Implications of Retrospective Measurement Error in Event History Analysis

Jose Pina-Sánchez, Johan Koskinen, and Ian Plewis
University of Manchester

Abstract

It is commonly accepted that the use of retrospective questions in surveys makes interviewees face harder cognitive challenges and therefore leads to less precise measures than questions asking about current states. In this paper we evaluate the effect of using data derived from retrospective questions as the response variable in different event history analysis models. Two common specifications in event history analysis are studied, single and repeated events. For each of these specifications four event history models are considered, an accelerated life Weibull, an accelerated life exponential, a proportional hazards Cox, and a proportional odds logit. The impact of measurement error is assessed by a comparison of the estimates obtained when the models are specified using durations of unemployment derived from a retrospective question against those obtained using validation data derived from a register of unemployment. Results show large attenuation effects in all the regression coefficients. These effects are relatively similar across models, while the consideration of repeated events reduces the average size of the attenuations.

Key words:

Measurement error, event history analysis, retrospective question, register data, unemployment.

Contents

1. Introduction

Retrospective questions are a widely used tool in surveys when there is an interest in capturing changes over time. These types of questions ask respondents for information about events from the past. They obtain information on a particular span of time at a single occasion, and are therefore usually cheaper than the alternative approach of repeatedly contacting respondents during that span of time as in 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, for example, changes in the wording of questions over time. Moreover, retrospective questions can capture information on the full history of an event for a particular period of time, whereas repeated questions on current state are only able to provide a series of snapshots. They can only capture within wave transitions if retrospective questions are included at each wave[1].

The major problem for retrospective questions stems from their higher propensity to generate measurement error (ME) 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.

In this paper we study the implications of using data collected from retrospective questions in statistical models used for longitudinal data. In particular we consider the consequences of using data derived from these questions as the response variable in event history analysis (EHA) models. Within the field of statistics, EHA could be defined as the branch of longitudinal data analysis where discretely defined outcomes are studied. The impact of ME is assessed by comparing estimates obtained from models that are specified

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

using durations of unemployment derived from a retrospective question against those obtained using validation data derived from a register of unemployment.

In choosing to study the consequences of ME in the response variable of EHA models we address an area which has not been widely researched. In the analysis of ME a majority of studies have focused on settings where the explanatory variables were the ones prone to ME, in what is known as the "errors in variables" problem. This focus on the predictors can be explained from the widespread consideration that ME affecting the response variable only affects the precision of the model and therefore it is a lesser problem. In addition, the study of ME was until recently restricted to analyses using linear models, with the seminal work of Fuller (1987) as the main reference. In the last decade the study of ME has been extended to other non-linear models, especially after the publication of Carroll's et al. (2006) work. However, the study of ME in EHA models has been identified by many authors as an area in need of further contributions, (Augustin (p. 2, 1999), Pyy-Martikainen and Rendtel (p.140, 2009), Skinner and Humphries (p. 23, 1999), and Jäckle (p.2, 2008)).

This paper is structured as follows: in the next two sections we examine the statistical models that are used in our analysis. The EHA outcome models are presented first in Section 2, which is divided in three subsections, each one dedicated to the different families of EHA models (parametric, semi-parametric, and non-parametric), and a final subsection introducing the differences between EHA models that consider one or more than one spell. In Section 3 we introduce the ME models which have been most often used in the literature; these models are subsequently criticised in 3.1 because of their inadequacy to reflect ME derived from retrospective questions. In Section 4 we present a summary of findings from other studies in the literature. In Section 5 we describe the characteristics of the data that we use in our analysis, and in Section 6 we present results from our analyses. Results from models for first spells are included in 6.1 and results from models for repeated events come in 6.2. Finally, we conclude in Section 7 with a summary of the results and how these relate to previous findings in the literature.

2. Event History Analysis

The group of techniques used in EHA are also known by the names of time to event, survival or duration analysis, depending on the scientific discipline where they are used. Regardless of which field these techniques are applied to, they all study the spells of time in a particular state of interest until a transition from that state is made. There are two fundamental aspects of the study of times-to-event: the survivor function and the hazard rate.

Here we present these two concepts using a mathematical illustration taken from Box-Steffensmeier and Jones (2004). Let the durations (the time spent in a particular state) be denoted by *T*, a positive random variable, which for the time being will be assumed to be continuous. Let realizations of *T*, denoting the duration of a particular unit be denoted by t. If these values are ordered by size a cumulative distribution function, $F(t)$, can be established as follows,

$$F(t) = \int_0^t f(t)d(t) = Pr(T \leq t)$$

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

$$f(t) = \frac{dF(t)}{d(t)}$$

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

$$f(t) = \lim_{\Delta t \to 0} \frac{Pr(t \leq T \leq t + \Delta t)}{\Delta t}$$

Hence, when a variable capturing durations is analysed, the cumulative distribution function indicates the probability of observing a case lower or equal to time t, while the density function gives the unconditional failure rate in an infinitesimally small differentiable period of time, that is, the proportion of individuals making a transition from one state to another.

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

$$S(t) = 1 - F(t) = Pr(T > t)$$

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

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

$$h(t) = \frac{f(t)}{S(t)}$$

Thus, the hazard rate gives the rate at which the unit's duration ends by *t*, given that the unit had survived until *t*;

$$h(t) = \lim_{\Delta t \to 0} \frac{Pr(t \leq T < t + \Delta t | T \geq t)}{\Delta t} \quad (1)$$

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

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

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

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

$$\hat{S}(t) = \prod_{t_j<t} \frac{n_j - d_j}{n_j} \quad (2)$$

where d_j represents the number of failures in a particular time t, the subscript j is used to identify subjects, $j = 1, \dots, N$, while the meaning of n_j differs according to whether censored cases are present in the dataset. Censored cases are those for which the transition from their original state has not been recorded before the end of the window of observation. When there is no censoring, n_j is just the number of survivors just prior to time t_j; with censoring, n_j is the number of survivors less the number of censored cases.

The survivor function and hazard rate are fundamental in the exploration of time-to-event processes, but they are only descriptive statistics, that is they only measure parts of the process without controlling for other variables. When interested in ascertaining the conditional association of different variables with the observed durations, EHA models need to be used. Most of them are developed from a specification of the hazard rate, which can be treated as having a dependency on time as well as a dependency on the regressors, denoted by x. Then we can re-express the hazard rate from equation 1 as

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

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

Any model used for the analysis of event data can be characterized by two features, the nature of the response variable and whether a specific statistical distribution is used. The first feature distinguishes two groups of EHA models, accelerated life models (AL henceforth) and hazard models. The former group models the observed event time (failure time) directly, whereas the latter models the hazard rate. In terms of statistical distributions, EHA models can be classified into three families: parametric, semi-parametric and non-parametric models. We now review the models that are studied in Section 6 using this last distinction.

2.1. Parametric Models

The basic logic underlying parametric event history models is to directly model the time

dependency exhibited in event history data. This is done by specifying a density distribution for the failure times. The exponential, Weibull, Gompertz, or Gamma models are examples of models belonging to the parametric family, although in our analysis we focus solely on the Weibull and exponential models. If the Weibull distribution is used the baseline function can only be monotonically increasing, monotonically decreasing, or flat with respect to time. The hazard rate is then expressed as;

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

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

$$h(t|x) = \alpha(t)^{\alpha-1}exp(\beta'x) \quad (3)$$

and now the scale parameter would be $exp(\beta'x)$, and the shape parameter is α. The impact of the covariates is to alter the scale parameter, while the shape of the distribution remains constant.

The Exponential and the Weibull model are members of the family of proportional hazard (PH) models; that is, unit changes in the covariates will imply constant changes in the risk of event occurrence in a direction and magnitude indicated by their slope coefficients. However, these models could also be expressed as modelling the log of the event duration, which makes them accelerated life (AL) models. The Weibull model would then be expressed as;

$$log(T) = \beta'x + \sigma\epsilon \quad (4)$$

where ϵ is a stochastic disturbance term with a type-1 extreme value distribution scaled by σ. Here, in contrast with their hazard specifications, slope coefficients are indicative of the change in log-durations (the logs of the expected life time) for different values of the covariates.

The specification for the exponential PH model is shown in equation 5 below, which differs from the PH Weibull model defined in equation 3, in its use of a constant hazard rate,

$$h(t|x) = h(t)exp(\beta' x) \tag{5}$$

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

$$log(T) = \beta' x + \epsilon \tag{6}$$

with a disturbance term ϵ following an extreme value distribution with mean zero and variance 1, which in this case is not affected by a scaling factor

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

2.2. Semi-Parametric Models

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

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

$$h(t|x) = h_0(t)\mathrm{e}xp(\beta' x) \tag{7}$$

The baseline hazard function is left completely unspecified, whereas, like the Weibull and exponential models, the regressors are assumed to have proportional effects on the hazard rate. The estimation of such a model is possible using partial likelihood, a variant of maximum likelihood where only the order of the failure-times is taken into account. As

such, each interval between two failures is modelled separately. For these two reasons the Cox model is often categorised as a semi-parametric model. Furthermore, because of its reliance on partial likelihood the Cox model can be extended easily to allow for explanatory variables that change in value over time.

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

2.3. Non-Parametric Models

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

$$y_{ti} = \begin{cases} 0 & t < y_i \\ 0 & t = y_i, \delta_i = 1 \\ 1 & t = y_i, \delta_i = 0 \end{cases}$$

The discrete-time hazard for interval t is the probability of an event during interval t, conditional on both the fact that the event has not occurred in a previous interval and on the set of covariates included in the model.

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

Since the dependent variable is binary, constructing models relating this variable to the covariates involves selecting one from a variety of suitable distributions for binary data. Two commonly used link functions are the logistic distribution and the standard normal

distribution. The use of these distributions gives rise to the logit and the probit models, respectively. Here we will use the former, which transforms the response variable into the log of the odds-ratio of failure,

$$log\left(\frac{\gamma_{ti}}{1-\gamma_{ti}}\right) = \beta' X_{ti} \tag{8}$$

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

When using the logistic transformation this model is also known as the proportional odds (PO from now on) model because as in the proportional hazards model, changes in the regressors are assumed to induce proportional changes in the response variable. In addition, on the right hand side of (8), besides the set of explanatory variables of interest, a series of temporal dummies are included in order to specify the baseline hazard function. Each of the dummy variables included represents a period of the time-frame, in what is called a piecewise-constant hazards model. The additional set of parameters required to estimate this baseline hazard function reduces the degrees of freedom, affecting precision, which is a weakness of non-parametric models.

2.4. Single vs Repeated Events

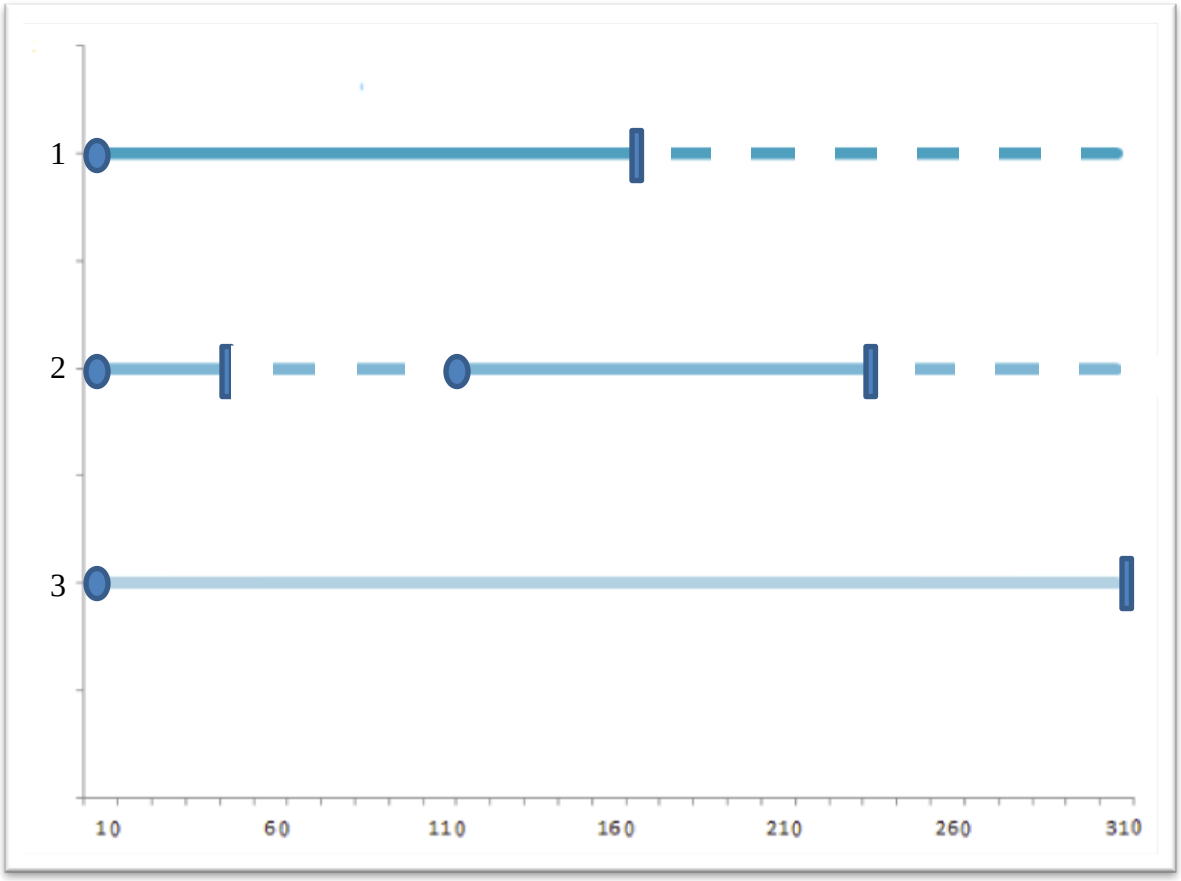
The descriptive statistics and models presented here refer to a particular setting of EHA, one where only the first transition from the state of interest is contemplated. Some life-course events are bound to have such a design, e.g. the study of time to first marriage. However, there are other settings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

To illustrate the difference between designs for single and repeated spells we use Figure 1 below, which represents the three first cases from the sample of work histories captured in the retrospective question that we use in the analysis (see Section 5). Time in unemployment is represented by a continuous line. States different from unemployment (employed, out of the labour force, etc.) have been aggregated into a single category and they are denoted by a dashed line. In both the setting for single and repeated events the first case shows a transition at day 160, and the third is right censored since it remained in unemployment until the end of the window of observation. The only difference is found

in case 2, for which the repeated events model would consider the first spell of unemployment terminating at day 40 and a second spell spanning from days 110 to 230.

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

Figure 1. Exemplar durations of three work histories



However, the use of repeated events models is not always straightforward. In particular, careful consideration needs to be given to issues of dependency of events within subjects, and to whether the repeatable events follow a sequential order, or whether they are out of order. For example, regarding the order of events, in a study of time to marriage we might

expect some covariate effects such as partner's wealth to have a stronger effect on subsequent marriages. Regarding the problem of dependency between spells of the same subject, two solutions can be explored. We can use marginal models which treat this dependency as noise, or we can use random effects to model it.

Marginal models maintain the assumption of an independent correlation structure, regardless of what the true structure is. These models concentrate on adjusting the estimated variance of the regression coefficients which[2], in the presence of dependency between observations, would be systematically underestimated. The most widely used variance adjustment method applied in marginal models is the "sandwich" variance estimator. Williams (2000) demonstrates that this is an unbiased estimator for cluster correlated data. It is calculated as the product of three matrices: the matrix capturing the observation-level score vectors[3] (the meat of the sandwich), which is pre- and post-multiplied by the model-based variance matrix (the bread of the sandwich).

Multilevel models specify the dependence structure through the inclusion of random effects. Survival models with random effects are also known as frailty models, or shared frailty models when there is more than one observation per group or cluster. These models are often used when interest is in modelling the heterogeneity arising from unobserved covariates which tends to attenuate baseline functions as subjects drop out the risk set.

In its simplest form only one random term is introduced to represent the possibility of random intercepts[4], that is, different intercepts for each subject. The random intercept (RI) is shared by all spells experienced by the same individual and can be interpreted as subject level unobserved heterogeneity. After controlling for the individual-specific unobservables represented by the random effect, we assume that the durations of episodes for the same individual are independent (Steele, 2005).

[2] This is not necessarily the case for models using generalised estimating equations, which are semi-parametric and therefore less sensitive to variance structure specifications than likelihood based models.
[3] Derived from the maximum likelihood estimation process in the parametric and non-parametric models, or from pseudolikelihood in the PH Cox model.
[4] See Steele (2008), and Kelly (2012) for a good review of how multilevel modelling can be applied to longitudinal data analysis, and the software available to estimate these types of models, respectively.

This extension can be formally illustrated using the PO logit model specified in equation 8 as an example. In particular, φ_i, the RI term, is introduced on the right-hand side of the model, and it is supposed to be Normally distributed with a mean of zero and constant variance. So the RI PO logit model is defined as follows,

$$log\left(\frac{\gamma_{ti}}{1-\gamma_{ti}}\right) = \beta' X_{ti} + \varphi_i$$

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

3. Measurement Error Models

ME models are those where an observed variable that is prone to error is specified in relation to the true variable from which it stems, and the error term that contaminates it. The classical additive model is by far the most frequently used model in the study of ME. It is simple, it applies to many settings where the ME is supposed to be random, and it is assumed by many of the methods designed for the adjustment of ME. In addition, it serves as the foundation upon which more sophisticated models can be built.

The classical model was expressed by Novick (1966) as a way of operationalizing his idea of the classical test theory: *"By classical test theory we shall mean that theory which postulates the existence of a true score, that error scores are uncorrelated with each other and with true scores and that observed, true and error scores are linearly related."* (Novick, p. 1, 1966). This can be formalized as follows,

$$X^* = X + U \tag{9}$$

where X^* is the observed variable, which is equal to its equivalent true variable, X, plus the ME term, U, that is normally distributed. The imposition of an additive relationship between the true variable and the error term is a specific one from the classical additive model, although other forms can be used within the classical framework. In fact, what classifies a ME model as classical is the following set of five assumptions regarding the ME term:

$$Classical\ Model \begin{cases} E(U) = 0 \ ; & null\ expectancy \\ Var(U_i) = Var(U) \ ; & homoskedasticity \\ Cov(X, U) = 0 \ ; & indep.\ error\ and\ true\ value \\ Cov(U_i, U_j) = 0 \ ; & independence \\ Cov(\Psi, U) = 0 \ ; & non - differentiality \end{cases} \quad (10)$$

1. Null expectancy refers to the assumption that the error term is non-systematic, or in other words, the expected value of the error term is zero, $E(U) = 0$.

2. The assumption of homoscedasticity indicates that the variance of the error term is assumed to remain constant across the sample, $Var(U_i) = Var(U)$.

3. The third assumption specifies that the correlation between the true value and the error term is assumed to be zero, $Cov(X, U) = 0$.

4. Furthermore, the correlation between different values of the error term is also assumed to be zero, $Cov(U_i, U_j) = 0$, where U_i, U_j represent any two values of the error term, for subjects i and j.

5. The last assumption, non-differentiality, only becomes relevant when X^* is used in a regression model. It indicates that, given the true value, the ME is not associated with the remaining variability in the response, $E(Y|X, X^*) = E(Y|X)$, or alternatively, $Cov(\Psi, U) = 0$, where Ψ represents the residual term from the regression model.

It is well known that the inclusion of variables prone to ME in regression models produces bias in estimates of the regression parameters, although Fuller (1987) demonstrated that this is not the case when classical ME affects the response variable in a linear model. If, instead of observing the true response variable Y, a different variable Y^*, subject to classical additive ME is observed,

$$Y^* = Y + V$$

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

$$Y^* = \beta_0 + \beta_1 X + (\Psi + V) \quad (11)$$

A similar result holds for other measurement error models such as the classical multiplicative. This is an extension of the classical model presented in (equation 9), based

on similar assumptions, but where the additive relation between the true variable and the error term is replaced by a multiplicative one,

$$Y^* = Y \cdot V \tag{12}$$

Some researchers (Holt, McDonald, and Skinner, 1991, Skinner and Humphreys, 1999, Augustin, 1999, and Dumangane, 2007) have suggested the use of classical multiplicative models in order to better specify MEs produced by memory failures. Here the same assumptions about the error term (equation 10) apply, with the distinction that $E(V) = 1$ and the error term uses a distribution bounded from 0 to ∞ such as the log-normal distribution. The multiplicative relationship is then used to take into account the possibility that longer durations are affected more intensely by the error-generating processes[5].

Skinner (1999, 2000) illustrates how classical multiplicative errors do not produce a bias in the systematic parts of AL models, only in the estimation of the distribution of the durations. Following the same rationale as in the case of classical additive errors affecting the response variable seen above, and for the case the AL exponential model (equation 6), we can see that the impact of classical multiplicative ME resides in the stochastic part of the model,

$$logY^* = \beta_0 + \beta_1 X + (\in + logV) \tag{13}$$

The problem is that contrary to what is commonly assumed this result cannot be safely generalized beyond the simple settings used to derive equations 12 and 13. When the outcome model is non-linear or the type of ME is not classical or both, there is a strong chance that modelling a response variable prone to ME is going to bias the regression coefficients.

One example where the classical model is entirely inadequate is in those cases where the unit of observation is formed by binary data as in the PO logit model. Specifically, the ME model where an error term is added to the true value would not make sense because categorical variables lack a scale. Here, the problem of ME becomes one of misclassification (MC henceforth), and the ME should be specified as the probability of

[5] This argument is elaborated in Section 3.1.

correctly seeing each of the categories of the variable. In the case of binary variables these are referred to as sensitivity and specificity,

$$\begin{cases} P(x^* = 1|x = 1) = \theta_{1|1}\ ; & True\ positive\ (Sensitivity) \\ P(x^* = 0|x = 0) = \theta_{0|0}\ ; & True\ negative\ (Specificity) \end{cases}$$

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

$$\begin{cases} P(x^* = 0|x = 1) = \theta_{0|1}\ ; & False\ negative \\ P(x^* = 1|x = 0) = \theta_{1|0}\ ; & False\ positive \end{cases}$$

Together these four probabilities can be used to specify a MC model for a binary variable, and they are summarized in the misclassification matrix, represented by θ in the equation below

$$\theta_{X^*|X} = P(X^* = x^*|X = x) \tag{13}$$

We continue this section with a presentation of the reasons why ME processes stemming from retrospective questions with a longitudinal component depart from the classical framework.

3.1. Types of Measurement Errors in Retrospective Reports of Unemployment

First of all, when considering ME derived from retrospective questions it is important to recognise the differences between these types of questions. There are retrospective questions which generate answers that can be operationalized as a single variable in the form of duration or count data, such as the number of times a spell was experienced in a particular time-frame, or the duration of an specific event. In these instances, the classical multiplicative model might be an appropriate ME specification. For example, if the number of spells of unemployment experienced in the last 12 months is asked, it could be expected that interviewees who have only had a few of them, say 0, 1, or 2 spells, might offer more accurate reports than those who experienced a higher number. Similarly, for questions asking about the time since last employed, it could be expected that workers recently made unemployed will recall more accurately than those who have been unemployed for a longer time. In other words, when the question asked requires an

answer that can be represented by a single number, memory failures might make the recall of high figures less precise, and thus, the classical multiplicative model becomes an appropriate specification for ME.

The problem of finding an appropriate ME specification becomes less straightforward for questions that request to report an entire event history for a specific period. These questions require ordering and dating events as they are reported[6], which makes the error generating mechanisms more complex. Pina-Sánchez, Koskinen, and Plewis (2012) distinguish between three types of ME typically stemming from retrospective questions where work histories are reported: miscounting the number of spells, mismeasuring spells' length, and misclassifying spells' categories.

We illustrate the differences between these types of ME graphically in Figure 2 below. Here, we have taken the three work histories presented in Figure 1 where every subject starts from the same category (unemployed) and only first spells are considered. We assume that the work histories presented in lines 1, 2, and 3, are the true ones, and are denoted by Y. The error term is represented by V and it is encompassed by the bracket immediately below, whereas the observed duration is represented by Y^*.

When spells are mismeasured the observed durations can appear shorter or longer than the true ones. This is represented by the first case shown in Figure 2, where the only spell of unemployment experienced within the window of observation has been reported to be 70 days longer than it really was. Mismeasurement errors are therefore no different from the ME affecting the simpler type of retrospective questions discussed before, and there is no reason to believe that they cannot be adequately specified under the classical multiplicative model

Miscounting the number of spells can result in omitting or over-counting spells. Over-counting spells is not a problem when using EHA models for single spells, since any reported spell beyond the first transition is not considered by the model. However, the omission[7] of spells could distort estimations based on this data. Take the second work

[6] See Section 5 to find out how these questions can be phrased.
[7] Levine (1993), comparing retrospective questions using a one year recall time with questions asking about the current work status, found that between 35% and 60% of persons failed to report at least one spell of unemployment.

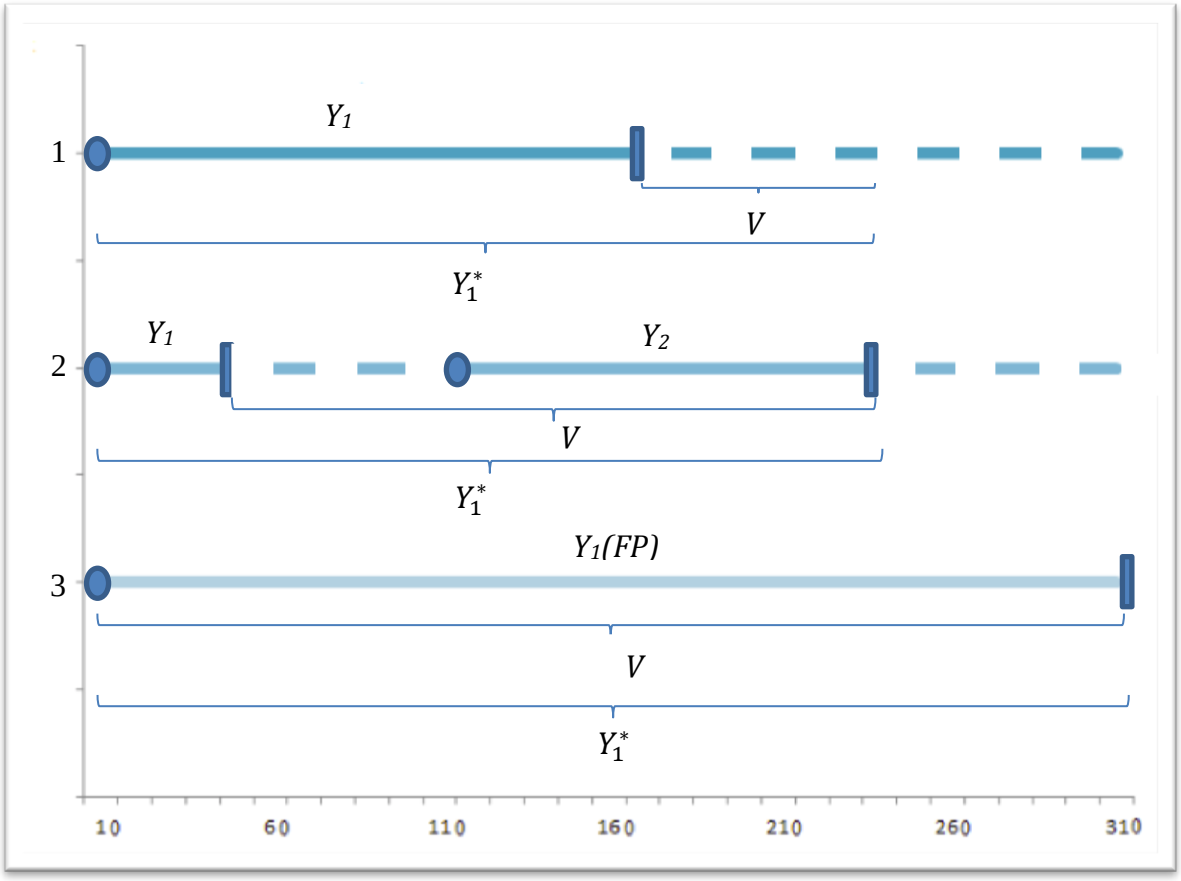
history presented in Figure 2: if the spell representing a different state than unemployment starting in day 30 was omitted, the two spells of unemployment that occur before and after would be linked, and the reported work history for that subject would look like a unique spell of unemployment. The consequences of these types of errors are twofold. The magnitude of ME as a proportion of the true duration is potentially larger than what is seen in errors derived from problems of mismeasurement. On the other hand, the assumption of independence between the true duration and the ME used in the classical framework is violated ($Cov(Y,V) \neq 0$). Given the fixed time covered by the window of observation, the longer the first spell of unemployment is the lower the probability of omitting subsequent spells.

Finally, we turn to problems of MC spells taking the form of reported false positives[8] (FP hereafter) and false negatives (FN). In FP cases the observed duration is entirely formed by the error term, $Y^* = V$; this is represented by the third work history shown in Figure 2. The fact that the observed duration is not formed by a combination (either additive or multiplicative) of true duration and noise renders the classical framework inadequate. In addition, for the first spells setting, FP cases represent an artificial increase of the sample size. Given our setting where only work histories that start from unemployment and first spells are considered, when FN are present the problem becomes one of missing data because FN durations are not observed. However, as long as the probability of committing FN is independent of the duration of unemployment (missing completely at random) it will only reduce the precision of the EHA model estimates.

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

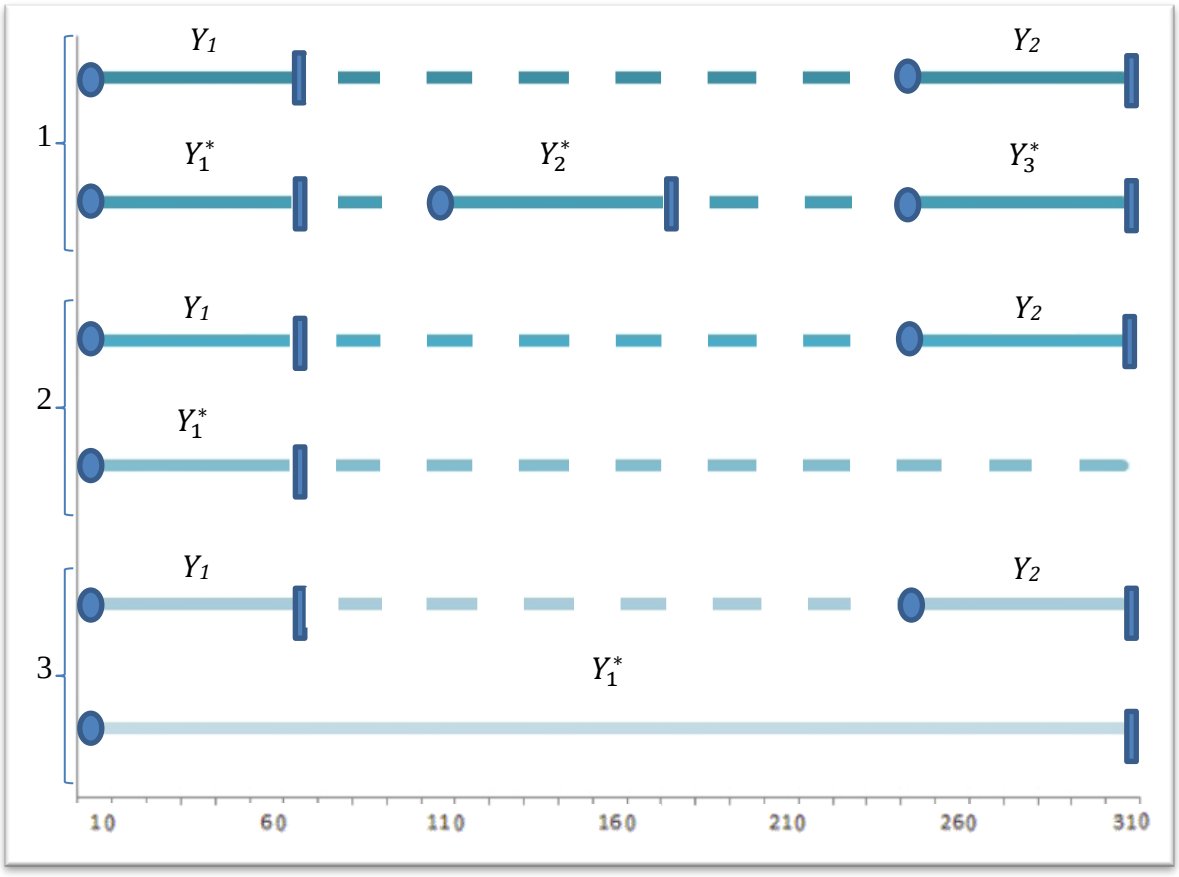
[8] Bound (2001) in a review of the literature concludes that in cross-sectional surveys 11-16% of respondents stated to be unemployed are likely to be misclassified.

Figure 2. Work histories affected by ME in a First Spells Setting



These are presented in Figure 3 below, where the second case from Figures 1 and 2 is used to illustrate them. The first case in Figure 3 shows an example of over-inclusion. This situation would lead to wrongly considering an additional spell of unemployment. The second case shows the second spell of unemployment being omitted; in this case the repeated nature of the work history would be lost. The third case shows an omission of the spell different from unemployment, which would result in linking the first and second spell of unemployment in a single one encompassing the whole window of observation. So, just like in the FP case described in Figure 2, all of these problems of miscounting will now represent a misspecification of the classical model since they are not the result of a combination of true durations and noise, but just the latter.

Figure 3. Work histories affected by ME in a Multiple Spells Setting



To sum up, retrospective questions are problematic not only because of the higher prevalence of ME in their answers, but also because of the more complex forms that this ME can take. In particular, we have shown that in questions that involve ordering and dating event histories such as unemployment, different forms of ME arise, which render the classical ME framework inadequate. Finally, we have pointed out that some forms of ME such as over-reporting spells of unemployment or FN, could be expected to have a bigger effect in EHA models for multiple spells than when only single spells are considered.

4. Literature Review

According to the research design used, we can identify two main groups of studies which

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

The only studies that have explored analytically the impact of ME on the response variable in EHA models are the working papers by Augustin (1999) and Dumangane (2007). They used AL Weibull models and assumed classical multiplicative errors affecting the recall of durations. In this case ME in the response was found to produce an attenuation bias in the regression coefficients. However, this particular setting does not account for other types of errors observed in retrospective questions on work histories, such as omission of spells, or misclassification of status (we examine this topic in detail in the next section). In addition, Augustin (1999) requires the assumption of no right censoring in the data and Dumangane (2007) assumes that the true duration and error distributions are independent. The set of assumptions used in these papers shows both the difficulty of studying the effect of ME in the response variable of EHA models analytically, and how the general expressions developed so far are not really representative of the problems found in retrospective data, which are prone to other types of ME besides mismeasured durations.

Another group of studies assessing the impact of ME in EHA are those that have carried out an empirical analysis. These studies compare estimates derived from a model that uses prone to ME data against the estimates obtained from replicating the same model but using data free of ME. Korn et al. (2010) studied the effects of ME in a PH Weibull model by means of simulating multiplicative log-normal errors in the response. The authors found small downward biases in the hazard rate as long as the ME remains non-differential and hazard rates relatively high. However, by simulating different levels of ME in the control and treatment groups, the authors also demonstrated that the degree of

attenuation augmented substantially when the ME was differential.

Considering non-parametric models for discrete data Meier, Richardson and Hughes (2003) assess the bias in the regression coefficients produced by simulating different levels of non-differential FP and FN. The authors conclude that the bias is always toward the null, and that FPs induce greater bias in estimation of the cumulative distribution function and regression coefficients than FNs when the failure rate is low. Moreover, for this case of EHA (non-parametric models using discrete data), additional findings can be obtained from studies on the impact of misclassification in the response variable in more general models for categorical data. In this respect, Magder and Hughes (1997) show how a response variable subject to MC could generate bias in the regression coefficients of a logistic regression. Neuhaus (1999) derived general expressions for the magnitude of the bias due to MC in the response for different type of regression models for binary data (logistic, probit, complementary log-log). The authors found that ignoring response MC leads to attenuated covariate effects when the errors are independent of the covariates. However, when MC probabilities depend on covariates, ignoring these errors can lead to bias away from the null.

An early attempt to look at the effect of retrospective ME in the response in EHA models was Holt, McDonald, and Skinner (1991). Here, the authors compared two AL Weibull models where duration of unemployment was regressed on age. A sample of durations was simulated and different types of differential and non-differential multiplicative ME was superimposed on them. A similar study was carried out for estimation of age at menarche. The comparison of free from ME with prone to ME models shown biases in the coefficient of age and in the baseline hazard function. A bias towards the null was found when the model of unemployment was used, while for the model of age at menarche that bias was in the opposite direction. In both studies the bias increased when the ME was correlated with age (the covariate). In addition, the baseline function was always underestimated.

Skinner and Humphreys (1999) used the example of age at menarche with no right-censored cases presented in Holt et al. (1991) but trying different types of ME. The authors simulated different types of non-differential errors (additive vs multiplicative,

homoscedastic vs heteroskedastic, in different combinations) on both the durations in unemployment and for age at menarche. Their findings proved that under the assumption of no censoring, the regression coefficients of a Weibull model are approximately unbiased when MEs affecting spells are independent of each other, of the spell durations and of the covariates. The estimator of the shape parameter that determines the duration dependence of the hazard is, however, biased. The authors traced the effects analytically and noted that if ME is related to covariates then the estimators of the corresponding coefficients are likely to be biased.

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

Other interesting studies that represent the effect of retrospective ME more closely are Peters (1998) and Jäckle (2008). The former compares survey data captured from prospective (questions measuring current states) and retrospective questions on different life-cycle events, time to first marriage, and time to first divorce. The durations of the two events are specified using a PH Weibull model. The regression coefficients for the models that were run on the retrospective data differ only slightly with respect to the ones obtained using prospective data. Jäckle (2008) used retrospective data and compared it to a gold standard (obtained from the "Improvement of Survey Measurement of Income and Employment" study). The author found that ME in the reporting and dating of receipt of unemployment benefits attenuated both the duration dependence and the regression coefficients from a Weibull model. The recall period used by the retrospective question was four months though, which is perhaps not long enough for the typical memory failures that characterize retrospective data to be seen.

One last study that used the more common recall frame of one year was the one from Pyy-Martikainen and Rendtel (2009). Here the authors compared data derived from a retrospective question on work status against a gold standard obtained from the Finnish register of unemployment. PH Cox and Weibull models for unordered repeated events

were specified for the duration of unemployment and both attenuation and augmentation bias were found in the regression coefficients. None of these biases changed the survey estimates by more than 30%, and they were found in the same direction and similar magnitude for both the Cox and Weibull models. Moreover, the comparison of the Cox and Weibull models shows that the baseline hazard was more accurately estimated by the former. The survey baseline hazard from the Weibull model is nearly constant while the register baseline hazard shows positive duration dependence leading to erroneous conclusions about the duration dependence; whereas the Cox baseline hazards from survey and register both display positive duration dependence.

The authors also studied some interesting extensions; first, they discretized the duration data and specified a cloglog model; second, they reran the PH Cox models but using a cause-specific frame which only considered spells from unemployment to employment. The results for these two settings were again very similar to what was obtained before; some of the bias was accentuated for the repeated spells models but their directions were still the same.

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

5. Data

The data we use has been obtained from the “Longitudinal Study of the Unemployed” 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[9]). This register provided individual-level data on the work status of the participants of three surveys, run in 1992, 1993 and 2001. The three surveys are relatively similar with respect to the composition of both the sample of participants and the questionnaire. The target sample was composed of subjects registered as unemployed on 28th February 1992 from ages 25 to 55.

In this study we use data derived from a retrospective question on work status from the 1993 survey. This question uses an event-occurrence framework; Lawless (2003) coined this term to define questions where events are asked to be reported in order of occurrence indicating the particular status, and their dates of start and end. In particular the question reads 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?* [10]

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

In order to simplify the observation scheme we set the beginning of the window of observation at February 28th and only consider subjects who started from a state of unemployment in both the register and the survey. This could be considered the most sensitive approach to follow for researchers who only had access to survey data. That is, in order to reduce the impact of ME, and making use of what is known regarding the sample design, it could be expected that subjects who appeared to have misclassified their work status on 28th February are discarded.

Under this restriction our sample shares the structure seen in state-based samples (Holt, McDonald and Skinner, 1991), where the sample frame is created out of individuals who are known to be in a particular state and mimics the scheme used in the examples in Section 3.1. Our final sample size captures 381 individuals (out of a total of 532 captured by both survey and register) and the window of observation encompasses spells from 28/2/92 to 30/03/93, where the ending date represents the earliest day interviews were taken. Right censoring is present in both datasets.

The explanatory variables in the models considered here are age, experience, and their

[9] PRESO is a register from the Swedish employment office (Arbetsmarknadsstyrelsen).
[10] This and the following quote are translations from the original in Swedish.

interaction term. Experience captures self-reported levels of experience in the type of work that the subject applied for on a scale with three levels (low, medium, and high). Both variables are drawn from the register; the value for age is taken in January 1993, while for experience the mean of the monthly reported levels in 1992 is used. Given that age is an important variable in the register the probability that it is prone to ME is very low. This is different for experience since it is a self-reported value. However, in our analysis we assume that both of them are free of ME. In our sample the mean age is 37 and the standard deviation 8.8, while for experience these are 2.59 and .60 respectively. Finally, regarding these two variables the ME can be considered non-differential since the Spearman correlation coefficients for the misclassification of person-day observations with age and experience were .01 and .03, respectively.

6. Results

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

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

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

$$BIAS = \hat{\beta}_s - \hat{\beta}_r \tag{14}$$

where *s* stands for survey and *r* for register. A second measure particularly useful for making comparisons between models and between explanatory variables that use different scales for their regression coefficients is the relative bias,

$$R.BIAS = \frac{\left|(\hat{\beta}_s - \hat{\beta}_r)100\right|}{\left|\hat{\beta}_r\right|} \tag{15}$$

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

$$RMSE(\hat{\beta}_s) = \sqrt{MSE(\hat{\beta}_s)} = \sqrt{E[\hat{\beta}_s - \hat{\beta}_r]} = \sqrt{Var(\hat{\beta}_s) + (BIAS)^2} \tag{16}$$

Finally, in order to facilitate comparisons between models in terms of the RMSE, we also use the relative root mean squared error,

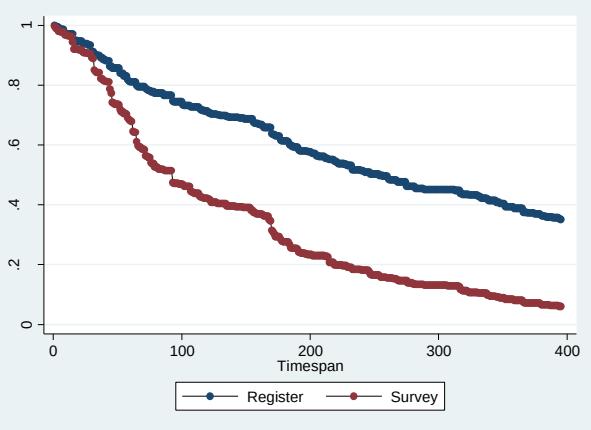
$$R.RMSE = \frac{\left(RMSE(\hat{\beta}_s) - RMSE(\hat{\beta}_r)\right)100}{RMSE(\hat{\beta}_r)} \tag{17}$$

6.1. Impact on EHA Models for Single Events

We start the analysis of the impact of ME in EHA by looking at the descriptive statistics of the first spells of unemployment. In this part of the analysis the sample used contains 381 subjects and spells, in both the register and the survey datasets. Figure 4 below shows the Kaplan-Meier estimate (defined in equation 2) of the survivor functions for the registered and reported time in unemployment. The two datasets show a similar path for the first 30 days; from that point until about day 100 the two measures diverge due to an accelerated failure rate in the survey; from then on the two survivor functions behave

roughly similarly and the gap between them is maintained. At the end of the window of observation 35% of the spells of unemployment in the register, 133 in total, were right-censored, whereas in the survey this was only 6% of the sample, 23 in total.

Figure 4. Survivor function for the register and survey data



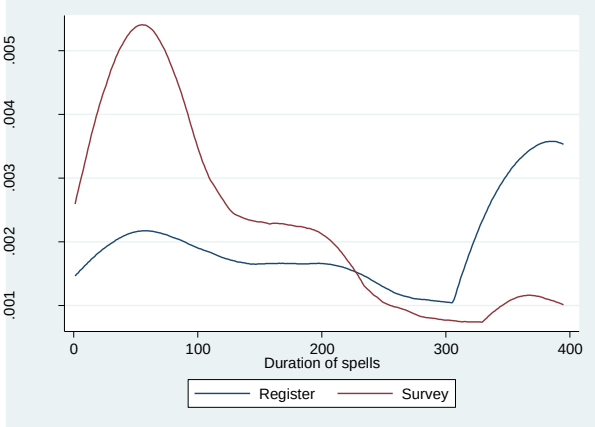
Measures of central tendency for the registered and reported durations also show substantial differences. These are included in Table 1 together with their standard deviations.

Table 1. Descriptive Statistics of the Unemployment Durations

	Mean	Median	Std. Dev.
Register	241	253	145
Survey	136	92	113

The mean duration in the register is 241, while the median is 253 days. In the survey these figures were 136 and 92, respectively. The higher median than the mean in the register indicates that the probability density function of durations is skewed to the left, whereas the opposite can be deduced for the distribution of durations from the survey. On the other hand, measures of dispersion are similar. These features can be seen graphically in Figure 5, where the probability density functions for spells of unemployment in the register and survey data are plotted.

Figure 5. Probability density function of single spells



Given the three types of ME affecting retrospectively reported work histories presented in Section 3.1 (miscounting, mismeasurement, and misclassification), and the setting considered here, where every subject starts from the same category (unemployed) and only first spells are contemplated, we conclude that the differences in the survivor functions are mainly due to mismeasurements. In particular, spell lengths of unemployment are under-reported in the survey. As we show in Figure 2, miscounting in the form of omitted spells would show longer spells of unemployment in the survey due to the artificial link of current to future spells of unemployment, yet, the exploratory analysis shows the opposite effect, a shortening of durations. Finally, the possibility that the survey data contain misclassified cases is ruled out since cases starting from a different status were discarded and multiple spells are not considered in this first part of the analysis.

The impact of using this prone to error data in EHA models is analysed next. A separate subsection is included for each family of EHA models, and at the end their relative performance in the presence of ME is assessed.

6.1.1. Impact on the Accelerated life-time Weibull and Exponential Models

In Tables 2 and 3 below we show the results obtained when comparing the AL Weibull models using register and survey data. In the model using the register data the main effects for both age and experience are negative and statistically significant, while their interaction effect is also significant but positive. So, considering the main effects of age

and experience, the older and more experienced the subjects the longer it takes to make a transition out of unemployment. However, this claim is nuanced by the positive interaction term, which indicates that subjects who are both old and highly experienced make that transition quicker.

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

Table 2. Single event AL Weibull model using register and survey data*

		Regression Coefficient	95% Confidence Interval		Standard Error
Register	Age	**-.087**	-.163	-.012	.039
	Experience	**-1.38**	-2.38	-.38	.51
	Age*Exp	**.038**	.010	.065	.014
	Constant	**9.08**	6.34	11.81	1.40
	α	.98	.88	1.10	.05
	LR Chi^2 (3)	**11.34**			
Survey	Age	.001	-.055	.058	.029
	Experience	-.09	-.835	.645	.377
	Age*Exp	.001	-.019	.022	.010
	Constant	**5.07**	3.06	7.07	1.02
	α	**1.11**	1.02	1.21	.05
	LR Chi^2 (3)	.98			

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

Table 3 shows the four measures set out at the beginning of this section to assess the impact of ME. The relative biases in the coefficients of the explanatory variables age and

experience are very large, 101.1% and 97.4%, respectively. These results indicate that the size of the bias is roughly the size of the true estimate. The interaction effect suffers from a similar effect, with a R.BIAS of 93.1%; it is only more moderate in the constant term, 44.2%.

Table 3. Bias in the single event AL Weibull model

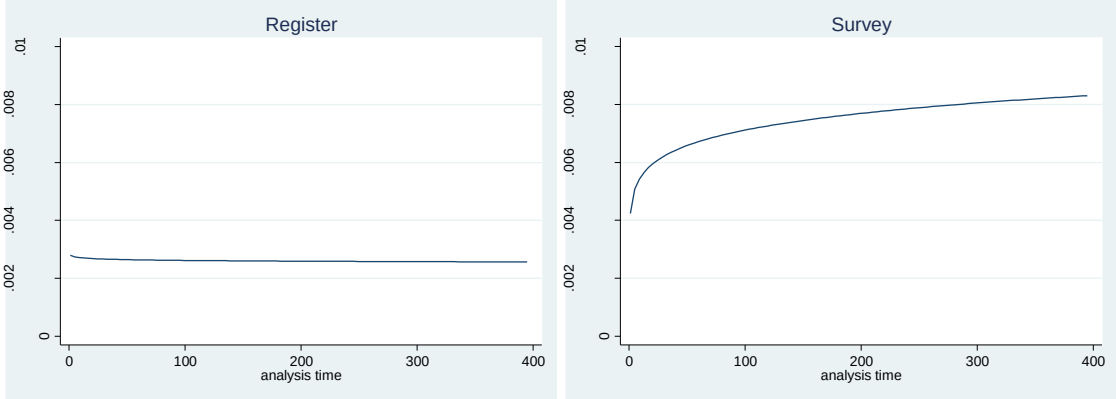
	BIAS	R.BIAS	RMSE	R.RMSE
Age	.088	101.1%	.093	137.6%
Experience	1.29	93.1%	1.34	161.9%
Age*Exp	-.037	97.4%	.038	173.8%
Constant	-4.01	44.2%	4.14	196.6%
α	.130	13.3%	.139	178.6%

The impact of ME on α, the parameter used in the Weibull model (see equations 3 and 4) to estimate the shape of the baseline hazard function, might seem relatively unimportant compared to what has been seen in the other coefficients since the true estimate is .98 and the one found using survey data is 1.11. However, as Skinner and Humphreys (1999) point out, in some settings, there is interest not only in the size of this estimate but also in the distinction between α<1, α=1, and α>1, or equivalently between a decreasing, constant or increasing hazard function respectively. For example, Chesher et al. (2002) anticipate that in the analysis of unemployment durations it is well known that uncontrolled across-individual heterogeneity in hazard functions can lead to the appearance of negative duration dependence. In our case we observe the opposite effect when the model is specified using survey data, while the model using register data show no effect in either direction. Here the impact of ME differs from what we have seen for the rest of coefficients, indicating a positive effect where there is none, which represents a type I error.

Figures 6 shows the shapes of the baseline hazard functions for the register and the survey data. In spite of the different signs of the slopes, it is worth noting that the shape of the baseline hazard function from the survey data mimics quite well the one from the register data. However, this result could be expected. Due to the constraints of the Weibull model, where only one shape parameter is used, baseline hazard functions are

bound to be either monotonically increasing or decreasing.

Figure 6. Weibull baseline hazard function for the register and survey data



Another characteristic to note from Figures 6 is the flatness of both hazard functions, which are almost constant across the window of observation. This feature suggests the possibility of using a simpler model to parameterize the baseline hazard function. In particular, the AL exponential (defined in equation 6) appears to be a good alternative because it assumes a constant baseline hazard function.

A likelihood ratio test between the two models using register data (taking the exponential model to be nested in the Weibull) corroborates this intuition. The test shows that the difference in deviances (.13) for 1 degree of freedom is not statistically significant (p-value=.72). The better specification of the exponential model can also be concluded from the lower SEs for age, experience and the constant term. Results are shown in Tables 4 and 5.

Table 4. Single event AL Exponential model using register and survey data

		Regression Coefficient	95% Confidence Interval		Standard Error
Register	Age	**-.087**	-.161	-.012	.038
	Experience	**-1.37**	-2.35	-.38	.50
	Age*Exp	**.037**	.010	.064	.014
	Constant	**9.04**	6.36	11.72	1.37
	LR Chi^2 (3)	**11.47**			
Survey	Age	-.003	-.063	.062	.032
	Experience	-.11	-.94	.71	.42
	Age*Exp	.002	-.021	.025	.012
	Constant	**5.08**	2.85	7.32	1.14
	LR Chi^2 (3)	.81			

In addition, the exponential model seems to perform marginally better at buffering the effects of ME; at least in terms of R.BIAS which is now lower for all the coefficients. It is possible that parametric EHA models are more sensitive to ME in the response when the baseline hazard function is misspecified.

Table 5. Bias in the single event AL Exponential model

	BIAS	R.BIAS	RMSE	R.RMSE
Age	.084	96.6%	.090	136.5%
Experience	1.25	91.7%	1.32	163.9%
Age*Exp	-.035	94.6%	.037	164.3%
Constant	-3.95	43.7%	4.12	200.9%

6.1.2. Impact on the Proportional Hazards Cox Model

We now turn to the PH Cox model (defined in equation 7). Estimates from the PH Cox model are often presented on a hazard rate scale, however, here we show the untransformed coefficients to facilitate comparisons between models.[11] Tables 6 and 7 below show the results of the PH Cox model using the register and survey data.

[11] That is we report $\hat{\beta}_i$ instead of $\exp(\hat{\beta}_i)$

Table 6. Single event PH Cox model using register and survey data*

		Regression Coefficient	95% Confidence Interval		Standard Error
Register	Age	**.086**	.011	.160	.038
	Experience	**1.35**	.36	2.33	.50
	Age*Exp	**-.037**	-.064	-.010	.014
	LR Chi^2 (3)	**11.03**			
Survey	Age	.002	-.061	.065	.032
	Experience	.13	-.69	.95	.42
	Age*Exp	-.002	-.025	.020	.012
	LR Chi^2 (3)	.77			

*Compared to the AL models signs of regression coefficients are now reversed since an increase in the hazard corresponds to a decrease in the expected (log-) duration, and vice-versa.

Results regarding the impact of ME on the PH Cox model show a very similar picture to what was found in the previous models. The regression coefficients are again heavily attenuated. Interestingly, the PH Cox model performs slightly better in terms of RMSE. This is surprising given the higher precision obtained from parametric models when they are correctly specified. This result suggests that, in spite of using an optimal parametric form for the true durations (from the register data), the baseline hazard function will probably change when there is ME, and less restrictive models such as the PH Cox are then a better choice.

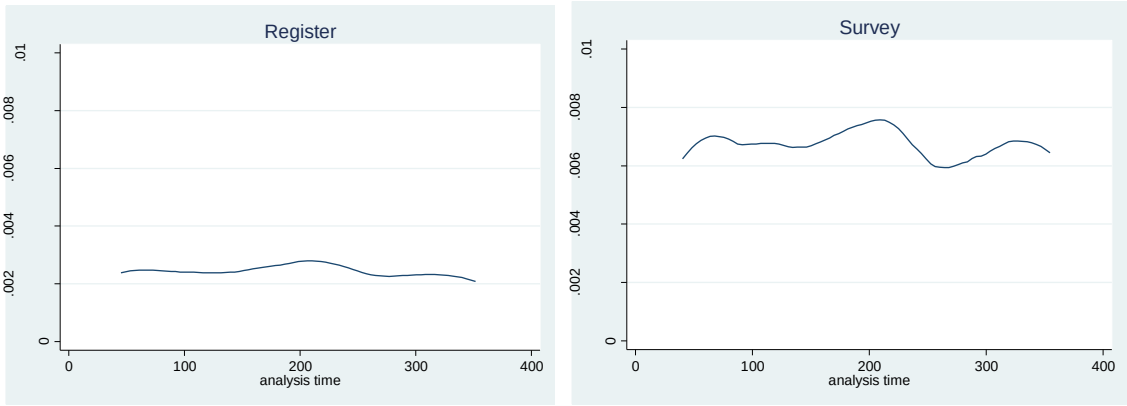
Table 7. Bias in the single event PH Cox model

	BIAS	R.BIAS	RMSE	R.RMSE
Age	-.084	97.7%	.090	136.5%
Experience	-1.22	90.5%	1.29	157.2%
Age*Exp	.035	94.6%	.037	164.3%

The Cox baseline functions for the survey and register data are displayed in Figure 7. These are calculated using STATA 11, where the baseline hazard ratio at each transition is estimated first, and then kernel density estimation is used to smooth them. From the comparison of the two functions it can be seen that the former is overestimated, just like it was in the Weibull model. Also, now that the baseline function is freely estimated, we can confirm that the baseline function from the register data is truly constant, which

corroborates the adequacy of the exponential model as opposed to the Weibull one. In addition, we can see that when using survey data there are a couple of bumps (at about day 220 and 330), which could lead to slightly misleading time-dependence inferences. These shocks were not captured by the exponential nor the Weibull baseline functions for survey data because of their parametric restriction. However, since the Cox baseline function using survey data remains roughly constant when the Weibull one shows a positive slope, we might say that the former reflects the true function more faithfully.

Figure 7. Cox baseline hazard function for the register and survey data



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

6.1.3. Impact on the Proportional Odds Logit Model

Finally we review the effect of retrospective ME in the response variable on a model from the non-parametric family, a PO logit model (defined in equation 8). Here, a series of temporal dummies are included in the model in order to specify the baseline logit-hazard function. Each of the dummy variables represents a period of the time-frame, in what is called a piecewise-constant hazards model. This is a reasonable solution when

coarse time units relative to the window of observation are used. However, for our case this raises some complications. First, the degrees of freedom are drastically reduced from the inclusion of 394 dummy variables, one for each day. Second, some of the days capture the same number of failures, which produces a problem of perfect multicollinearity in the model. In order to prevent these two problems we used temporal dummies that aggregated failures by weeks.

The results for the PO logit model are shown in Tables 8 and 9. The dummy variables representing the 56 weeks considered in the window of observation are not included in the tables for reasons of space, but they are shown in Figure 8 below as the dots composing the baseline hazard functions. In addition, the sample size is now 89,842 person-day cases in the register, and 50,366 in the survey[12].

Table 8. Single event PO logit model using register and survey data

		Regression Coefficient	95% Confidence Interval		Standard Error
Register	Age	**.086**	.012	.161	.038
	Experience	**1.36**	.38	2.35	.50
	Age*Exp	**-.037**	-.064	-.010	.014
	Constant	**-9.41**	-12.24	-6.57	1.45
	LR Chi^2 (61)	**77.29**			
Survey	Age	.002	-.061	.065	.032
	Experience	.14	-.69	.97	.42
	Age*Exp	-.003	-.025	.020	.012
	Constant	**-5.76**	-8.09	-3.42	1.19
	LR Chi^2 (61)	**161.33**			

The outcomes of the two models are again very similar to what was found in the previous EHA specifications. The expected lower precision due to the 56 additional parameters that needed to be estimated to reproduce the baseline hazard function was not as problematic as first thought. In fact, the same SEs as in the AL exponential and the PH Cox were obtained for age, experience, and the interaction effect when the register data is

[12] The two datasets differ in their sample size because of the transformations required in the specification of EHA models for discrete data; from a dataset capturing one case for each subject to another capturing person-week cases.

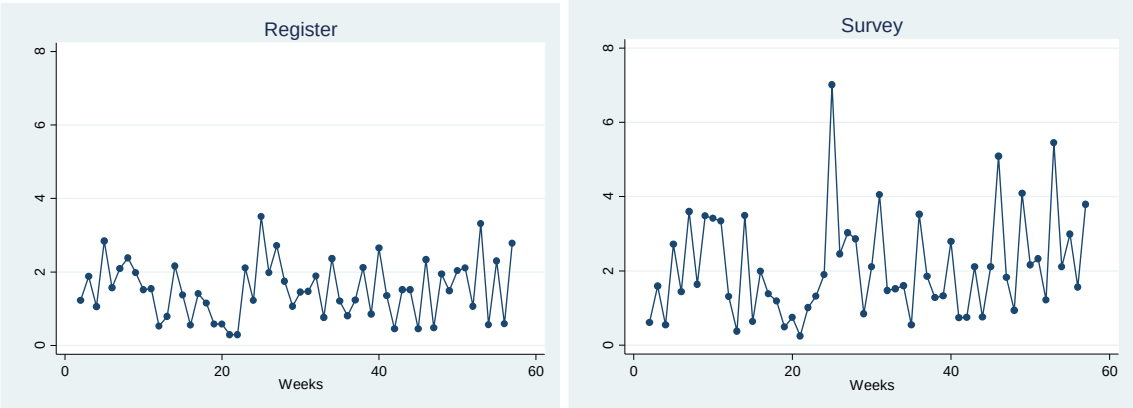
used.

Table 9. Bias in the single event PO logit model model

	BIAS	R.BIAS	RMSE	R.RMSE
Age	-.084	97.7%	.090	136.5%
Experience	-1.22	89.7%	1.29	157.0%
Age*Exp	.034	91.9%	.036	157.5%
Constant	3.65	38.8%	3.84	165.4%

Because of the effect of aggregating days into weeks, both baseline hazard functions differ from the PH Cox ones, in particular the function for the register data is not constant. Also, unlike in the previous cases where the effect of ME was expressed as a higher baseline function, here what we see is more volatility between time-periods (weeks), resulting in a more jagged baseline function.

Figure 8. PO baseline hazard function for the register and survey data*,**



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

6.1.4. Summary of the Impact on Models for Single Spells

In this section we have seen that the consequences of using retrospective data in EHA models for single spells are not negligible and are very similar across different models. Strong attenuation effects were found in all the regression coefficients. In Table 10 we summarise these results by taking the average R.BIAS and R.RMSE over the coefficients

of age, experience and their interaction term[13], for each of the four models studied.

Table 10. EHA models' performance in the presence of retrospective ME

	R.BIAS	R.RMSE
AL Weibull	97.2%	157.8%
AL exponential	94.3%	154.9%
PH Cox	94.3%	152.7%
PO logit	93.1%	150.4%

In addition to the strong attenuation effects, illustrated by measures of R.BIAS not lower than 93%, the similarity of the effects across models is striking. None of the models seems to buffer the effects of ME better than the others. In fact, the between model variability in terms of R.BIAS and R.RMSE is 1.8% and 3.1% respectively, while the average effect within models is 94.7% for the former and 153.9% for the latter.

The analysis presented so far has focused on assessing the impact of ME in each model separately. However, it could be argued that some EHA models are superior to others for the type of data that we are using here. For example, non-parametric models like the PO logit are recommended when there are fewer time units, whereas here we have seen that the exponential model is a better specification than the Weibull model. In order to assess which model performs better in the presence of ME we need to compare them against a common benchmark. That is, the use of a common benchmark allows us to analyse not only comparisons between the same models using error free and prone to error data, but also comparisons between different models when prone to error data is used.

Here, we use results from the PH Cox model based on register data as that benchmark. There are both empirical and theoretical reasons for this choice. First, the PH Cox model has, along with the AL exponential and the PO logit, the lowest SEs in their regression coefficients (see Tables 4, 6 and 8). Second, since the baseline hazard function is freely estimated it cannot be misspecified. Finally, tied events, a flaw affecting models for continuous time such as the Cox model, are not a major issue here. The window of

[13] In order to make comparisons possible we excluded the constant term from this analysis since the PH Cox model does not estimate it.

observation covers 395 days, which makes the time-unit approximately continuous, and rarely do two spells or more end on the same day.

This process to assess the relative impact of ME on the different EHA models implies the assumption that the Cox model using register data produces the true estimates. The comparisons can be formally defined by equations 15 and 17, where $\hat{\beta}_r$ is now substituted by $\hat{\beta}_{r,Cox}$.

Table 11. EHA models' performance compared to the PH Cox

	R.BIAS	R.RMSE
AL Weibull	97.1%	161.6%
AL exponential	94.2%	154.7%
PO logit	93.0%	150.5%

Results are shown in Table 11 above, where it can be seen that the PO logit performs marginally better than the rest. It is also interesting to note that the AL Weibull offers the worst performance. These results reinforce the idea put forward when discussing the effect of ME in the baseline function, EHA models that do not make use of a restrictive parametric form seem to do better at buffering the effect of ME in the response variable. This seems to be especially true when the parametric form used is not the most appropriate, as it is shown by the worse performance by the Weibull than the exponential model.

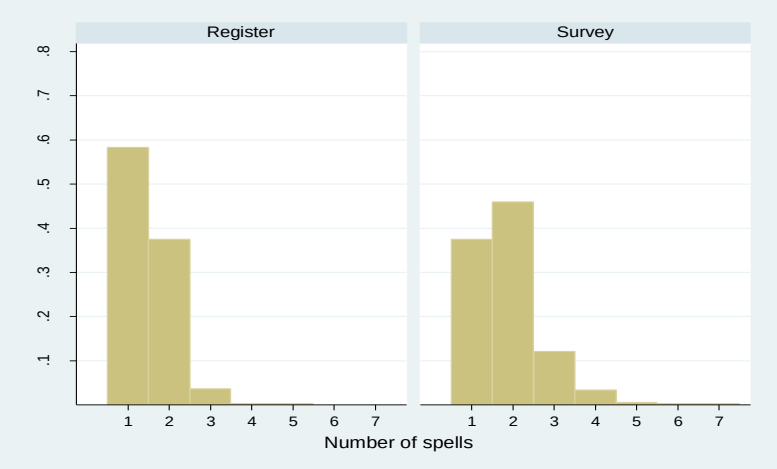
In the first part of the analysis we were interested in assessing the relative performance of different EHA models in the presence of retrospective ME, when only first spells are considered. We proceed with the analysis by looking at the implications derived from using retrospective data for the same EHA models as before but when multiple spells are considered.

6.2. Impact on EHA Models for Repeated Events

The consideration of repeated spells makes both the specification of the durations of unemployment, and the types of ME affecting them, more complex. As was seen in

Section 3.1 models for repeated spells can also include both FP and FN cases. This remains valid regardless of the restriction that we used to study only cases which are known to start from unemployment in both the register and the survey. Similarly, problems of miscounting are more complex than before. In fact, omitted and/or over-reported spells make the sample size in the survey differ from the one in the register. In the analysis presented in the previous section both the register and survey data set contained a sample size of 381 spells. Now, from the inclusion of repeated events and as a consequence of miscounting, there are 559 spells of unemployment in the register, and 706 in the survey. Figure 9 below shows a histogram for the number of registered and reported spells.

Figure 9. Histogram of the number of spells in the register and the survey

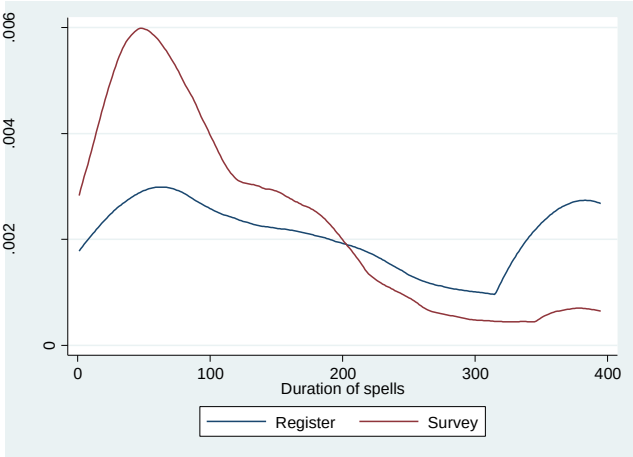


The mean duration of spells in the register is 204 days with a standard deviation of 140, while in the survey the mean duration was 116 days, and the standard deviation 97. These mean durations are lower than the ones in Section 6.1, which implies that the additional spells included in this setting are shorter than the first ones.

Figure 10 displays the probability density functions for the spells of unemployment in both the survey and the register. The density function for the survey peaks for spells lasting around 50 days, whereas the register function has two peaks, one at day 50 and another where spells are right-censored. Similar shapes were found in Figure 5, where the probability density functions for the case of single spells was shown. However, it seems that now the register and survey functions have converged, which might anticipate a

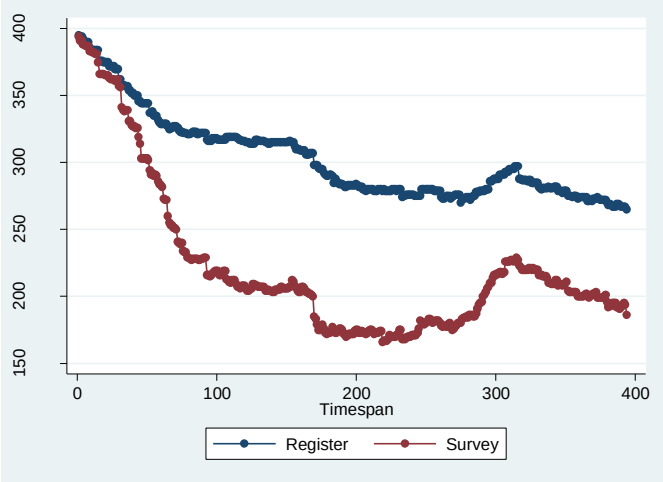
lower effect of ME. On the other hand, an analysis based solely on shapes of the probability density functions can be misleading, since these functions compare proportions, that is, they do not take into account the higher number of spells recorded in the survey, which might also be having an effect.

Figure 10. Probability density function of repeated spells



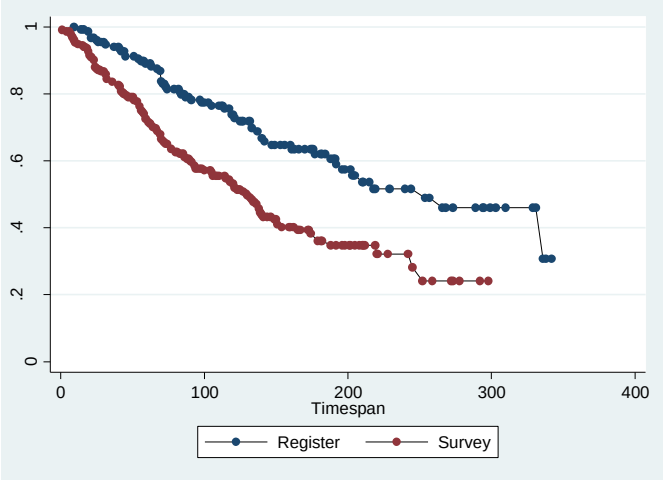
In order to make a better graphical assessment of the impact of ME in EHA models for repeated events we use a scatter plot capturing the number of subjects unemployed at each period of the window of observation (Figure 11). The first part of the graph shows a similar picture to that of the survivor functions from Figure 4, the rates of unemployment for both the register and the survey fall in a similar pattern for the first 30 days, subsequently the survey shows a sharper decay, widening the difference between the survey and register rate more than in the single spells' setting. In addition, after approximately day 90, when some of the first spells have failed and repeated ones are entering the study, the two functions remain relatively stable and they even show both growth and certain convergence.

Figure 11. Number of subjects unemployed across the window of observation



The mild convergence between the two functions at the second half of the timespan can only be due to the higher number of cases re-entering unemployment in the survey, since as we saw before the mean length of spells is shorter when considering repeated events, indicating that the first spells are longer. In Figure 12 we plot survivor functions for the second spells found in the register and the survey; here we observe an accelerated failure rate in the survey, similar to what was found in Figure 4 for the case of first spells. The survivor function for the register is based on 159 spells with an average length of 131 days, while the one from the survey had 238 spells with an average length of 101 days. Unlike the first spells, these start at different times of the window of observation, and because of that censored cases are seen at different times for the survey and the register.

Figure 12. Survivor functions restricted to second spells



In what follows we assess the impact of using the survey dataset on different EHA models when repeated events are considered. In all of the following models we assumed that the spells of unemployment were of the same type and unordered. That is, spells of unemployment were not treated differently because of their location in the window of observation or their position with respect to their order of appearance. To account for within subject dependencies we use robust SEs. Specifically we use the sandwich estimator, which is the default option used in STATA version 11. This choice of model for repeated events replicates what has been used before in the literature on labour market studies (Gash, 2008; Pyy Martikainen and Rendtel, 2009). Finally, the impact of ME in the correlations of spells within subjects is briefly explored at the end by comparing RI PO logit models. As before, we start with the parametric family of EHA models.

6.2.1. Impact on the Accelerated life-time Weibull and Exponential Models

The Weibull model for the register data is now slightly different than it is for the single spells case; all the regression coefficients remain significant, while age, experience, their interaction, and all the SEs are now smaller (see Table 12 below). The lower effect size of the explanatory variables can be interpreted as the length of second and subsequent spells being less associated with age, experience or their interaction effect, while the smaller SEs represent the improvement in precision after an increase in the sample size.

Table 12. Repeated events AL Weibull model using register and survey data

		Regression Coefficient	95% Confidence Interval		Standard Error
Register	Age	**-.084**	-.150	-.017	.034
	Experience	**-1.22**	-2.10	-.34	.45
	Age*Exp	**.033**	.009	.058	.012
	Constant	**8.91**	6.51	11.3	1.22
	α	**1.01**	.92	1.10	.04
	LR Chi^2 (3)	**8.43**			
Survey	Age	-.005	-.054	.043	.025
	Experience	-.15	-.77	.47	.32
	Age*Exp	.003	-.015	.020	.009
	Constant	**5.36**	3.66	7.05	.86
	α	**1.10**	1.01	1.19	.04
	LR Chi^2 (3)	.78			

Table 13 summarizes the impact of ME in this model. Similar to the case of single spells all the coefficients are attenuated, and they are not significant. The impact in terms of R.BIAS is slightly smaller than for the single spells model for all coefficients, whereas in terms of R.RMSE the effect is lower for age and the constant and bigger for experience and the interaction term The similarity of the results when compared to the single spells model is rather unexpected, since in Section 3.1 we anticipated that the forms of ME could be more complex in a multiple spells design.

Table 13. Bias in the repeated events AL Weibull model

	BIAS	R.BIAS	RMSE	R.RMSE
Age	.08	94.0%	.083	143.7%
Experience	1.07	87.9%	1.12	149.5%
Age*Exp	-.03	90.9%	.031	161.0%
Constant	-3.56	39.9%	3.66	199.3%
α	-.090	8.2%	.088	119.1%

Finally, unlike we saw in the previous section, the shape parameter is now significant and positive when survey data is used. However, its coefficient is still very close to 1, which let us think again of the appropriateness of using an exponential specification. Tables 14

and 15 show the results for the AL exponential model[14].

Table 14. Repeated events AL Exponential model using register and survey data

		Regression Coefficient	95% Confidence Interval		Standard Error
Register	Age	**-.084**	-.151	-.017	.034
	Experience	**-1.22**	-2.11	-.34	.45
	Age*Exp	**.034**	.009	.058	.012
	Constant	**8.93**	6.53	11.33	1.23
	LR Chi^2 (3)	**8.40**			
Survey	Age	-.006	-.056	.043	.025
	Experience	-.16	-.80	.48	.33
	Age*Exp	.003	-.015	.021	.009
	Constant	**5.39**	3.64	7.13	.89
	LR Chi^2 (3)	.71			

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

Table 15. Bias in the repeated events AL Exponential model

	BIAS	R.BIAS	RMSE	R.RMSE
Age	.078	92.9%	.082	140.9%
Experience	1.07	87.0%	1.11	147.7%
Age*Exp	-.031	91.2%	.032	169.0%
Constant	-3.54	39.7%	3.65	198.0%

6.2.2. Impact on the Proportional Hazards Cox Model

Results for the PH Cox model are included in Tables 16 and 17. The estimates for both

[14] The use of the exponential model is again justified by a likelihood ratio test (p-value=.86) used to compare it against the Weibull specification.

the model using register and survey data are again very similar to what has been seen in this section for the case of the AL Weibull and exponential.

Table 16. Repeated events PH Cox model using register and survey data

		Regression Coefficient	95% Confidence Interval		Standard Error
Register	Age	**.083**	.017	.149	.034
	Experience	**1.21**	.34	2.08	.45
	Age*Exp	**-.033**	-.057	-.009	.012
	LR Chi^2 (3)	**8.32**			
Survey	Age	.007	-.044	.059	.026
	Experience	.17	-.48	.84	.34
	Age*Exp	-.003	-.022	.015	.009
	LR Chi^2 (3)	.76			

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

Table 17. Bias in the repeated events PH Cox model

	BIAS	R.BIAS	RMSE	R.RMSE
Age	-.076	91.6%	.080	136.2%
Experience	-1.03	85.5%	1.09	144.1%
Age*Exp	.030	90.9%	.031	161.0%

6.2.3. Impact on the Proportional Odds Logit Model

In the estimation of the PO logit for the repeated events 110,563 and 82,599 person-day cases were used in the register and survey models respectively. Results for the PO logit are presented in Tables 18 and 19. Regression coefficients are similar to the ones found in the rest of the models from this section. The same applies to the SEs which, just as for the single spells, do not fall in spite of having to estimate the baseline function week by week.

Table 18. Repeated events PO logit model using register and survey data

		Regression Coefficient	95% Confidence Interval		Standard Error
Register	Age	**.083**	.016	.150	.034
	Experience	**1.21**	.32	2.09	.45
	Age*Exp	**-.033**	-.057	-.009	.012
	Constant	**-9.26**	-11.84	-6.68	1.32
	LR Chi^2 (61)				
Survey	Age	.005	-.047	.056	.026
	Experience	.15	-.50	.81	.33
	Age*Exp	-.002	-.021	.017	.009
	Constant	**-6.13**	-8.07	-4.19	.99
	LR Chi^2 (61)	**207.7**			

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

Table 19. Bias in the repeated events PO logit model

	BIAS	R.BIAS	RMSE	R.RMSE
Age	-.078	94.0%	.085	141.8%
Experience	-1.05	87.4%	1.10	145.1%
Age*Exp	.031	93.9%	.032	169.0%
Constant	3.13	33.8%	3.28	149.3%

6.2.4. Impact on the Random Intercepts Proportional Odds Logit Model

Here, the effect of retrospective ME in a random effects model is explored. The specific model of study is a RI PO logit estimated using Markov chain Monte Carlo. This last model is estimated using MLwiN[15], unlike the previous models where we used STATA. In order to assure that convergence was achieved and to reduce simulation error we used

[15] Although it was called from the STATA interface using the command runmlwin developed by Leckie and Charlton (2011).

5,000 iterations after burning-in another set of 5,000 iterations. Results are presented in Table 20 and 21 below.

Table 20. RI PO logit model using register and survey data

		Regression Coefficient	95% Credible Interval		Standard Deviation
Register	Age	**.068**	.036	.100	.018
	Experience	**1.05**	.62	1.43	.22
	Age*Exp	**-.029**	-.039	-.017	.006
	Constant	**-8.81**	-10.12	-7.41	.65
	Var(RI)	**.083**	.002	.280	.077
	Bayesian DIC	4281.7			
Survey	Age	.010	-.010	.036	.013
	Experience	.23	-.049	.552	.161
	Age*Exp	-.004	-.014	.003	.004
	Constant	**-6.40**	-7.45	-5.48	.518
	Var(RI)	**.017**	.001	.093	.023
	Bayesian DIC	6308.7			

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

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

Table 21. Bias in the RI PO logit model

	BIAS	R.BIAS	RMSE	R.RMSE
Age	-.058	85.3%	.059	230.2%
Experience	-.826	78.3%	.842	282.5%
Age*Exp	.025	86.2%	.025	322.0%
Constant	2.41	27.4%	2.46	278.1%
Var(RI)	-.066	79.5%	.070	-9.2%

The variance of the RI term is significant in both models but it is much smaller in the one using survey data. Precisely, in terms of R.BIAS, there is an attenuation of 79.5%. This result indicates that the between subjects unobserved heterogeneity decreases in the presence of retrospective ME. Similarly, it could be argued that the impact of ME makes work histories less distinguishable, or that the retrospective ME is not concentrated in the reports of a few subjects, but affects most of them.

6.2.5. Summary of the Impact on Models for Repeated Spells

As was the case for models using single spells, the coefficients for age, experience and their interaction effect were significant in all the models using register data and became non-significant when the survey data is used. We also observe very similar attenuation effects to what we saw in models for single spells. However, contrary to our hypothesis expecting more complex forms of MEs as a result of considering repeated events (Section 3.1), we have found that the average R.BIAS is lower than before, in each of the models analysed.

Similarly, the impact of ME in terms of R.RMSE for models considering repeated events is not higher than for the single spells case. The only exception in this respect is the RI PO logit model, which shows almost twice the impact than any other model. This is shown in Table 22, where the average results for age, experience and their interaction term are compared. Comparing models for repeated events, it is interesting to note that the PH Cox comes the first and second in terms of R.RMSE and R.BIAS, respectively. The lowest impact in terms of R.BIAS comes from the RI PO logit. However, we have seen that this might be due to the smaller effects found when register data is used, which leaves less room for attenuation towards the null.

Table 22. EHA models' performance in the presence of retrospective ME

	R.BIAS	R.RMSE
AL Weibull	90.9%	151.4%
AL exponential	90.4%	152.5%
PH Cox	89.3%	147.1%
PO logit	91.8%	160.5%
RI PO logit	83.3%	278.2%

In order to properly assess not only the impact of ME in each of the models, but which one performs better in the presence of ME, we need to make comparisons against a common benchmark. For that we use once more the PH Cox model based on the register data. This is the model less prone to problems of misspecification, and as we just saw the model that suffers the least when both R.BIAS and R.RMSE are contemplated. We present results from these comparisons in Table 23.

Table 23 EHA models' performance compared to the PH Cox

	R.BIAS	R.RMSE
AL Weibull	90.88%	151.77%
AL exponential	90.18%	153.28%
PO logit	91.78%	161.48%
RI PO logit	85.63%	91.50%

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

7. Conclusion

In this paper we have explored the implications of using EHA models where the response variable is affected by ME derived from a retrospective question. Evidence of large attenuation biases in the regression coefficients is found across different EHA models. These findings go against the common belief that ME in the response variable only affects the SEs of the model's estimates, and also contrast with what has been seen so far in the literature.

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

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

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

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

Using the Cox model as a benchmark we found that the exponential model is less affected by ME than the Weibull both in terms of R.BIAS and R.RMSE and for both the case of single and multiple spells. It seems that parametric forms that are correctly specified for the true data might buffer the effect of ME better than when they are misspecified. Moreover, semi and non-parametric models perform similarly to the exponential model, even in terms of R.RMSE. This is an interesting result since parametric models, when correctly specified, are expected to obtain more precise estimates. Moreover, ascertaining the shape of the baseline hazard function is complicated, and in general, it could be expected that parametric forms will perform worse than we have seen here. Hence, when the shape of the baseline hazard function cannot be identified, as is the case in settings

that use durations measured with errors, the use of semi- and non-parametric forms are recommended.

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

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

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

In this study we have used data derived from a retrospective question on work histories for a period of 395 days, yet our findings could be generalized to other cases where retrospective data is used to derive different lifecourse events. In particular, this would be the case for events that, because of their relatively low saliency, can be subject to recall

errors in the form of mismeasuring, miscounting, and misclassification of spells in the same way as spells of unemployment are.

However, more research is necessary since some of the findings presented here need to be tested. Augustin (1999) pointed at this area as an underesearched one in need of more contributions, *"In contrast to its practical importance, ME has not yet attracted much attention in duration analysis."* (Augustin, p. 2, 1999); and this is something that Pyy-Martikainen and Rendtel (2009) have recently confirmed *"Despite the recognition of the existence of measurement errors in survey-based data on event histories, little is known about their effects on an event history analysis."* Pyy-Martikainen and Rendtel (2009, p.140). In particular we would like to extend our study to cases where the ME affecting the response variable in EHA is associated with the explanatory variables. Non-differential ME can generate biases that are not necessarily towards the null, but it is not clear what are the levels of association that could cause a change in the direction of the bias. Another setting of interest would be the extensions of the models seen here to the case of competing risks. This would allow contemplating the impact of retrospective data in EHA in greater detail, in particular the influence of misclassified cases.

References

Augustin, T. (1999), Correcting for Measurement Error in Parametric Duration Models by Quasi-likelihood. *Munchen Institut fur Statistik,* Working Paper, available at: http://epub.ub.uni-muenchen.de/1546/1/paper_157.pdf [accessed 24 January 2011].

Box-Steffensmeier J., and Jones, B. (2004). *Event Hystory Modeling: A Guide for Social Scientists*. Cambridge: Cambridge University Press.

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.

Carroll, R. et al. (2006), *Measurement Error in Nonlinear Models; a Modern Perspective*. Boca Raton: Chapman and Hall.

Chesher, A., Dumangane, M., and Smith, R. (2002), Duration Response Measurement Error, *Journal of Econometrics*, Vol.(111), pp. 169-194.

Cox, D. R. (1972), Regression Models and Life-Tables, *Journal of the Royal Statistical Society. Series B,* Vol.34, No. 2, pp.187-220.

Cox, D. R. (1975), Partial Likelihood, *Biometrika*, Vol. 62, No.2, pp. 269-276.

Dumangane, M. (2006), Measurement Error Bias Reduction in Unemployment Durations, *Centre for Microdata Methods and Practice*, Working Paper 3, available at: http://www.cemmap.ac.uk/wps/cwp0603.pdf [accessed 9 September 2011].

Fuller, W. (1987), *Measurement Error Models,* New York: John Wiley and Sons.

Gash, V. (2008), Bridge or Trap? Temporary Workers' Transitions to Unemployment and to the Standard Employment Contract, *European Sociological Review*, Vol.24, N0.5, pp.651-668.

Holt, D., McDonald, J.W. and Skinner, C.J. (1991). The Effect of Measurement Error on Event History Analysis. In: P. Biemer, ed. *Measurement Error in Surveys*. New York: John Wiley. Ch. 32.

Hughes, M. (1993), Regression Dilution in the Proportional Hazards Model, *Biometrics*, Vol.40, pp. 1056-1066.

Jäckle, A. (2008), Measurement Error and Data Collection Methods: Effects on Estimates from Event History Data, *Institute for Social and Economic Research (ISER),* Working Paper 2008-13, available at: https://www.iser.essex.ac.uk/publications/working-papers/iser/2008-13 [accessed 15 May 2013].

Kelly, P. (2012), A Review of Software Packages for Analyzing Correlated Survival Data, *The American Statistician*, Vol. 58, No. 4, pp. 337-342.

Korn, E., Dodd, L., and Freidlin, B. (2010), Measurement Error in the Timing of Events: Effect on Survival Analyses in Randomized Clinical Trials, *Clinical Trials*, Vol.7, No.6, pp.626-633.

Lawless, J. (2003), *Statistical Models and Methods for Lifetime Data*, Hoboken: John Wiley and Sons.

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

Leckie, G. and Charlton, C. (2011), Runmlwin: Stata Module for Fitting Multilevel Models in the MLwiN Software Package. Centre for Multilevel Modelling, University of Bristol.

Magder, L. and Hughes, J. (1997), Logistic Regression When the Outcome is Measured with Uncertainty, American Journal of Epidemiology, Vol.146, No.2, pp.195-203.

Meier, A., Richardson, B., and Hughes, J. (2003), Discrete Proportional Hazards Models for Mismeasured Outcomes, B*iometrics*, Vol. 59, No. 4, pp. 947-954.

Nakamura, T. (1992), Proportional Hazards Model with Covariate Subject to Measurement Error, *Biometrics*, Vol.48, No.3, pp.829-838.

Neuhaus, J. (1999), Bias and Efficiency Loss Due to Misclassified Responses in Binary Regression, *Biometrika Trust*, Vol.86, No.4, pp.843-855.

Novick, M.R. (1966), The Axioms and Principal Results of Classical Test Theory, *Journal of Mathematical Psychology*, Vol. 3, N. 1, pp. 1-18.

Peters, E. (1988), Retrospective Versus Panel Data in Analyzing Lifecycle Events, *The Journal of Human Resources*, Vol.23, No.4, pp.488-513.

Pina-Sánchez, J., Koskinen, J., and Plewis, I. (2012), Measurement Error in Retrospective Reports of Unemployment, *CCSR*, Working Paper, available at: http://www.ccsr.ac.uk/publications/working/ [accessed 18 June 2012].

Prentice, R. (1982), Covariate Measurement Errors and Parameter Estimation in a Failure-time Regression Model, *Biometrika*, Vol.69, No.2, pp. 331-42.

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.

Skinner, C. and Humphreys, K. (1999), Weibull Regression for Lifetimes Measured with Error, *Lifetime Data Analysis*, Vol.5, pp.23-37.

Skinner, C. (2000), Dealing with Measurement Error in Panel Analysis. In: D. Rose, ed. *Researching Social and Economic Change*. New York: Routledge. Ch. 6.

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.

Steele, F. (2005), Event History Analysis, *National Centre for Research Methods*, Briefing Paper, available at: http://eprints.ncrm.ac.uk/88/1/MethodsReviewPaperNCRM-004.pdf [accessed 26 February 2013].

Steele, F. (2008), Multilevel Models for Longitudinal Data, *Journal of the Royal Statistical Society: A*, Vol.171, No. 1, pp. 5-19.

Williams, R. L. (2000), A Note on Robust Variance Estimation for Cluster-Correlated Data. *Biometrics*, Vol.56, pp. 645–646.