4.3), misclassifications between categories shouldn't be understood as they would be in a cross-sectional setting. Within-subject errors are probably not independent and previous classification errors can determine the probability of misclassification at a particular time-point. That is, evidence of misclassification between statuses presented in this descriptive analysis has to be regarded with caution because they are affected by previous errors which can have a cumulative effect.

This difficulty of establishing "pure" misclassification of status in retrospectively reported work histories might be why this type of ME has been less explored than problems of omission or misdating. Pyy-Martikainen and Rendtel (2009, p.139) acknowledge this gap, *"Even though spell outcomes may also be misreported (e.g. misclassification of a transition out of labour force as a transition to employment), this topic has received little attention in the literature*". However, the same authors recognise the difficulty of linking survey and register cases at the spell level. As an alternative they suggest limiting the study to those subjects who have one spell of unemployment

[15] If they are entitled to receive any kind of benefit that doesn't expire it would be reasonable to assume so.

in both the survey and the register. We were not able to replicate this approach in our study because of the reduced sample size[16]. Instead we contemplate the whole sample first and then look at misclassification for spells reported on the first and last days of the time frame.

We start the analysis of misclassification presenting a simple crosstabulation for the original categories that were used in LSA-1993 and PRESO. We analyse only LSA-1993 and not LSA-2001 because the shorter time frame for LSA-1993 reduces the extra noise derived from lack of match due to errors of spell omission or misdating. As in the previous section, the sample size consists of the 395 subjects who have jointly reported and registered data for 204,587 person-days.

Table A1 in Appendix I represents the entire crosstabulation for LSA and PRESO status. Here, in order to facilitate its interpretation, we present the most informative sections of that crosstabulation in Tables 5 and 6. In addition, we only include column percentages; that is the percentage of person-days found in each cell over the total in that category for PRESO.

In Table 5 it can be seen how categories that are clearly defined engender a better recall. In particular, we include three categories from the PRESO in decreasing order of clarity; labour market training, unemployed and replacement scheme. Arguably the recall accuracy between states of labour market training (77%), unemployment (64%) and replacement scheme (45%) should not be due to biases of social desirability since none of the three categories is illustrative of a status-giving position. 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'.

[16] Only 61 subjects in LSA and 19 in PRESO reported a single spell of unemployment in the 1993 study.

Table 5. Crosstabulation of person-day cases of PRESO and LSA-1993 for three PRESO categories (labour market training, unemployed, and replacement scheme)

PRESO / LSA	Labour market training	Unemployed	Replacement scheme
Unemployed	1,864 (13%)	98,844 **(64%)**	424 (12%)
Employee	928 (7%)	32,594 (21)	1,543 **(45%)**
Job training	**10,847** **(77%)**	6,297 (4%)	1 (0%)
Entrepreneur	0 (0%)	8,192 (5%)	0 (0%)
Homeworker	2 (0%)	1,491 (1%)	0 (0%)
Parental leave	285 (2%)	3,374 (2%)	0 (0%)
Employment development	0 (0%)	89 (0%)	1,421 **(42%)**
Other	153 (1%)	3,921 (3%)	0 (0%)
Total	14,079 (100%)	154,802 (100%)	3,389 (100%)

*Each cell captures the absolute number of person-day cases and between brackets the percentages of those cases over the column total (PRESO total).

As opposed to labour market training, which is a very specific category, unemployment is often confused with other categories such as being out of the labour force or working in informal situations. However, the most illustrative example of misunderstanding as a source of ME can be seen from replacement scheme, a category which respondents seem to identify as a normal job (employee) or as some sort of subsidized work (employment development).

Another interesting pattern (Table 6) can be identified from the better recall seen in permanent jobs (64%) than in part-time employment (61%) and in temporary work (52%).

Table 6. Crosstabulation of person-day cases of PRESO and LSA-1993 cases for three PRESO categories (permanent job, part-time employed, and temporary job)

PRESO / LSA	Permanent job	Part-time employed	Temporary job
Unemployed	1,203 (30%)	2,969 (27%)	1,597 (19%)
Employee	2,532 **(64%)**	6,669 **(61%)**	4,329 **(52%)**
Job training	161 (4%)	203 (2%))	290 (3%)
Entrepreneur	85 (2%)	354 (3%)	614 (7%)
Homeworker	0 (0%)	0 (0%)	80 (1%)
Parental leave	0 (0%)	799 (7%)	281 (3%)
Employment development	0 (0%)	0 (0%)	1 (0%)
Other	0 (0%)	0 (0%)	1,153 (4%)
Total	3,981 (100%)	10,994 (100%)	8,345 (100%)

*Each cell captures the absolute number of person-day cases and between brackets the percentages of those cases over the column total (PRESO total).

Presumably this difference would be derived from social desirability or saliency reasons; however it is hard to distinguish which of the two sources has a bigger effect.

Tables 7 and 8 present results from the 1993 study for the first and last spells in the time frame[17]. Table 7 gives the percentage of subjects that 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 8 represents that same ratio but for the interview dates of each subject (from March to April 1993).

[17] Again only data from the 1993 study is used, but now because the dates for the 2001 study (January 1990 and December 2001) only captured cases from 57 and 51 subjects respectively

Table 7. Misclassification of spells of unemployment in January 1992*

PRESO / LSA	Unemployed
Unemployed	**87**
Employee	7
Job training	2
Entrepreneur	2
Homeworker	0
Parental leave	0
Employment development	0
Other	2
Total	100

*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 independent of errors in the report of subsequent spells. This could be a common effect of event-occurrence frameworks (see Section 3.1) where the questionnaire seems to emphasis the recalls of the status for the first spell, while simply asking to date the start and report the status of following events.

Table 8. Misclassification of spells of unemployment for the day of the interview*

PRESO / LSA	Unemployed
Unemployed	**55**
Employee	25
Job training	7
Entrepreneur	5
Homeworker	2
Parental leave	3
Employment development	0
Other	3
Total	100

*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, which is what we do in the last part of this section.

Person-Day/Month Analysis

We now consider 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 (misclassification) 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).

Table 9 displays the number of units contained in the two time frames, the number of subjects that have contributed at least one time unit to the study (which could be understood as level 2 units) and the number of person-period cases (level 1 units).

Table 9. Sample size, broken down by unit level

Type of Units	1993	2001
Maximum number of days/months	594	132
Number of subjects (level 2 units)	395	383
Total sample size (level 1 units)	205,278	21,658

The total sample size does not equal the number of subjects by the number of time units because most accounts in PRESO were deactivated[18] during certain periods causing gaps in their work histories.

Table 10 displays the number and the percentage of person-period cases that were correctly reported (coded as 0) and those that were misclassified (1).

[18] See Section 2-Data.

Table 10. Frequency and percentage of ME in person-day observations

	LSA-1993		LSA-2001	
	Frequency	**Percentage**	**Frequency**	**Percentage**
0 (Match)	140,319	**68.4**	13,455	**62.1**
1 (Misclassification)	64,959	31.6	8,203	37.9
Total	205,278	100	21,658	100

The percentage of cases correctly reported in LSA-2001 is lower (62.1%) than the one in LSA-1993 (68.4%), which contributes to the validation of hypothesis 1 indicating that recall accuracy deteriorates with time. It only decayed by six points, however, when the time frame is about ten times longer, making the effect less substantively important.

The binary variable describing misclassification of person-period units is the response variable in the following logit models. Since we use a very similar group of explanatory variables, this analysis replicates that of Mathiowetz and Duncan (1988). The main difference lies in the way that non-independence between person-day observations within subjects is dealt with. Whereas Mathiowetz and Duncan (1988) used jacknife replications of their sample in order to calculate the variance of the regression coefficients across all replicates and with that adjust their standard errors, we use a random effects (RE) logit model[19]. In addition to offering a simple way to control for dependence between observations, hierarchical models distinguish between residuals from different levels, and this information can be used to estimate the level of correlation at the subject level.

This model is an extension of the logit model for omission (equation 1), with a new residual term, ζ_i, which captures the unexplained variability for the level 2 units (subjects), and the subscript j is now used to differentiate amongst person-day units

$$y_{ij}^{*} = x_{ij}\beta + \zeta_i + \varepsilon_{ij} \tag{3}$$

[19] A fixed effects model was disregarded because of the substantive interest in regressors that did not vary with time such as gender or type of interview.

The error term ε_{ij} is again assumed to be independent and to follow a logistic distribution with mean zero and variance $\sigma_\varepsilon^2 = \pi^2/3$. The subject specific error term (ζ_i) is also assumed to be independent and follows a normal distribution with mean zero and variance σ_ζ^2. The model is estimated by maximum likelihood, using a Gauss-Hermite quadrature to approximate the integral over the random term ζ_i in the log-likelihood function.

In addition, this model provides information about the intra-cluster correlation (ICC), which captures the correlation among the latent responses by the same person,

$$\rho \approx \sigma_\zeta^2/(\sigma_\zeta^2 + \frac{\pi^2}{3}) \tag{4}$$

These two models use the same set of explanatory variables as the previous models with the addition of a few new ones. A binary variable is now used to capture whether the interviewee was unemployed in specific periods. This variable is used to account for the different composition of times of unemployment in the two time frames, which could hinder comparisons if error-generating mechanisms differ for false positives and false negatives. Two interaction terms were also derived from this variable and included in both models. Moreover, the time component of the response variable was exploited including time-span, a variable that measures the distance of each period from the interview date.

A non-linear relationship was detected for the variable measuring time-span with the proportion of misclassified cases in the sample, and a quadratic term for time-span was therefore included in the two models. Figure 14 and 15 display scatter plots of these quadratic relationships in the two datasets.

Figure 14. Scatter plot of the probability of committing misclassification and time-span from the interview in LSA-1993

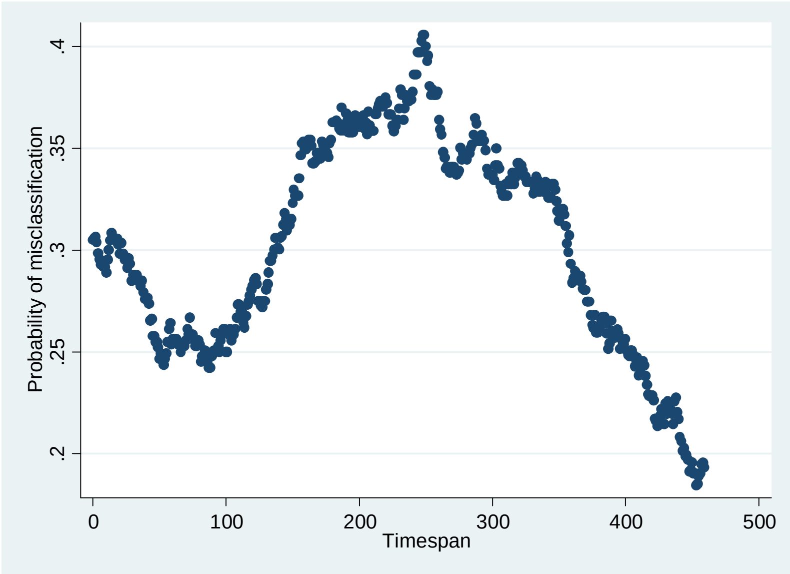
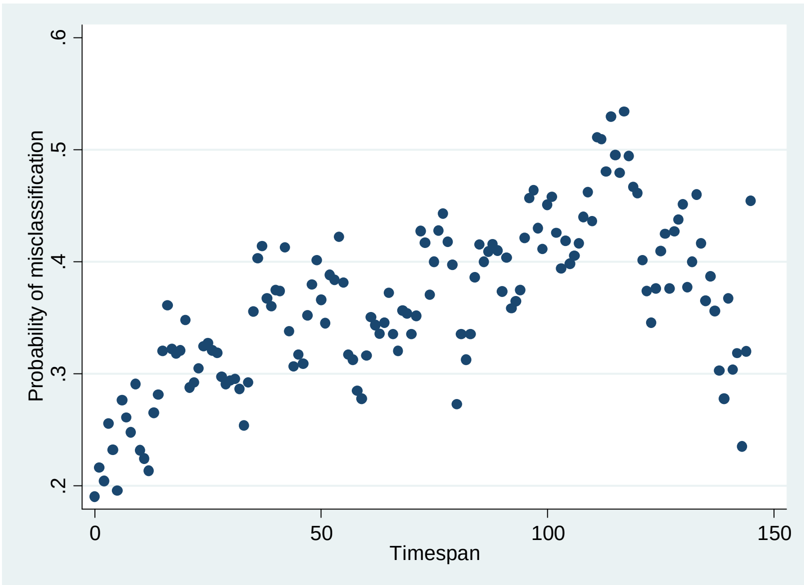


Figure 15. Scatter plot of the probability of committing misclassification and time-span from the interview in LSA-2001



Results of the models[20] for the 1993 and 2001 studies are presented in Table 11 below. We again see the demographic variables (age and gender) are relatively less important than measures capturing saliency such as cumulative unemployment. The inclusion of a variable controlling for registered cases of unemployment and the two interactions stemming from it were also found

[20] The random effects models were replicated using MCMC estimation in MLwiN in order to contrast their robustness. Similar regression coefficients were obtained in both models, although standard errors were higher in the MCMC models.

significant in both models. In addition, the intra-cluster correlation differs remarkably between models, .78 in LSA-1993 and .24 in LSA-2001, and that is in spite of both of them using the same set of explanatory variables. The higher unexplained variance at the individual level in the 1993 model could be due to the shorter time-span and time units that are used.

The variable spells of unemployment (hypothesis 4) shows the expected effect, but, in contrast with what was found in the previous models, it is not significant in either of the datasets. The reason for this might be the inclusion of unemployed, which captures the same information as spells of unemployment for subjects who were only unemployed once.

Table 11. Results from the RE logit model for misclassification in the 1993 and 2001 studies

	1993	2001
Age	-.03 (.02)	-.011 (.007)
Age^2	<.001 (<.001)	<.001 (<.001)
Female	.29 (.38)	-.23 (.13)
Phone interview	.87 (.50)	.24 (.17)
Experience	**-.49** (.05)	.06 (.05)
Cumulative unemployment	**.002** (<.001)	**.024** (.002)
Spells of unemployment	.02 (.02)	.02 (.02)
Unemployed	**1.81** (.15)	**2.18** (.21)
Unemployed X Female	**.22** (.06)	**.22** (.08)
Unemployed X Experience	**-.38** (.04)	**-.31** (.06)
Time-span	**.46** (.01)	.006 (.012)
Time-span^2	**<-.001** (<.001)	<-.001 (<.001)
Level 1 units	153,560	21,240
Level 2 units	395	383
Intra-cluster correlation	**.78** (.01)	**.24** (.02)
Wald chi2(12)	**3,393**	**2,046**

* Results for the regression coefficients represent the untransformed effect of a unit change in the variable on the log-odds of omission.

Hypothesis 2, declaring that groups less embedded in the labour market are more prone to commit errors, cannot be fully supported. Age and gender were not found to be statistically significant in either of the models. However, level of experience has a significant effect in improving recall accuracy in LSA-1993. In fact the interaction term between periods of unemployment and level of experience has a significant effect improving reports in the two datasets. This effect indicates that during spells of unemployment the level of experience has an even stronger effect at improving reports. In other words, it seems that subjects with a more developed career recall their periods of unemployment better than periods spent in other states. In addition, the interaction term between unemployed and female was positive and statistically significant in the two studies, indicating that women do not recall their periods of unemployment as well as they do for other states.

These last results rule out the possibility of a bias of social desirability having an impact in categories of the population more embedded in the labour market. However the overall mechanism of social desirability should not be discarded. The main effect of the variable unemployed is to increase the probability of producing ME, meaning that spells of unemployment have a higher propensity of being misclassified than other spells. Also, the fact that cumulative unemployment was found to be significant in both LSA-1993 and 2001 leads one to think that social desirability bias is more concentrated on the long-term unemployed (hypothesis 3).

The most interesting result stems from the inclusion of the variable time-span. The coefficients for the two time-span terms show the direction that the scatter plots anticipated in the two datasets, and they are found significant in LSA-1993. Contrary to the expected negative effect of time in recall accuracy stated in hypothesis 1, time seems to have a quadratic effect, and observing figures 14 and 15, the overall effect of time seems to affect recall negatively in LSA-2001 but not in LSA-1993.

Our hypothesis that ME increases with distance in time is valid. However when reporting work histories (or any other series of events), 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 probabilities of committing errors in a particular time frame two opposing error generating mechanisms might be at work, with the time mechanism (hypothesis 1) increasing the deterioration of reports the further they are from the date of the interview, and the effect of correlated errors being predominant for the closer periods. This argument on the lack of independence between the errors of different periods contradicts other views on the topic, e.g. Bound (2000, p.68) *"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".* Further research on the topic is required.

5. Conclusion

In this paper we offer a description of the presence of ME in retrospectively collected work histories, and a review of the significance of some the different sources of ME that have been theorized in the literature. We have implemented an original design which acknowledges the different types of ME that retrospective reports of work histories are prone to: miscounting of the number of spells, mismeasuring of spells' length, misdating of spells' starts and misclassification of statuses. In doing so we have combined Mathiowetz and Duncan's (1988) analysis, where only the prevalence of misclassification was modelled, and Pyy-Martikainen and Rendtel's (2009) study where only problems of omission of spells and underreported durations were modelled.

Each of these types of ME are associated with different ways work histories can be characterised. Miscounting the number of spells has direct implications if work histories are to be used in count data analyses (e.g. Poisson model), mismeasuring spell length affects duration data (e.g. accelerated time Weibull model), and misclassification of person-period observations affects data in categorical form (e.g. proportional odds model). Table 12 shows the main results from the descriptive parts of the analysis[21], reporting a combination of the sample means for the survey, register, 1993, and 2001 datasets. The prevalence of ME varies strongly according to how ME is conceptualized. Work histories retrieved from retrospective questions with a recall time frame of one year (LSA-1993) seem to have acceptable levels of reliability when used as count or duration data, since neither reported number of spells nor duration of spells differ more than 20% from what was found in the register. On the other hand, if these work histories are going to be used in a person-period data form, there is a .32 probability of finding either a false positive or a false negative.

---

[21] Results from misdating have not been included because only a graphical analysis was performed.

Table 12. Results of the exploratory analysis of the 1993 and 2001 surveys

	Number of spells		Spell length		Misclassification	
	1993	2001	1993	2001	1993	2001
Mean in PRESO	1.7	8.2	153	456		
Mean in LSA	1.4	3.2	167	238		
Ratio LSA / PRESO	.82	.39	1.09	.52	**.68**	**.62**

In addition, work histories retrieved from retrospective questions using a longer time frame (in this case 12 years) can be very misleading, especially if they are to be used as count or duration data. The average reported total time spent in unemployment is 52.2% of the registered one, and the average number of unemployment spells reported in LSA-2001 is 39% of the one registered in PRESO. These last two results corroborate the simplifying effect detected in Manzoni et al. (2011) who found that extending the recall time frame to 10 years was associated with two thirds of the spells being omitted.

These results contrast with the relative good measures obtained from LSA-1993 that we have seen. However, much narrower differences between the 1993 and 2001 question are found when considering work histories from a person-period perspective. After recoding categories into whether unemployed or not, 68.4% of the person-days units in the 1993 study and 62.1% of the 2001 study matched. The relatively better performance of the 2001 study for person-periods data than for the report of the number of spells (count data) might be due to the use of coarser time-units. The use of months instead of days forced spells shorter than 28 days to be omitted (see Section 3.2 and 3.3), however, the fact that the percentage of matched person-period cases is only six points lower in 2001 than in 1993 in spite of using a time frame almost ten times longer, shows the benefits of using a coarser time unit.

The consistently better results for the 1993 study partially corroborates the first hypothesis that we set to test at the beginning of this section; ME increases with recall time. LSA-2001 has shown poorer performances independently of how ME was defined which supports the view that questions using a longer time frame are more prone to ME. However, when modelling time on the probability of misclassifying status in the 1993 study it was shown that its effect is not linear. That is, periods that are further away from the interview date are not necessarily associated with higher probabilities of being misclassified. This finding agrees with results from Mathiowetz and Duncan

(1988) and calls for a revision of the way this hypothesis is generally formulated in the literature (Bound et al. 2001, Solga 2001).

Hypothesis 2, misunderstanding of categories, stated that groups of the population that are relatively more embedded in the labour market (middle-age men) provide more accurate reports than less engaged groups (young people and women). Here results for the different models vary. Table 13 below presents the results for the six models that have been specified. None of the coefficients for age or gender were found statistically significant in any of the six models. This outcome might however be related to the characteristics of the sample; younger and older subgroups of the population were deliberately omitted in the sample design. These subgroups were the ones found to be more prone to produce ME and the ones that defined the quadratic effect of age on ME in Pyy-Martikainen and Rendtel (2009). The lack of significance in the coefficients for female might be related to lower gender inequalities in Sweden. In 1991, 81.3% of Swedish women from ages 20 to 50 were employed (Korpi and Stern, 2006). On the other hand, the interaction between unemployment and female in the misclassification models corroborates the findings from Morgenstern and Barret (1974), who suggested that because women are less embedded in the labour market they tend to confound spells of unemployment with being out of the labour force.

Alternative evidence on the idea of how being strongly embedded in the labour market helps in the recognition of the different work statuses and hence improve reports can be derived from the effect of the variable experience. The use of this variable is an original feature introduced in our study; it allows us to ascertain whether those people that are well embedded in the labour market actually make better reports without relying on using age and gender as proxies, which is the way it has been approached in the literature so far. Experience was found to be significant and negatively associated with the probability of misclassification in the 1993 study. This result partially corroborates hypothesis 2. In addition, the descriptive study on misclassification points to the fact that, independent of personal characteristics, those categories that could be understood as being less clearly defined showed higher levels of misclassification.

Hypothesis 3, social desirability, indicated that groups of the population more susceptible to feel the social stigma derived from being unemployed (e.g. long-term unemployed), and survey modes that involved more personal contact (face-to-face interview), have a higher chance of omitting, shortening and misclassifying spells of unemployment. The effect of cumulative unemployment which captures the amount of time spent unemployed, shows contradictory results. In the logit models for omission it was found that the longer the time in unemployment the lower the probability of omission. An effect on the same direction was found for the models that consider looks at the differences between the reported and registered durations in unemployment in absolute value (Appendix III). However, the opposite effect was found for the models of underreported time in unemployment and misclassification, which corroborates the results from Pyy-Martikainen and Rendtel (2009), and Mathiowetz and Duncan (1988), respectively.

The contradictory results from the analysis of this variable do not allow hypothesis 3 to be validated or rejected. However, additional evidence supporting hypothesis 3 can be obtained from the descriptive analysis of spell length. Here it was found that long spells of unemployment (>1000 days) in the 2001 study had been underreported. Furthermore, the variable unemployed, capturing registered periods of unemployment, was statistically significant and positively associated with the probability of misclassification.

Regarding the two survey modes, phone interview was only found to be significant in one of the six models; it was positively associated with underreports of unemployment, which is to say that subjects underreported the length of their spells of unemployment by three days when interviewed by phone than when it was face-to-face. This result refutes the hypothesized effect of social desirability, and the general lack of statistical significance between survey modes (telephone vs face-to-face) corroborates what was found in Pyy-Martikainen and Rendtel (2009).

Hypothesis 4 indicated how the more complex the work history is the more difficult it is to report it correctly. This hypothesis is validated by our results since spells of unemployment, the variable capturing the number of spells of unemployment experienced by the subject was found to be positively associated with the different types of ME. In particular it has a strong effect predicting the probability of omitting spells of unemployment. In the 1993 model for spell length it was also found to be significant and in the expected direction. In the 2001 model on omission and in the models for misclassification it was not found to be significant though. Results from this variable and from cumulative unemployment also corroborate Mathiowetz and Duncan's (1988) conclusion where they stated that characteristics of the work histories (length and number of spells) are better predictors of ME than socio-demographics characteristics (age and gender). This effect has been impossible to detect within the literature based on replicated studies because predictors describing the characteristics of the work histories require validation data.

Table 13. Compilation of results from all the models presented in the paper

	Logit for Omission		OLS for Spell Length		RE Logit for Misclassification	
	1993	**2001**	**1993**	**2001**	**1993**	**2001**
Age	-.02 (.02)	-.01 (.03)	-.03 (.03)	-.10 (.07)	-.03 (.02)	-.01 (.007)
Age^2	-.001 (.002)	<.001 (.003)	.001 (.004)	.005 (.008)	<.001 (<.001)	<.001 (<.001)
Female	.37 (.27)	1.04 (.60)	.18 (.54)	2.16 (1.12)	.29 (.38)	-.23 (.13)
Phone interview	.48 (.36)	-.06 (.53)	**1.81** (.71)	-.60 (1.67)	.87 (.50)	.24 (.17)
Experience	-.22 (.19)	.01 (.35)	-.14 (.41)	.08 (.86)	**-.49** (.05)	.06 (.05)
Cumulative unemployment	**-.002** (.001)	**-.05** (.02)	**.016** (.001)	**.006** (.001)	**.002** (<.001)	**.024** (.002)
Spells of unemployment.	**1.41** (.18)	**.83** (.18)	**.80** (.17)	.24 (.17)	.02 (.02)	.02 (.02)
Unemployed					**1.81** (.15)	**2.18** (.21)
Unemployed X Female					**.22** (.06)	**.22** (.08)
Unemployed X Experience					**-.38** (.04)	**-.31** (.06)
Time-span					**.46** (.01)	.006 (.012)
Time-span^2					**<-.001** (<.001)	<-.001 (<.001)
Intra-cluster correlation					**.78**	**.24**
R^2*	.18	.31	.51	.31		
Chi^2**	**89.78**	**71.43**			**3,393**	**2,046**
Level 2 units					395	383
Level 1 units	389	429	334	243	205,278	21,658

*For the logit models the pseudo R2 is given.
**The chi squared values stem from a likelihood ratio test with 7 degrees of freedom for the models on omission and a Wald test with 12 degrees of freedom for the ones of misclassification.

We have used data from retrospective questions on work histories, but we believe that some of the findings could be generalized to many other topics that are collected retrospectively and that are prone to similar ME generating mechanisms. However in order to assess the true generalizability of this study the main problems that affect its validity need to be reviewed. This is what we do in this final part.

First, the sample composition is not entirely representative of Swedish society. In addition to the criteria mentioned in Section 2 (being unemployed in February 1992, older than 25, etc.), there are others that have been imposed by requirements of the analysis. For example, people that decided not to report the specific day (LSA-1993) or month (LSA-2001) of a particular spell were excluded from the analysis on misdating, spell length and misclassification. It might be argued that this decision excluded those interviewees that were more honest and therefore less prone to ME, or inversely it could be said that these interviewees were actually those that invested the least effort and hence the more prone to ME.

Nonetheless the biggest concern regarding the validity of the study stems from whether the assumption of PRESO being a gold standard is plausible. *"Errors of validation can occur if the validation source is erroneous, if the definition used by the validation source differs from the definition inherent in the survey question, or if there are coding or transcription errors […]"*(Duncan and Hill, 1985, pp.509). The PRESO file does not escape these problems

1) Administrative data is doubtless affected by coding errors; PRESO contained 72 subjects that have at least one spell dated to occur previously than the following event in time. For example subject 3 registered one spell in 95-03-02 and the following one in 95-02-02.

2) There are problems in the way unemployment is defined. Discouraged workers consider themselves unemployed, however often they stop visiting unemployment offices, this could be understood as a case of competing definitions of unemployment, although in this analysis we just considered it ME derived from misunderstanding. A more serious problem is derived from the changes in the definition of unemployment by the AMS (see Section 3.2) which might produce artificial variations in the levels of misclassification across time.

3) Moreover, PRESO might be prone to systematic errors if fraudulent practices were extended, e.g. some subjects might be working in the black market and still appear registered as unemployed in order to receive their benefits.

Finally, in order to be able to match spells' dates they need to be measured in the same unit. This is not a problem for the 1993 study since both LSA-1993 and PRESO use days as the time unit. However, LSA-2001 records events in months, and therefore PRESO need to be discretized into months to make the match possible. 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 generates additional 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 unit does not allow short spells to be captured. In particular, 13.7% of the spells of unemployment were lost because they were shorter than 28 days.

Whereas the first problem of errors in the coding can be considered to be randomly generated, the other two points affect particular groups of the population and might display a systematic pattern that endangering the validity of some of the findings. Further research on the accuracy of the registers of unemployment in general and in PRESO in particular would be necessary to make better assessments of ME.

References

Bound et al. (2001), Measurement error in survey data. In: J.J. Heckman and E. Leamer, ed. *Handbook of Econometrics, Volume 5 (*Ch. 59). North Holland.

Duncan, G. and Hill, D. (1985), An Investigation of the Extent and Consequences of Measurement Error in Labor-Economic Survey Data; *Journal of Labor Economics*, Vol.3, N.4, pp.508-532.

Jacob, S. (2002), Reliability and Recall of Unemployment Events Using Retrospective Data; *Work, Employment and Society*, Vol.16, pp.536-548.

Jurges, H. (2007), Unemployment, Life Satisfaction and Retrospective Error, *Royal Statistical Society,* Vol.170, pp.43-61.

Korpi, T., Stenberg, S.-Å. (2001), Massarbetslöshetens Sverige: Arbetslöshetens karaktär och effekter på individers levnadsförhållanden. In: Fritzell, J., Gähler, M., Lundberg, O., ed. *Välfärd och arbete i arbetslöshetens årtionde*, Ch.53.

Korpi, T., and Stern, C. (2006), Women's Employment in Sweden: Globalization, Deindustrialization, and the Labor Market Experiences of Swedish Women 1950 – 2000. In H. Blossfeld, and H. Hofmeister, ed. *Globalization, Uncertainty, and Women's Careers. An International Comparison*. Edward Elgar.

Lawless, J. (2003), Event History Analysis and Longitudinal Surveys. In R. Chambers & C. Skinner, ed. *Analysis of Survey Data*. Chichester: Wiley.

Levine, P. (1993), CPS Contemporaneous and Retrospective Unemployment Compared, *Monthly Labor Review,* Vol.116, pp.33-39.

Levine, E.J. and Wright, R.E. (1996) "Unemployment Insurance and Moral Hazard", *Review of Policy Issues*, vol. 2, no. 2, pp. 63-72.

Manzoni, A., Luijkx, R. and Muffels, R. (2011), Explaining Differences in Labour Market Transitions Between Panel and Life-Course Data in West-Germany, *Qual Quant*, Vol.45, pp.241-261.

Mathiowetz, N. and Duncan, G. (1988), Out of Work, Out of Mind: Response Errors in Retrospective Reports of Unemployment, *Journal of Business and Economic Statistics*, Vol.6, pp. 221-229.

Mathiowetz, N. Brown, Ch. and Bound, J. (2001), Measurement Error in Surveys of the Low-Income Population. In: Ver Ploeg, M., Moffitt, R. and Citro, C. *Studies of Welfare Populations: Data Collection and Research Issues* (Ch. 6). Commission on Behavioral and Social Sciences and Education (CBASSE).

Morgenstern, R., and Barrett N. (1974), The Retrospective Bias in Unemployment Reporting by Sex, Race and Age, *Journal of the American Statistical Association*, Vol.69, N.346, pp.355-357.

Paull, G. (2002), Biases in the Reporting of Labour Dynamics, *The Institute of Fiscal Studies*, Working Paper 02/10, available at: http://econpapers.repec.org/paper/ifsifsewp/02_2f10.htm [accessed 16 December 2010].

Pyy-Martikainen, M., and Rendtel, U., (2009), Measurement Errors in Retrospective Reports of Event Histories: A Validation Study with Finnish Register Data, *Survey Research Methods*, Vol.3, No.3, pp.139-155.

Solga, H. (2001), Longitudinal Survey and the Study of Occupational Mobility: Panel and Retrospective Design in Comparison, *Quality and Quantity*, Vol.35, pp.291-309.

Tourangeau, R. (1984). Cognitive science and survey methods. In T. B. Jabine, M. L. Straf, J.M. Tanur & R. Tourangeau (Eds.), *Cognitive aspects of survey methodology: Building a bridge between disciplines* (pp. 73-100). Washington DC: National Academy Press.

Appendix I: Misclassification in the 1993 study

Each cell captures the absolute number of cases and between brackets the percentages of those cases over the column total (PRESO total).

Table A1. Crosstab of person-day cases in absolute and relative terms for the complete categories of LSA-1993 and PRESO

PRESO / LSA	Replacement scheme	Unemployed	Part-time employed	Temporary job	Permanent job	Public temporary job	Employability rehabilitation programme	Labour market training	Other	Total
Unemployed	424 (12%)	98,844 (64%)	2,969 (27%)	1,597 (19%)	1,203 (30%)	681 (10%)	199 (20%)	1,864 (13%)	6 (1%)	107,787 (53%)
Employee	1,543 (45%)	32,594 (21)	6,669 (61%)	4,329 (52%)	2,532 (64%)	6,028 (84%)	557 (57%)	928 (7%)	21 (2%)	55,201 (27%)
Job training	1 (0%)	6,297 (4%)	203 (2%))	290 (3%)	161 (4%)	368 (5%)	142 (14%)	10,847 (77%)	464 (54%)	18,773 (9%)
Entrepreneur	0 (0%)	8,192 (5%)	354 (3%)	614 (7%)	85 (2%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	9,245 (5%)
Homeworker without remuneration	0 (0%)	1,491 (1%)	0 (0%)	80 (1%)	0 (0%)	17 (0%)	0 (0%)	2 (0%)	11 (2%)	1,601 (1%)
Parental leave without remuneration	0 (0%)	3,374 (2%)	799 (7%)	281 (3%)	0 (0%)	0 (0%)	0 (0%)	285 (2%)	59 (7%)	4,798 (2%)
Employment development	1,421 (42%)	89 (0%)	0 (0%)	1 (0%)	0 (0%)	53 (1%)	0 (0%)	0 (0%)	0 (0%)	1,564 (1%)
Other	0 (0%)	3,921 (3%)	0 (0%)	1,153 (4%)	0 (0%)	13 (0%)	82 (8%)	153 (1%)	296 (35%)	5,618 (3%)
Total	3,389	154,802	10,994	8,345	3,981	7,160	980	14,079	857	204,587

Appendix II. Residuals Diagnosis for the Models on Spell Length

LSA-1993

Figure A1. Qplot of the residuals

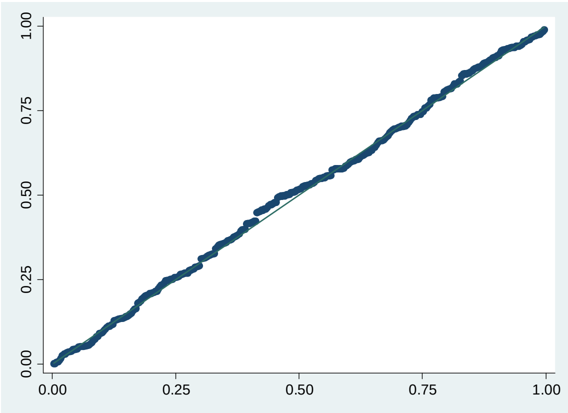
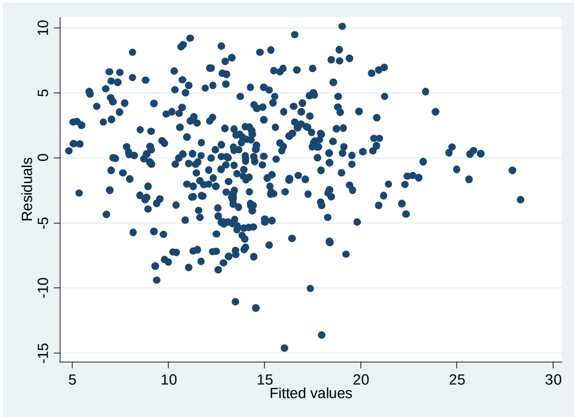


Figure A2. Scatter plot of the residuals vs fitted values



LSA-2001

Figure A3. Qplot of the residuals

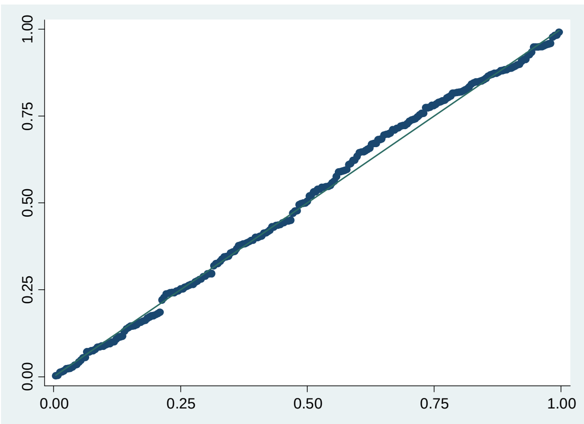
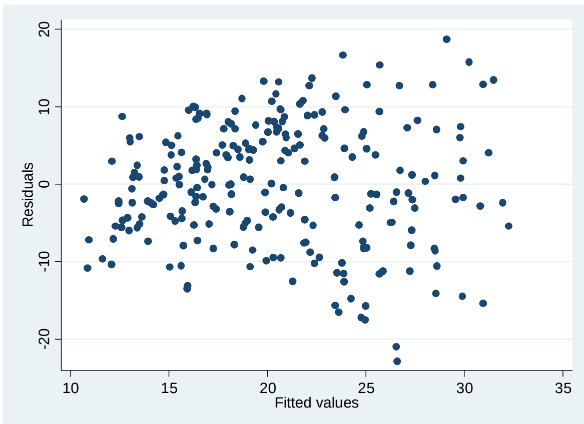


Figure A4. Scatter plot of the residuals vs fitted values



Appendix III. OLS Models for the Difference of Duration in Absolute Value

Table A2. Results for the OLS regression on the differences in reported and registered durations in unemployment in absolute value

	1993	2001
Age	-.04 (.03)	-.09 (.66)
Age^2	.003 (.004)	<.001 (.007)
Female	.20 (.53)	**2.40** (1.09)
Interview format	1.36 (.69)	.81 (1.57)
Experience	-.22 (.40)	-.07 (.87)
Cumulative unemp.	**-.015** (.001)	**-.003** (.001)
Spells of unemp	**.86** (.17)	**.43** (.17)
Constant	8.31	7.80
R2	.46	.20
Sample size	394	319

Appendix IV. Description of the Exploratory Variables Used

Table A3 includes descriptive statistics of the exploratory variables used in the paper. The sample size varied across models; here I have considered the values for the 1993 misclassification model, except for the variables reflecting work histories which have been completed with values for the 2001 misclassification model.

Table A3. Descriptive Statistics of the Exploratory Variables Used in the Misclassification Models

	Mean	Standard Deviation	Minimum/Maximum
Age	37	8	26/55
Female	.33	.47	0/1
Interview format	.15	.36	0/1
Experience	2.4	.69	1/3
Cumulative unemployment 1993*	213	159	1/1329
Cumulative unemployment 2001*	19	17	1/95
Spells of unemployment 1993*	1.6	1.02	1/5
Spells of unemployment 2001*	4	3	1/18
Days 1993	393	158	1/1502
Months 2001	74	42	1/149
Unemployed 1993	.75	.43	0/1
Unemployed 2001	.59	.49	0/1

*The cumulative time in unemployment and the number of spells of unemployment presented here differs from results presented in Sections 4.1 and 4.2 because of the aggregations of PRESO data into months in the 2001 misclassification model, and because of changes in the sample size and the window of observation affecting both 1993 and 2001 misclassification models.