DE GRUYTER OPEN

Journal of Official Statistics, Vol. 35, No. 1, 2019, pp. 203–229, http://dx.doi.org/10.2478/JOS-2019-0010

Adjusting for Measurement Error in Retrospectively Reported Work Histories: An Analysis Using Swedish Register Data

Jose Pina-Sánchez[1], *Johan Koskinen*[2], *and Ian Plewis*[2]

We use work histories retrospectively reported and matched to register data from the Swedish unemployment office to assess: 1) the prevalence of measurement error in reported spells of unemployment; 2) the impact of using such spells as the response variable of an exponential model; and 3) strategies for the adjustment of the measurement error. Due to the omission or misclassification of spells in work histories we cannot carry out typical adjustments for memory failures based on multiplicative models. Instead we suggest an adjustment method based on a mixture Bayesian model capable of differentiating between misdated spells and those for which the observed and true durations are unrelated. This adjustment is applied in two manners, one assuming access to a validation subsample and another relying on a strong prior for the mixture mechanism. Both solutions demonstrate a substantial reduction in the vast biases observed in the regression coefficients of the exponential model when survey data is used.

Key words: Bayesian statistics; measurement error; mixture model; retrospective data; unemployment.

1. Introduction

Many different forms of measurement error (ME) have been treated in the literature (Berkson 1950; Black et al. 2000; Neuhaus 1999; Novick 1966). Moreover, in some instances, different forms of ME interact within the same data-recording strategy. We consider here such a case, motivated by a retrospective data set, where ME is best modelled using a combination of errors. We demonstrate the application of a mixture of ME using validation samples of 20% and 40%, as well as in the absence of validation samples.

Retrospective questions are useful for collecting information about life-course events over a span of time from a single contact with a respondent. They are cheaper to administer than alternative data collection schemes that rely on repeated updates of the current state (Solga 2001). On the other hand, retrospective questions are prone to ME due to memory failures. These memory failures can take different forms.

For questions retrieving specific events, for example, "*When was the last time you went to the dentist?*" (Office for National Statistics 2006) – we can detect *telescoping effects*

[1] School of Law, University of Leeds, Leeds, LS2 9JT, United Kingdom. Email: j.pinasanchez@leeds.ac.uk
[2] Social Statistics, University of Manchester, Manchester, M13 9PL, United Kingdom. Emails: Johan.Koskinen@manchester.ac.uk and Ian.Plewis@manchester.ac.uk
Acknowledgments: We want to thank Sten-Åke Stenberg for granting us access to the "Longitudinal Study of the Unemployed", a unique data set without which this research project would not have been possible.

(Golub et al. 2000; Johnson and Schultz 2005). Neter and Waksberg (1964) coined this term to refer to the temporal displacement of an event, whereby people perceive recent events as being more remote than they are (backward telescoping or time expansion) and distant events as being more recent than they are (forward telescoping or time compression). Other researchers (Bradburn et al. 1994; Huttenlocher et al. 1988; Rubin and Baddeley 1989) have argued that, rather than distorted time perceptions, recall errors take the form of random ME around the reported date with the size of the error being proportional to the distance between the time of the interview and the actual date. That is, the further away the date of the event to be reported, the harder it is to recall and therefore the bigger the ME.

For questions retrieving count data – that is, those enquiring about the number of times an event has been experienced over a period of time, for example, *"How many times during the last two years have you put together self-assembly furniture at home?"* (Office for National Statistics 2008). – *interference effects* have been detected. This term was coined by Crowder (1976) and refers to the probability of recalling a particular event being inversely related to the number of times the respondent experiences similar events (Shiffrin and Cook 1978).

All these types of memory failure result in a significant loss of reliability and validity in most of the variables collected retrospectively, which in turn can have severe consequences in the form of loss of statistical power and biased estimates when used in statistical models. Nonetheless, although undoubtedly inconvenient, these effects can be mitigated through the implementation of methods for the adjustment of ME. In particular, much of the literature has focused on the implementation of adjustments where multiplicative models (see Section 2) are used to specify the distribution of recall errors (Augustin 1999; Dumangane 2007; Glewwe 2007; Holt et al. 2011; Pickles et al. 1996, 1998; Pina-Sánchez 2016; Skinner and Humphreys 1999).

Multiplicative models can be successfully used to map different ME processes stemming from memory failures in the reports of specific events. However, things become much more complicated when dealing with retrospective reports of different kinds of histories. Such reports require the identification and timing of the same or different events in chronological order. Take this question from the 2003 Improving Survey Measurement of Income and Employment project (Jenkins and Lynn 2005) as an example, *"Have you received Job Seeker's Allowance at any time since <DATE OF PREVIOUS INTERVIEW>?"* If yes, interviewees were then asked *"For which months since <MONTH OF INTERVIEW> have you received Job Seeker's Allowance?"*. Here, we might expect spells of receiving the benefit having misdated start or end times. However, we should also expect ME in the form omitting spells that did occur, over-reporting spells that did not occur and misclassifying spells by, for example, reporting receipt of one kind of benefit when it was, in fact, a different benefit. These other types of errors can generate severe distortions since, unlike misdated starts or ends of spells, they give rise to durations of spells that are unrelated to their true values. Under such circumstances, the impact of ME on the estimates of, say, a regression model, should be expected to be more serious than when using spells solely subject to multiplicative errors, while the potential application of adjustment methods is more limited and more complex to implement.

In this article, we use administrative data from the Swedish register of unemployment linked to survey reports of individuals' work histories (described in Section 3). Under the assumption that the former is perfectly measured, we assess: a) the prevalence and forms of ME in the retrospective reports of work histories (Subsection 4.1), b) the consequences of using durations of spells of unemployment captured from those work histories as the response variable in an exponential model (Subsection 4.2), and c) the effectiveness of different approaches based on Bayesian methods to adjust for the consequences of ME in such an exponential model (Section 5). In particular, we demonstrate the potential of an adjustment using a Bayesian mixture model. We calibrate the ME model separately using validation samples and by relying on past information. Section 6 concludes with a list of recommendations regarding the collection of work histories retrospectively, and with a discussion of possible further improvements in the adjustment model, using auxiliary data.

2. Modelling Recall Errors

In order to adjust for the consequences of using variables prone to ME, many analysts rely on the assumption that the error mechanism is classical. The classical ME model was first formally defined by Novick (1966) as $X^* = X + V$; where X^* is the observed variable, equal to the true variable, X, plus the ME term, V, with the following five assumptions:

Table 1. Assumptions of the classical model.

$E(V) = 0$;	null expectancy
$Var(V_i) = Var(V)$;	homoscedasticity
$V \sim N\,(0,\ Var(V))$;	normally distributed
$Cov(X,\ V) = 0$;	indep. error and true value
$E(Y\,\vert X,\ X^*) = E(Y\,\vert X)$;	non-differentiality

The classical model nicely reflects the type of ME that we can expect to find in continuous variables prone to random errors, for example when measuring temperature using an unreliable thermometer. In addition, the classical model is often used as the foundation upon which more complex ME processes are specified. One such a case is the previously mentioned multiplicative model, where the additive relationship between the true value and the error is substituted by a multiplicative one, so, $X^* = XV$.

The same assumptions about the error term described above apply, except that the error term is now log-normally distributed with a mean of one. The ME has a symmetric effect across the true values and maintains the scale used in duration and count data, from 0 to ∞. More importantly, since the effect of V on X^* is proportional to the value of X, the multiplicative model can be used to reflect a random ME process stemming from memory failures observed in retrospective questions capturing events or counts of events (see the examples from the Office for National Statistics in the introduction). That is, the greater the number of events experienced, or the further the event from the time of the interview, the higher the prevalence of ME. Note, as well, that this same model can also be used to account for systematic (i.e., nonrandom) recall errors, such as those found under the presence of backward and forward telescoping effects, by shifting the distribution of V to the right or left so its mean goes below or above one.

However, this model cannot adequately account for the type of ME observed in questions where interviewees are requested to report retrospectively not only the duration of a specific spell, but an entire history. As anticipated in the introduction, such reports are subject to additional forms of ME. In particular, the spells comprising the reported history will not only be affected by problems of their start or end being misdated, or any of the other typical forms of bias found in survey data, such as social desirability, or acquiescence bias, but also by more problematic issues of omission/overreporting and misclassification. Each of these forms of ME can distort the true durations of the spells in different ways. In Figure 1, we present three examples of the potential implications of ME when the retrieved durations are considered as the response variable of an event history model based on first spells (as opposed to multiple spells). The examples are taken from the data set presented in the following section, where all subjects start from a state of unemployment and the window of observation covers 395 days. The true spells of unemployment are now denoted by Y and are represented by the continuous lines. The error is represented by V and is encompassed by the bracket immediately below, whereas the observed durations are denoted by $Y*$.

When spells are misdated, the observed durations can appear shorter or longer than the true ones. This is represented by the first case shown in Figure 1, where the only spell of unemployment experienced within the window of observation has been reported to be 90 days longer than it really was. Misdating errors are the only types of recall errors that can occur when subjects known to be in one particular state are requested just to report the duration or the end date of that spell. This simple case could be well represented by the classical multiplicative model, where shorter durations will be associated with more accurate reports.

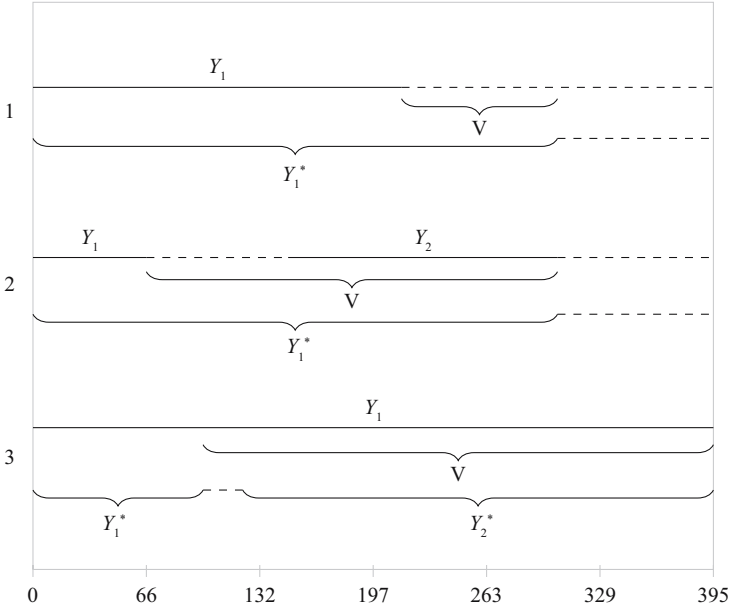

Fig. 1. Work history durations affected by different types of measurement error.

The omission of spells could more seriously distort estimation based on this data. Take the second work history presented in Figure 1, for example: if the spell representing a status different from unemployment starting in day 66 was omitted, the two spells of unemployment that occur before and after would be artificially linked, and the reported work history for that subject would look like a unique spell of unemployment. The same situation could also be reproduced as a result of misclassification of spells, specifically if the spell between Y_1 and Y_2 was a false positive case of unemployment (e.g., a registered case of employment reported as a spell of unemployment). An equally problematic form of ME occurs as a result of overreported spells. The third case in Figure 1 represents an example of a spell different from unemployment being mistakenly reported. This spell only covers 20 days (from day 100 to 120), but it has a severe effect, since it splits a true duration that was right censored and generates an observed duration only 100 days long.

Finally, we could also categorise a ME effect from misclassified spells in the form of false negative spells of unemployment. Given our particular setting, where all work histories start from unemployment and only first spells are considered, the effect of false negative spells will resemble that of missing data since these cases will be completely unobserved. The difference lies in the fact that in the presence of missing data we know which cases are missing. As long as the probability of committing false negatives is independent of the duration of unemployment and other observed explanatory variables, it will only affect the precision of model estimates.

3. 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-Åke Stenberg, and with the collaboration of the register of unemployment (PRESO). This register provided individual-level data on the work status of the participants of three surveys, run in 1992, 1993, and 2001. The sample was designed to capture 830 jobseekers randomly selected amongst those registered as unemployed on 1992-02-28 who met the following criteria: aged between 25 to 55 years, of Nordic nationality, no occupational disability, and seeking a full-time job. The three surveys are relatively similar with respect to the composition of both the sample of participants and the questionnaire, although for reasons of attrition the response rate for the three surveys decreased across time from 64.7%, to 59.4% and 50%. In this study, we use data derived from a retrospective question on work status from the 1993 survey, which reads as follows:

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

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

In order to simplify the reported work histories to be analysed, we set the beginning of the window of observation at 1992-02-28 and only consider subjects who started from a state of unemployment in both the register and the survey. Under this restriction, our sample mimics the structure seen in state-based samples (Holt et al. 2011), where the sample frame is created out of individuals who are known to be in a particular state. Our final sample size captures 381 individuals (from a total of 532 captured by both survey and

Table 2. Descriptive statistics of the sample.

	Mean/ Median	Standard deviation/ Interquartile range	Minimum	Maximum
Age	37	8.8	26	55
Experience	2.6	0.6	1	3
Register durations*	253	303	1	395
Survey durations*	92	144	1	395

*Since these variables are subject to censoring, their medians and interquartile ranges are reported.

register). The window of observation encompasses all spells from 1992-02-28 to 1993-03-30, where the end date represents the earliest day that interviews were held for the second wave of the survey. Right censoring is present in both the survey and register data sets.

In addition to drawing the duration of spells of unemployment from PRESO, the register also provides the age and experience of the 381 subjects. Given that *age* is an important variable in the register, the probability that it is prone to ME is very low. This is not so for *experience*, which 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). However, in our analysis we assume that both *age* and *experience* are free of ME. The value for *age* is taken in February 1993, which gives us a mean sample age of 37 and a standard deviation of 8.8, while for *experience* the mean of the monthly reported levels in 1992 is used, with a mean and a standard deviation in our sample of 2.6 and 0.6 respectively. The rest of the descriptive statistics for the variables used in our study are reported in Table 2.

4. Prevalence and Impact

In a study using this same data set, Pina-Sánchez et al. (2014) found evidence of the different types of errors discussed in Section 2. For example, 57% of the first spells reported were misdated by more than 31 days, and 30% were misclassified (interviewees reported to be employed or out of the labour force when, in fact, they were registered as unemployed, or vice versa). In addition, a tendency to omit spells of unemployment was detected since the ratio for the mean number of spells of unemployment reported over those registered during the window of observations was 1.4/1.7.

Here, we carry forward this analysis to assess the prevalence and effect of the ME found in spells of unemployment retrieved from work history reports. First, we look at differences between spells of unemployment from the same subjects and across the same window of observation captured by the survey and the register to estimate the prevalence of ME in the former. Second, we assess the impact of such ME on event history analysis, by looking at differences in regression coefficients and their measures of uncertainty between the same models when survey instead of register data is used. These analyses are based on the key assumption that spells from the register are free of ME. This is a realistic, yet not entirely perfect, assumption. Even data from registers can be affected by ME (Kapteyn and Ypma 2007; Pavlopoulos and Vermunt 2015). As a result, in the few

instances where the register is inaccurate, we might be wrongly considering spells of unemployment to be misreported.

4.1. Prevalence of Measurement Error in Durations of Unemployment

A comparison of medians (Table 2) shows that the durations of spells of unemployment are substantially longer in the register (253 days) than in the survey (92 days). This longer duration of registered spells is also reflected by the 133 cases (35% of the sample) that are right censored (remained unemployed by the end of the window of observation), compared with only 23 (6% of the sample) in the survey. These differences can be appreciated in Figure 2(a), where the survivor functions for the two types of durations are represented using Kaplan-Meier estimates (P(S) indicates the probability of not having made a transition out of unemployment at a particular time). The two data sets 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.

The different failure rates observed for the reported and registered durations indicate the presence of a systematic component in the ME process. In Figure 2(b) we plot the density of the error term assuming it is multiplicative, $V = Y^*/Y$. Here we can observe that, although a majority of errors are centred around one (as could be expected in a classical multiplicative framework), the distribution is bimodal and shows a substantial number of extreme values. The calculation of the error term might be biased since the right-censored cases were taken to be equal to 395. However, this could not account for the extreme values seen here. Hence, we can conclude that the ME process is not multiplicative.

The ME can also be assessed using scatter (c) and density plots (d). The former can be used to assess the effect of ME on a case-by-case basis, while the latter represents the

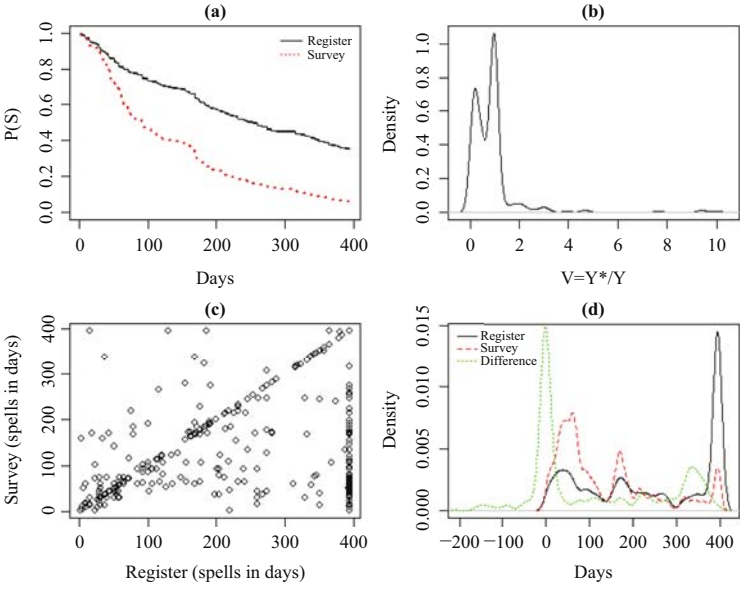

Fig. 2. Survey, register durations, and the effect of measurement error.

probability density functions of the error term as the difference between the registered and reported durations. First, we can see how a substantial proportion of cases are relatively unaffected by ME. This is reflected by the points lying around the diagonal line of the scatterplot, where survey and register durations are equal. In particular, 162 subjects, 42% of the sample, reported durations within ± 15 days of what was captured in the register. This pattern is also manifested by the dotted line depicting differences between registered and reported durations in the density plot, which shows a majority of cases for which the observed ME appears to be well approximated by a Normal distribution centred around zero, as could be expected from classical additive ME. However, the same density function also shows that for a considerable proportion of the sample, durations have been markedly shortened, together with a few other spells reported to be artificially long. For these cases, there does not seem to be a particular relationship between Y^* and Y since – except for the cases subject to classical additive ME – they are roughly uniformly distributed across the window of observation (as depicted by the dotted line).

The complexity of the ME seen here can be regarded as the outcome of the different ME mechanisms affecting the retrospective report of work histories (presented in Figure 1). In the light of the sample restrictions — namely the selection of spells for which the respondents reported correctly to be unemployed — we can differentiate two ME mechanisms: 1) spells where the difference between the registered and reported durations are distributed around zero, which could be considered to be due to problems of misdating; and 2) spells that are spread across the window of observation, which could be due to the problems of omission or false positives (case 2 in Figure 1), or to problems of overreported spells (case 3 in Figure 1).

4.2. *Impact of Measurement Error in an Exponential Model*

To assess the consequences of using durations of unemployment prone to this type of ME, we now compare the results obtained for two accelerated-life exponential models, one using durations from the register as the response variable (the true model) and the other relying on the reported durations (the naïve model) – see Pina-Sánchez et al. (2013) for a detailed review of the impact of this type of ME in different types of event history analysis models. An exponential model is used instead of other commonly used Weibull or Cox specifications (Kettunen 1997; Lancaster 1979; Pyy-Martikainen and Rendtel 2009) for reasons of parsimony. Given the monotonic increase of the cumulative hazard functions for the reported and registered durations (Figure 3) and the complexity of the adjustments carried out in the following sections, it was deemed preferable to use the simplest plausible model specification. We include the same set of explanatory variables, *age*, work *experience*, and their interaction term in these two models. These variables could be considered nondifferential (see Table 1) with respect to the ME observed here, since the Pearson correlation coefficients between the ME (defined as $V = Y^* - Y$) with *age* and *experience* were 0.07 and -0.01.

To facilitate comparisons with the adjustments presented in the following section, the true and naïve models are estimated using Bayesian methods. We specify a model with a hierarchical dependence structure that lends itself to straightforward estimation in software such as WinBUGS (Lunn et al. 2000), JAGS (Plummer et al. 2006), or Stan

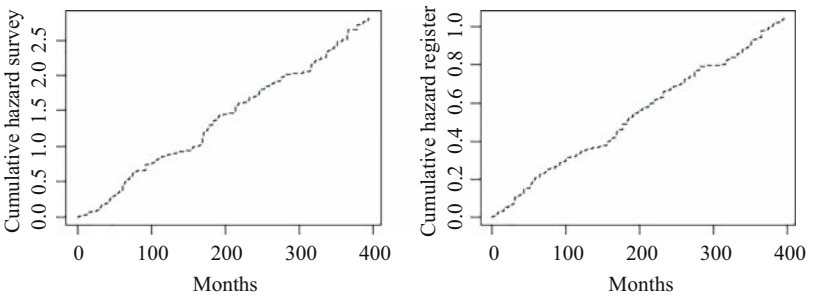

Fig. 3. Cumulative hazard functions for the reported and registered spells.

(Carpenter et al. 2017). All of these Bayesian packages are based on the Markov chain Monte Carlo (MCMC) approach (Geman and Geman 1984) that may be used when the full conditional posterior distributions of all unobservables are known distributions that are simple to draw from "directly".

The joint distribution for the true model can be formally expressed as

$$p(Y, \beta|X) = p(\beta)\prod_i p(Y_i|\mu_i), \tag{1}$$

where X represents the three explanatory variables, *age*, *exp*, and their interaction term *ae*, and μ_i is the expected value of Y_i. We assume that Y_i is exponentially distributed where we model μ_i using the link function $X_i\beta = \log(\mu_i)$. To complete the specification of the model we give diffuse priors to the regression estimates included in μ. For our 381 observations, we expect the likelihood to overwhelm the prior and we assume, a priori, that $\beta \sim N_4(0, 100^2 I_4)$. To estimate the naïve model, we only have to substitute Y for Y^* in Equation 1. Results from both the true and naïve models are shown in Table 3.

Although all the estimated coefficients have the same sign in both the naïve and true models, the effect of ME is clearly reflected by substantial attenuation in the estimates for the model using survey data. In particular, the effects of *age*, *experience* and their interaction on the durations of spells of unemployment are at least 90% smaller in the naïve model than in the true model. This has been measured using the relative bias (R.BIAS) for coefficient $\hat{\beta}_k^{(\text{survey})}$ relative to the true regression coefficient, $\hat{\beta}_k^{(\text{register})}$, $R.BIAS = 100\left|\hat{\beta}_k^{(\text{survey})} - \hat{\beta}_k^{(\text{register})}\right| / \left|\hat{\beta}_r\right|$, for $k = 1, 2, 3, 4$.

Table 3. Results of the exponential true and naive models and the impact of measurement error in terms of R.BIAS and R.RMSE.*

	Register	Survey	R.BIAS	R.RMSE
constant	9.10 (1.36)	5.12 (1.15)	43.7%	204.6%
age	−0.088 (0.038)	−0.001 (0.032)	89.9%	143.9%
experience	−1.39 (0.50)	−0.13 (0.42)	90.6%	171.1%
age × exp	0.038 (0.014)	0.002 (0.012)	94.7%	163.7%

*Results for all the models presented in this article are calculated from the posterior distributions formed from two chains of 400,000 iterations after the first 10,000 from each chain are burnt-in.
**Posterior standard deviations are shown within brackets.

The R.BIAS is useful for making comparisons between explanatory variables using different scales. In addition, in order to take into account the impact on the precision of the regression coefficients, we use the relative root mean squared error (R.RMSE), which is the root mean squared error of a regression coefficient obtained in the naïve model, $RMSE\left(\hat{\beta}_k^{(\text{survey})}\right) = \sqrt{SD\left(\hat{\beta}_k^{(\text{survey})}\right)^2 + (BIAS)^2}$, over that of the same coefficient in the true model, $R.RMSE = 100\frac{\left|RMSE\left(\hat{\beta}_k^{(\text{survey})}\right) - SD\left(\hat{\beta}_k^{(\text{register})}\right)\right|}{SD\left(\hat{\beta}_k^{(\text{register})}\right)}$.

In our study, the naïve model underestimates the standard deviations of all the regression coefficients. However, due to the attenuation of those estimates, the posteriors for the naïve model cover zero except for the constant term. The combination of bias and imprecision in the naïve model makes the impact of ME in terms of RMSE range from being doubled for the case of the constant term, to an increase of 43.9% (for *age*).

5. Adjustment

The great impact of the ME found in spells of unemployment derived from retrospectively reported work histories just shown is even more worrying if we take into account the difficulty of adjusting for such a complex process. Most of the standard methods designed for the adjustment of ME rely on rather simple assumptions regarding the behaviour of the ME. For example, methods such as SIMEX (Cook and Stefanski 1994), or the methods of moments (Fuller 1987) tend to assume that the error is classical (see Table 1), whereas other methods that can easily account for systematic ME, such as multiple imputation (Brownstone and Valletta 1996; Cole et al. 2006; Freedman et al. 2008; Messer and Natarajan 2008; Peytchev 2012; Rubin 1987) or regression calibration (Carroll and Stefanski 1990; Freedman et al. 2008; Glesjer 1990; Messer and Natarajan 2008; Veronesi et al. 2011; Wang et al. 1997) assume that the ME process is homoscedastic.

The problem of ME that changes in size according to the true value is typical in the retrospective report of start or end of spells. Several studies in the literature have investigated the adjustment of such types of errors using multiplicative models (Augustin 1999; Biewen et al. 2008; Dumangane 2007; Glewwe 2007; Holt et al. 2011; Pickles et al. 1996; Pickles et al. 1998; Pina-Sánchez 2016; Skinner and Humphreys 1999). However, as seen in the previous section, the type of ME observed in spells of unemployment derived from retrospectively reported work histories cannot be approximated using a multiplicative model. In particular, we have seen how, in addition to typical problems of spells being misdated, we should expect more serious problems of omission/overreport and misclassification of spells, which can generate cases where the observed and true durations are unrelated.

To adjust for such complex types of ME, we rely on the flexibility of the Bayesian approach. The possibility of specifying the ME freely to map the error-generating mechanism adequately is the key element that gives Bayesian methods an advantage over the previously discussed methods in terms of the flexibility and applicability in treating missing data (Rubin 1996). See for example, Clayton (1992) for a general framework; Richardson and Gilks (1993) demonstrated how the missing data allows for treatment of a wide variety of ME; Dellaportas and Stephens (1995) deal with both Berkson and classical

errors-in-variables ME for regressors; Butts (2003) treat ME in perceptions of social interaction; Ghilagaber and Koskinen (2009) correct for ME stemming from retrospectively defined covariates.

In what follows, we explore the possibility of extending the naïve model presented using a mixture measurement model with the aim of differentiating between two ME processes. The first process will reflect cases being either correctly reported or misdated. The second will deal with spells that result from misclassification, omission or overreporting of events, and for which the observed duration cannot offer much, if any, meaningful information about the true duration. In order to inform the mixture model about the allocation of cases to each of these processes, we explore two approaches: one relies on having access to a validation subsample, while the other relies on specific prior information.

5.1. Adjustments Relying on Validation Subsamples

For the first of these adjustments, we use random validation subsamples of 20% and 40% of the true durations captured in the original sample of 381 respondents. Compared to other studies seeking to adjust for ME, 20% is a relatively small subsample – for example Cole et al. (2006) used validation subsamples of 150 cases, accounting for 25% of the original sample – but large enough to make the model identifiable regardless of the configuration of the validation subsample. This was not always the case for smaller validation subsamples of 5% and 10%. A 20% validation subsample will, on average, only have 44 noncensored and 32 censored cases.

We present a general hierarchical mixture measurement model and then proceed to define the parameter structure and prior distributions. We let $\Lambda = \left(\beta, \theta, \sigma_1^2, \sigma_2^2, \pi\right)$ denote the collection of parameters of interest and introduce a latent variable $T = (T_i: i = 1, \ldots, 381)$. With prior distribution $p(\Lambda)$, the joint distribution of data and the unobservables is

$$p(Y^*, Y, \beta, \Lambda, T|X) = \prod_i p\left(Y_i^*|Y_i, \Lambda, T_i\right) p(T_i|X_i, Y_i, \Lambda) \tag{2}$$

$$\times \prod_i p(Y_i|X_i, \Lambda) \tag{3}$$

$$\times\ p(\Lambda), \tag{4}$$

where the model (3) for the true response is defined as an exponential distribution $p(Y_i|X_i, \Lambda) = p(Y_i|X_i, \beta)$ as before, but conditionally on a response Y_i, we define the ME distribution (2) for Y_i^* as a mixture of different ME models.

5.1.1. Mixture Measurement Model

The mixture model permits us to account for different types of ME simultaneously. We could, for example allow some cases to follow a multiplicative ME, while others follow an additive ME model. Here, we limit the case to two types of ME. For each respondent $i = 1, \ldots, 381$, we assume that they belong to one of two unobserved categories indicated by the latent variable $T_i \in \{0, 1\}$. If $T_i = 1$, we assume a standard additive ME

model $Y_i^* \big| \left[T_i = 1, Y_i, \sigma_1^2\right] \sim N\left(Y_i, \sigma_1^2\right)$. Conditional on $T_i = 0$, we assume that Y_i^* is completely unrelated to the true duration Y_i. Denoting this distribution $f\left(Y_i^*\right)$ and writing $\pi_i = \Pr(T_i = 1 | \pi, X_i, Y_i)$, the two processes correspond to a mixture

$$\pi_i \phi\left(Y_i^*; Y_i, \sigma_1^2\right) + (1 - \pi_i) f\left(Y_i^*\right),$$

where $\phi(\cdot;\ a,\ b)$ represents the pdf of a normal distribution with mean a and variance b.

The intuition is that additive ME might work for a subset of data but not for another subset. For a purely additive model, this heterogeneity would have to be accounted for entirely by the variance σ_1^2. A number of distributions for $f(\cdot)$ are conceivable and the shape of the distribution may contribute information and help discriminate between additive and random ME. To allow for $f(\cdot)$ to depend on for example X, would not alter the general structure of the model as long as the parameters of $f(\cdot | X)$ were distinct from β. Here we assume a normal distribution $f\left(Y_i^*\right) = \phi\left(Y_i^* \big| \theta, \sigma_2^*\right)$ which is convenient and can be interpreted as a variance decomposition of the ME. The conditional predictive distribution for Y_i^* with unknown T_i simplifies to

$$Y_i^* \big| \left[\pi_i, Y_i, \theta, \sigma_1^2, \sigma_2^2\right] \sim N\left(\pi_i Y_i + (1 - \pi_i)\theta, \pi_i^2 \sigma_1^2 + (1 - \pi_i)^2 \sigma_2^2\right).$$

Since respondents cannot report negative durations, Y_i^* can be truncated to the left in 0. Here we chose to relax this constraint and permit Y_i^* to have support $\mathbb{R}$. The reason for this is partly because the MCMC becomes less efficient with the truncation. Another consideration is that the truncation confounds the variance partition and makes interpretation of ME in terms of classical ME difficult. The distribution of the true value Y_i given Y_i^* and Λ, will be nonnegative even if Y_i^* is not truncated (see Section 7 Appendix for details).

5.1.2. Mixture Proportions

The latent variable, T, denoting the part of the mixture model to which cases are allocated, is set to follow a Bernoulli distribution $\Pr(T_i = 1 | Y,\ X,\ \Lambda) = \pi$, independently for all respondents conditional on the mixture proportion π. A priori π determines what proportion of cases follow the classical ME and what cases are unrelated. A posteriori, the predictive distribution for individual memberships also incorporates the information in the other variables. Like other latent variable models, it is possible to model π_i but we chose the more straightforward case.

5.1.3. Prior Distributions

To balance the evidence in data for the different ME processes and to adjust proportions, we need to pay careful attention to prior distributions. To reflect the structure of the model, we set prior distributions

$$p(\Lambda) = p(\pi) p(\beta)\, p\left(\theta, \sigma_1^2, \sigma_2^2\right),$$

assuming a priori independence between different blocks of parameters. As $\pi \in (0, 1)$, a convenient and common choice is $Beta(\zeta_1, \zeta_2)$, with standard reference priors being $\zeta_1 = \zeta_2 = 1/2$ (Jeffreys, 1946) or a uniform distribution $\zeta_1 = \zeta_2 = 1$. Here we chose the latter.

The analysis of the full register data was not very sensitive to the prior for β. With 20% or 40% validation samples, the scant information in data will, however, give more weight to the prior distribution. For regression parameters β, it is common in generalised linear models to use the prior distribution $\beta \sim N_p(0, \text{diag}(\lambda))$, for $\lambda = (\lambda_1, \lambda_2, \lambda_3, \lambda_4)^T$. This may be interpreted as a prior that shrinks the coefficients towards the origin. We set a relatively conservative $\lambda = (100^2, 10^2, 10^2, 10^2)^{\mathrm{T}}$, where the variance λ_1 for β_{cons} reflects the greater variation in this parameter and prevents too great attenuation of the intercept. Other alternatives include the class of conjugate families (Chen and Ibrahim 2003), or other forms of reference priors – for example, Ibrahim and Laud (1991) – that can account for the difference in scale and correlation between covariates.

As the ME model attempts to parse out the relative contribution of variance in $Y* - Y$, from the two kinds of ME, the model is potentially sensitive to the prior specifications for $p\left(\theta, \sigma_1^2, \sigma_2^2\right)$. Since we do not have any theory to guide us and given the structure of the mixture model, we assume that $\left(\theta, \sigma_2^2\right)$ are independent of σ_1^2 a priori. Here, we consider two different kinds of prior distributions.

The first type is a standard conjugate prior

$$\frac{1}{\sigma_1^2} \sim Ga(\alpha_1, \gamma_1),$$

and for the parameters of f we chose a standard normal-inverse-gamma prior

$$\theta \big| \sigma_2^2 \sim N\left(\theta_0, \sigma_2^2/n_0\right) \quad \frac{1}{\sigma_2^2} \sim Ga(\alpha_2, \gamma_2),$$

for hyper parameters $\alpha_1,\ \alpha_2 > 2,\ \gamma_1,\ \gamma_2 > 0$, and $n_0 > 0$. To make sure that both the variances and the precisions have proper distributions with the first two moments finite, we set $\alpha_1 = \alpha_2 = 3$. This is to ensure that the conditional posteriors are well defined for the case when $\sum_i T_i = 0$ or $\sum_i T_i = 381$, that is, when the model allocates all observations to either one of the two latent classes. Throughout, we will set $n_0 = 1$ and $\theta_0 = 187$, which is exactly the centre of the range of observable values on Y_i^*. Because of the (conditional) conjugacy, updates of α_1,α_2, γ_1,γ_2, and θ are efficient. Similarly, updates of unobserved Y_i are straightforward given the result in the Section 7 Appendix.

To relax the dependence between θ and σ_2^2 somewhat, and make $f()$ slightly more robust to violations of normality, we also consider a prior where θ and σ_2^2 are independent a priori. More specifically, we set the prior for $\theta \sim N(\theta_0, \tau^2)$, $\theta_0 = 187$, and independently thereof $\sigma_s^2 \sim U(1, a_s)$, for some choice of upper bound a_s, $s = 1, 2$ (Gelman et al. 2006, 520).

5.2. *Performance of the Adjustment*

To investigate the performance of the adjustment, we assess the sampling error associated with the choice of validation subsample, as well as the sensitivity to prior specifications. We estimate the model from 50 samples of 20% and 40% of the register data for prior specifications, as in Table 4.

5.2.1. 20% Validation Sample

A comparison of the posteriors for β with the naïve analysis and the gold standard is provided in Figure 4 that provides the 95% credibility intervals (CI) for the adjustments

Table 4. Specifications for hyper parameters for σ_1^2 and σ_2^2.

	γ_1	γ_2	$E(\sigma_1^2)$	$E(\sigma_2^2)$	$V(\sigma_1^2)$	$V(\sigma_2^2)$
Prior I	5	10	2.5	5	6.25	25
Prior II	10	100	5	50	25	2500
Prior III	100	400	50	200	2500	40000

under priors I, II, and II, as well as the CIs and means for the validation sample only and register analyses. While the posteriors for π, σ_1^2, σ_2^2, and θ clearly are sensitive to the prior specifications, the posteriors for the regression parameters β appear robust to the priors for the ME. With as few as 76 validation cases, it appears that we get valid inference for both β and T and there seems too be a very weak dependence on the actual sample judging by the small variation across samples. The posterior means for the adjustments seem to be close to those obtained using the validation sample only, but the CIs for the latter are much wider than for the adjustments. Using the survey only yields less uncertainty and narrower CIs (as shown in Table 3) than the adjustment, but the bias of the survey results means that the CIs do not cover the true posterior mean for any parameter.

All regression coefficients for the adjustment are attenuated and the CIs are somewhat wider. For β the adjustment CIs are close to identical for Priors I, II, and III. There is a

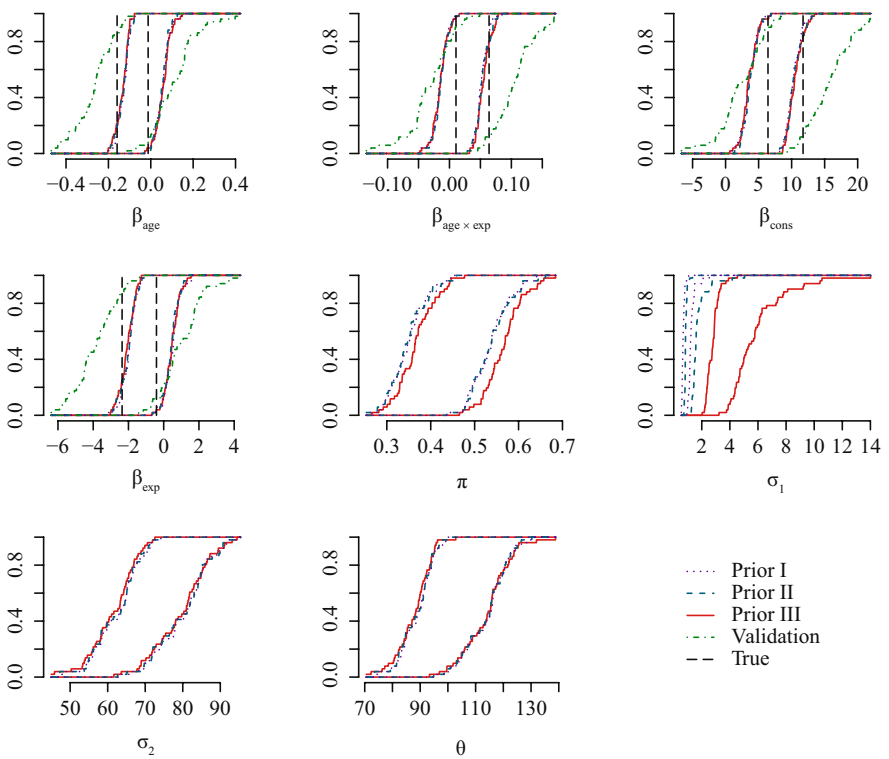

Fig. 4. 20% validation samples: CDF of Credibility intervals under prior specifications I, II, III, for 50 replicates, compared to the naïve estimates and the validation sample only.

healthy overlap between the adjustment CIs and the true CIs, but the interval lengths for the latter are shorter. The relative bias for *age* ranges between 5 and 129, with a mean of 58; for the constant, the relative bias ranges between 1 and 46 with a mean of 23.

We may note that on balance, the first part of the mixture model $T_i = 1$ makes a relatively small contribution in the adjustment. The fact that the model can differentiate between these cases and those much more seriously affected by ME is critical for the success of the adjustment. Using the 20% validation subsample, the mixture model estimates that the proportion of cases set to $T_I = 1$ (π) is generally in the range 30% to 60% which covers 46%, the proportion of cases in the validation subsample where reported durations lay within ± 15 days of what was captured in the register.

While the inference for β seems robust to the prior specification using the tight coupling of θ and σ_2^2 of the conjugate prior, we also fitted the adjustment model using the second kind of prior with $a_1 = 50$ and $a_2 = 500$ (Prior IV). Again, the posterior means and standard deviations are not affected by the prior specification for the ME. Prior IV partitions the variance of the ME more clearly but is very similar to Priors I and II (Figure 5).

5.2.2. 40% Validation Sample

Increasing the size of the validation subsample to 40% yields more information both for the regression parameters β and the ME process. We re-estimate the model under Prior II

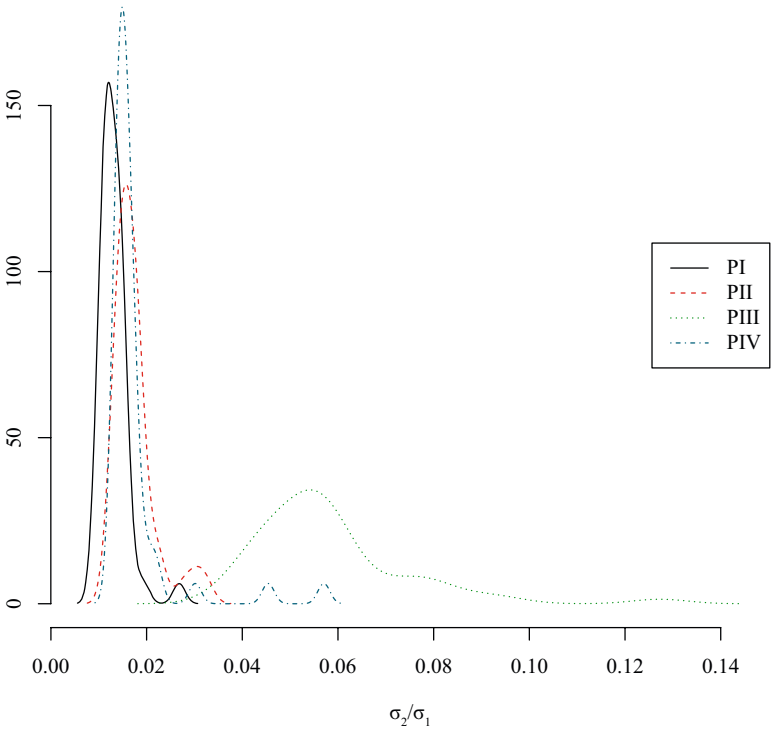

Fig. 5. Comparison of posterior ratio between variance contributions of ME different types of priors across 20% validation samples.

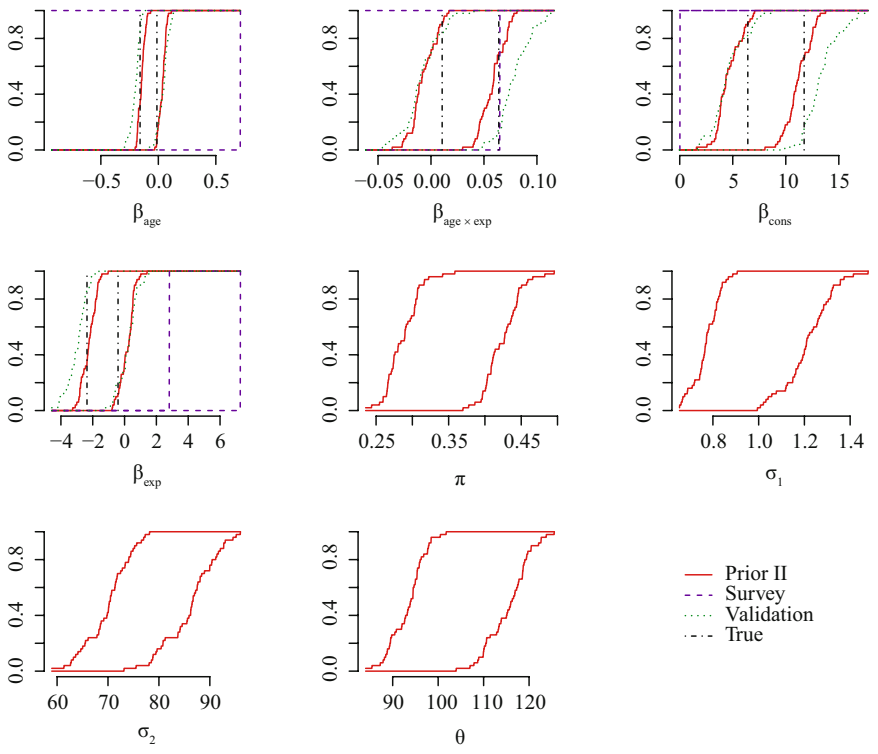

Fig. 6. 40% validation samples: CDF of Credibility intervals under prior specification II, for 50 replicates, compared to the naive estimates and the validation sample only.

(given that the inference for 20% was robust to choice of prior). The posterior means for β are still somewhat attenuated, but the CIs (Figure 6) are now very close to the intervals for the register data. This improvement is not matched by the validation-only analyses. For example, the average interval length for *age* for the adjustment is 0.18 and for validation only it is 0.24.

5.3. Adjustments Without Validation Sample

Without a validation subsample, the previous model would have been unidentifiable since it would be impossible, using the survey data alone, to determine the probability of cases belonging to each part of the mixture model. However, the requirement of having access to validation data can be relaxed using more informative priors. In particular, we obtain results by assuming that the probability of cases falling within each of the two parts of the mixture model is fixed. Prior distributions become a very useful tool in the presence of models that cannot be identified – as is often the case when dealing with ME problems. *"One intuitive way of thinking about Bayesian inference in the absence of parameter identifiability is that the prior distribution plays more of a role than usual in determining the posterior belief about the parameters having seen the data"* (Gustafson 2003, 64). However, this greater reliance on the priors implies a reduced role for the data and, in consequence, an increased possibility of incurring in model misspecification.

The more informed the researcher is about the value of the model parameters, the lower the probability of misspecification will be. Hence, studies designed to assess the presence of ME are essential to carrying out adequate adjustments. A number of studies have explored the problems of ME affecting retrospectively reported work histories (Biemer 2011; Kreuter et al. 2010; Manzoni et al. 2011 and 2010; Poterba and Summers 1984 and 1995; Pyy-Martikainen and Rendtel 2009). However, we are only aware of one study (Pyy-Martikainen and Rendtel 2009) using a register as validation data to assess the different types of ME found in these types of questions, and even here we should note some important differences from our study. Pyy-Martikainen and Rendtel (2009) looked at a representative sample of the Finnish population, studied spells reported over a period of five years, and the question used only required respondents to report the work status experienced in each month of the year. We, on the other hand, study a sample composed of Swedish jobseekers, observed for a period of one year, and responding to a question where every spell needed to be identified in chronological ordered and dated.

These differences in terms of the sample composition, window of observation, and question format, are so important that it would probably be unwise to borrow specific point estimates from Pyy-Martikainen and Rendtel (2009) for our adjustment. However, we can take the overall lower prevalence of ME found in Pyy-Martikainen and Rendtel (2009) into consideration as one of the scenarios that we explore. In particular, we carry out adjustments ranging from $\pi = 0.5$ – a rounded estimate of the proportion of reported durations that lay within ± 15 days of the true ones – to $\pi = 0.7$ – an estimate that assumes a lower prevalence of cases being misclassified, omitted or overrepresented. We assume Prior IV for the ME process.

Judging from Figure 7, inference for β is little affected by our choice of π but the added uncertainty of not having a validation sample is reflected in spread in the posteriors. The bias (relative to the register means) increases for all parameters as π increases. This is more clear in Table 5, and the bias for the two extremes are compared in Figure 8.

We see that the models explored succeeded in reducing the bias found in the naïve analysis (Figure 8), while some of the added uncertainty is reflected in inflated standard deviations. The model assuming $\pi = 0.5$ performed better at reducing the R.BIAS and

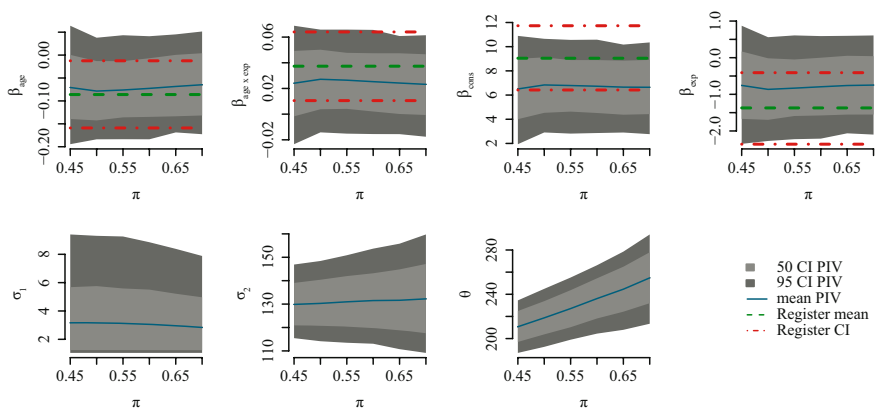

Fig. 7. No validation samples: Credibility intervals (50% and 95%) under prior specification IV against proportion π.

Table 5. Adjustment using mixture models with fixed proportions.*

	Register	Survey	$\pi = .5$	$\pi = .7$
constant	9.10 (1.36)	5.12 (1.15)	6.85 (1.98)	6.64 (1.94)
age	−0.088 (0.038)	−0.001 (0.032)	−0.079 (0.056)	−0.065 (0.058)
experience	−1.39 (0.50)	−0.13 (0.42)	−0.87 (0.72)	−0.75 (0.70)
age × exp	0.038 (0.014)	0.002 (0.012)	0.027 (0.020)	0.023 (0.020)
σ_1			3.13 (2.22)	2.80 (1.83)
σ_2			130.4 (8.7)	132.2 (12.9)
θ			218.9 (13.0)	254.5 (19.8)

*Posterior standard deviations are shown in brackets.

R.RMSE. The R.RMSE for the constant is reduced from 204% to 120%, and from 144% to 49% for *age*.

It seems that a value of π closer to what can be observed in the sample helps in the reduction of the R.BIAS observed in the naïve model. However, the lower the value of π the higher the proportion of cases being treated by the second part of the mixture model. Finally, if we compare the adjustments based on 20% and 40% validation subsamples, they generally have slightly higher R.BIAS and R.RMSE on average than models relying on a fixed $\pi \in (0.5, 0.7)$, but with a range that covers those for the latter. For example, R.BIAS for *age* for the 20% adjustment is between 4.7 and 129 with an average of 58; and the 40% adjustment is between 0.3 and 124 with an average of 40. R.RMSES for *age* for the 20% adjustment is between 14.9 and 215 with an average of 85; and the 40% adjustment is between 11 and 205 with an average of 58.

6. Discussion

As has been noted by different authors (Augustin 1999; Jäckle 2008; Pyy-Martikainen and Rendtel 2009), the understanding of and adjustment for ME in longitudinal data remains an understudied area. *"Despite the recognition of the existence of measurement errors in*

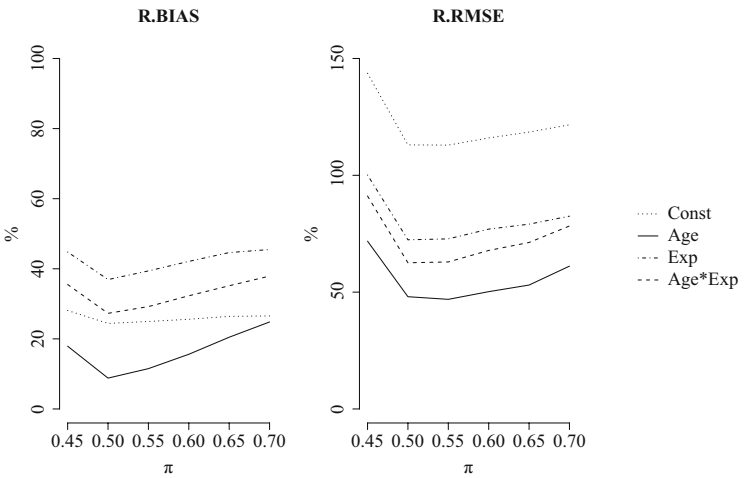

Fig. 8. Effectiveness of the adjustment using a mixture model with fixed π.

survey-based data on event histories, little is known about their effects on an event history analysis." (Pyy-Martikainen and Rendtel 2009, 140). Our intention here has been to contribute to this topic by studying the prevalence, impact, and adjustments of the type of ME observed in spells of unemployment derived from retrospectively collected work histories using register data. From this analysis, we would like to underline three main points, each being a consequence of the previous one:

1) Unlike the typical problems of time displacement found in reports of the starts or ends of specific spells, here we have noted that reports of work histories can also be prone to ME in the form of misclassification omission and overreporting of spells. Furthermore, these different types of ME can occur simultaneously, generating complex ME patterns.
2) Misclassification, omission and overreporting of spells are more problematic forms of ME than misdated spells because they tend to result in observed durations that are unrelated to the true values, hence endangering both the validity of these measures and the analyses based on them. In particular, we have shown how, when spells of unemployment derived from retrospectively reported work histories are used as the response variable of an event history model, its regression coefficients are severely attenuated by up to 90% of their true value.
3) Because of the complexity of the ME processes studied, standard adjustment methods are inadequate. In particular, it is important to note that the vast majority of studies aiming to adjust for the ME found in retrospective reports of dates have been based on elaborations of multiplicative ME, which are inadequate in the presence of spells being omitted, overreported, or misclassified.

Here, we have proposed to adjust for different ME mechanisms simultaneously through the use of a Bayesian mixture measurement model. Specifically, the mixture model allows us to differentiate between spells that are correctly reported or affected by minor problems of misdating, and those for which the observed durations are not related to their true value. To inform the mixture model about the proportion of cases that needed to be predicted according to each of the mechanisms, we explored two approaches.

First, we assumed that the researcher could have access to a validation subsample. The method performs well for 40% validation sample, but also works for 20% validation sample. Here, 20% means only 76 cases from the register (32 on average of which are censored), but even such a small figure is often inaccessible to most researchers due to reasons of confidentiality. As an alternative to using validation data, we demonstrate the use of fixing, a priori, the proportion of subjects in one of the two ME processes when no validation subsample is available. These adjustments were even more effective. For a model where the probability of being in the first part of the mixture model is fixed at 0.5 – reflecting the percentage of spells observed to be correctly reported or simply mildly misdated – we obtained average reductions of the R.BIAS of 65%, whereas fixing that probability at 0.7 showed average reductions of 54%. In addition, these models fixing the probabilities of the mixture model at 0.5 and 0.7 reduced the R.RMSE by 41% and 47%, respectively.

The adjustment methods presented here are potentially useful strategies to reduce the impact of the types of ME that can be expected in the report of life-course histories. The

adjustments have been kept relatively simple and could be implemented by other researchers using data that are similarly prone to combinations of different types of ME. It is, for example, straightforward to model the mixture proportions as a function of observables and to relax assumptions for the form of unrelated ME. The question remains as to how to choose the necessary priors to make the model identifiable when no validation data is available. Findings from studies analysing the prevalence of ME in similar survey questions could help to inform those decisions, but even in the context of a well-analysed question, a sensitivity analysis, or the comparison of results obtained from the use of different priors, is the most prudent solution. For example, future studies specifying durations of unemployment stemming from similar retrospective questions to the one analysed here could use our approach to explore the robustness of their findings when the percentage of spells taken to be misclassified, omitted, or overreported grows from 0% to 20% and 40%. Future work is needed to explore to what extent replicate data from the posterior predictive distributions can be used to assess what ME has best fit to observed data (Gilks et al. 1996).

The quality of the adjustment could also be improved using the, nowadays, more frequently coded paradata. For example, the probability of cases being considered by each part of the mixture model could be made conditional on certain factors known to be associated with the misclassification, omission or overreporting of spells of unemployment, such as whether the interview was taken on the phone rather than face-to-face (Mathiowetz and Duncan 1988). Such adjustments could be further improved if the researcher has access to key socio-demographic characteristics of the respondent. For example, it has been consistently detected that individuals less engaged in the labour market are more prone to omit spells of unemployment (Bound et al. 2001; Jürges 2007; Levine 1993; Morgenstern and Barrett 1974; Paull 2002).

We would also like to add a note of comfort regarding the high levels of both the prevalence and the impact of the ME analysed here, since there are reasons to believe that they might be unusually high. First, the sample under study here is not representative of the Swedish population, as it is composed of strictly unemployed subjects, who, for reasons of social desirability might have a higher tendency to omit their spells of unemployment. Second, the register of unemployment is not infallible. We detected coding errors in the form of nonsensical dates, for example spells that started before the previous spell had ended, or others dated the 32nd day of a month. These cases were dropped from the sample, but other coding errors in the register might have gone undetected, which should make us aware that a proportion of the differences between the register and the survey were actually reflecting ME in the former and not in the latter. Finally, the format of the question used is cognitively quite demanding. Nowadays, longitudinal surveys like the European Union Statistics on Income and Living Conditions (EU-SILC), or the Swedish Level of Living Survey (LNU), use questions where respondents are only asked to report their work status in each of the months of the previous year. These questions will fail to detect transitions shorter than a month, but since they are easier to answer we could also expect lower levels of ME than in questions where all work-related spells are asked to be reported and dated with day-level detail.

The different format of these questions affects the level of measurement of the variables to be retrieved from them, and with that, the modelling strategies to be used. Rather than

obtaining duration data amenable to parametric event history model specifications – like the exponential model presented here – person-period categorical data is obtained, which is more suitably specified using nonparametric models (Box-Steffensmeier and Jones 2004). The details of the adjustment presented here would not be directly applicable in those instances. Instead, we would direct the interested reader to relatively recent studies from Manzoni et al. (2010) and Biemer (2011) implementing adjustments based on latent Markov models and repeated measures.

7. Appendix – Conditional Predictive Distribution for True Values

For the un-truncated normal distribution, the location and scales are independent and normal conjugacy applies. This makes it straightforward to model θ, σ_2^2, and σ_1^2 such that data are able to determine class memberships (T_i: $i = 1, \ldots, 381$). The conditional posterior predictive distribution of the true value Y_i given Y_i^* and Λ, is a normal distribution truncated to the left in 0. Note that this only holds when Y_i^* itself is not truncated.

Let $\eta = \pi Y + (1 - \pi)\theta$, $\tau = \frac{1}{2(\pi_i^2\sigma_1^2 + (1-\pi_i)^2\sigma_2^2)}$, and set $z = Y^*$, then $y|[z, \Lambda]$ has pdf

$$p(y|z, \Lambda) = \frac{p(z|y, \Lambda)p(y|\Lambda)}{\int p(z|y, \Lambda)p(y|\Lambda)\mathrm{d}z} = \frac{\exp\{-\tau(z - \eta)^2 - y/\mu\}1(y > 0)}{\int \exp\{-\tau(z - \eta)^2 - y/\mu\}1(y > 0)\mathrm{d}y}.$$

The integrand in the denominator can be written

$$\exp\{-\tau(z - \eta)^2 - y/\mu\}\mathbf{1}(y > 0) = h(z, \Lambda)g(y; z, \Lambda)$$

where $h(z, \Lambda)$ is a only a function of z and Λ, and

$$\begin{aligned} g(y; z, \Lambda) &= \exp\left\{-\tau\pi^2\left(y^2 - y\lambda + \frac{ts}{\pi^2}\right)\right\}\mathbf{1}(y > 0) \\ &= \exp\left\{-\tau\pi^2(y - \lambda)^2 - \tau\pi^2\left(\lambda^2 - \frac{ts}{\pi^2}\right)\right\}\mathbf{1}(y > 0) \\ &= (\pi\tau)^{-1/2}\phi(y; \lambda, \upsilon^2)c(z, \theta, \mu)\mathbf{1}(y > 0) \end{aligned}$$

Where $s = z + (1 - \pi)\theta$,

$$\lambda = \frac{\pi t - \frac{1}{\mu\tau}}{\pi^2}, \quad \upsilon^2 = \frac{1}{2\pi^2\tau}, \quad t = z - (1 - \pi)\theta,$$

and

$$c(z, \theta, \mu) = e^{-\tau\pi^2(\lambda^2 - ts/\pi^2)}.$$

Since $c(z, \theta, \mu)$ is a function of z, θ, and μ, only, and $h(z, \Lambda)$ is a only a function of z and Λ, we have

$$p(y|z, \Lambda) = \frac{\phi(y; \lambda, \upsilon^2)\mathbf{1}(y > 0)}{\int \phi(y; \lambda, \upsilon^2)\mathbf{1}(y > 0)\mathrm{d}y} = \frac{\phi(y; \lambda, \upsilon^2)\mathbf{1}(y > 0)}{\Phi(-\lambda/\upsilon)}.$$

We recognise this as a normal distribution $N(\lambda, \nu^2)$, truncated in the left at 0. Consequently, even when f () is inconsistent with data in the sense that Y_i^* may take negative values, the conditional predictive distribution for the true values Y_i is still consistent with data.

Consider now the case when the observations with error Y^* are themselves truncated to the left in 0. For the normal distribution $\phi(\cdot; a, b)/\Phi(-a/b)$ truncated to the left in 0, the variance is a function of a (for example, for fixed b, setting $a < 0$ increases the variance). For our data, the truncation results in θ becoming negative so that the variance in $Y_i^* \big| \left[\theta, \sigma_2^2, T_i = 0\right]$ is determined by the right tail of the truncated distribution rather than the variance of the distribution. Negative mean θ and the dependence of $V\left(Y_i^* \big| \theta, \sigma_2^2, T_i = 0\right)$ on both θ and σ_2^2 makes the interpretation in terms of classical ME difficult and sampling using MCMC inefficient.

8. References

Augustin, T. 1999. Correcting for Measurement Error in Parametric Duration Models By Quasi-likelihood. Technical Report, Max Plank Institute.

Berkson, J. 1950. "Are There Two Regressions?" *Journal of the American Statistical Association* 45(250): 164–180. Doi: https://doi.org/10.2307/2280676.

Biemer, P.P. 2011. *Latent Class Analysis of Survey Error*. Wiley.

Biewen, E., S. Nolte, and M. Rosemann. 2008. "Perturbation by Multiplicative Noise and the Simulation Extrapolation Method." *Advances in Statistical Analysis* 92: 375–389. Doi: https://doi.org/10.1007/s10182-008-0089-7.

Black, D.A., M.C. Berger, and S.A. Scott. 2000. "Bounding Parameter Estimates with Nonclassical Measurement Error." *Journal of the American Statistical Association* 95(451): 739–748. Doi: https://doi.org/10.2307/2669454.

Bound, J., C. Brown, and N.A. Mathiowetz. 2001. "Measurement Error in Survey Data." In *Handbook of Econometrics*, edited by J. Heckman and E. Leamer. Vol. 5: 3705–3843. New York: Elsevier.

Box-Steffensmeier, J.M. and B.S. Jones. 2004. *Event History Modeling: A Guide for Social Scientists*. Cambridge: Cambridge University Press.

Bradburn, N.M., J. Huttenlocher, and L. Hedges. 1994. "Telescoping and Temporal Memory." In *Autobiographical Memory and The Validity of Retrospective Reports*, edited by N. Schwarz and S. Seymour, 203–215. New York: Springer.

Brownstone, D. and R.G. Valletta. 1996. "Modelling Earnings Measurement Error: A Multiple Imputation Approach." *The Review of Economics and Statistics* 78(4): 705–717. Doi: https://doi.org/10.2307/2109957.

Butts, C.T. 2003. "Network Inference, Error, and Informant (in) Accuracy: A Bayesian Approach." *Social Networks* 25(2): 103–140. Doi: https://doi.org/10.1016/S0378-8733(02)00038-2.

Carpenter, B., A. Gelman, M. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M.A. Brubaker, J. Guo, P. Li, and A. Riddell. 2017. "Stan: A Probabilistic Programming Language." *Journal of Statistical Software* 76(1): 1–32. Doi: https://doi.org/10.18637/jss.v076.i01.

Carroll, R.J. and L.A. Stefanski. 1990. "Approximate Quasilikelihood Estimation in Models with Surrogate Predictors." *Journal of the American Statistical Association* 91: 242–250. Doi: https://doi.org/10.2307/2290000.

Chen, M.H. and J.G. Ibrahim. 2003. "Conjugate Priors for Generalized Linear Models." *Statistica Sinica* 13: 461–476. Available at: http://www3.stat.sinica.edu.tw/statistica/oldpdf/a13n212.pdf (accessed February 2019).

Clayton, D.G. 1992. "Models for the Analysis of Cohort and Case-control Studies with Inaccurately Measured Exposures." *Statistical Models for Longitudinal Studies of Health*: 301–331.

Cole, S., H. Chu, and S. Greenland. 2006. "Multiple-imputation for Measurement-error Correction." *International Journal of Epidemiology* 35: 1074–1081. Doi: https://doi.org/10.1093/ije/dyl097.

Cook, J. and L. Stefanski. 1994. "A Simulation Extrapolation Method for Parametric Measurement Error Models." *Journal of the American Statistical Association* 89: 1314–1328. Doi: https://doi.org/10.2307/2290994.

Crowder, R.G. 1976. "The Interference Theory of Forgetting in Long-term Memory." In *Principles of Learning and Memory*, edited by R.G. Crowder. Oxford: Lawrence Erlbaum.

Dellaportas, P. and D.A. Stephens. 1995. "Bayesian Analysis of Errors-in-variables Regression Models." *Biometrics* 51(3): 1085–1095. Doi: https://doi.org/10.2307/2533007.

Dumangane, M. 2007. Measurement error bias reduction in unemployment durations. Technical report, CEMMAP. Doi: https://doi.org/10.1920/wp.cem.2006.0306.

Freedman, L.S., D. Midthune, R.J. Carroll, and V. Kipnis. 2008. "A Comparison of Regression Calibration, Moment Reconstruction and Imputation for Adjusting for Covariate Measurement Error in Regression." *Statistics in Medicine* 27: 5195–5216. Doi: https://doi.org/10.1002/sim.3361.

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

Gelman, A. et al. 2006. "Prior Distributions for Variance Parameters in Hierarchical Models (Comment on Article by Browne and Draper)." *Bayesian Analysis* 1(3): 515–534. Doi: https://doi.org/10.1214/06-BA117A.

Geman, S. and D. Geman. 1984. "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images." *IEEE Transactions On Pattern Analysis and Machine Intelligence* 6: 721–741. Doi: https://doi.org/10.1109/TPAMI.1984.4767596.

Ghilagaber, G. and J. Koskinen. 2009. "Bayesian Adjustment of Anticipatory Covariates in the Analysis of Retrospective Data." *Mathematical Population Studies* 16(2): 105–130. Doi: https://doi.org/10.1080/08898480902790171.

Gilks, W., S. Richardson, and D. Spiegelhalter. 1996. *Markov Chain Monte Carlo in Practice*. Chapman and Hall.

Glesjer, L. 1990. "Improvements of the Naive Approach to Estimation in Nonlinear Errors-in-variables Regression Models." In *Statistical Analysis of Error Measurement Models and Application*, edited by P. Brown and W. Fuller, 99–114. Providence: American Mathematics Society. Doi: https://doi.org/10.1090/conm/112.

Glewwe, P. 2007. "Measurement Error Bias in Estimates of Income and Income Growth Among the Poor: Analytical Results and a Correction Formula." *Economic Development and Cultural Change* 56: 163–189. Doi: https://doi.org/10.1086/520559.

Golub, A., B.D. Johnson, and E. Labouvie. 2000. "On Correcting Biases in Self-reports of Age at First Substance use with Repeated Cross-section Analysis." *Journal of Quantitative Criminology* 16: 45–68. Doi: https://doi.org/10.1023/A:1007573411129.

Gustafson, P. 2003. *Measurement Error and Misclassification in Statistics and Epidemiology*. Boca Raton: Chapman and Hall.

Holt, D., J.W. McDonald, and C.J. Skinner. 2011. "The Effect of Measurement Error on Event History Analysis." In *Measurement Error in Surveys*, edited by P. Biemer, 665–685. New York: John Wiley.

Huttenlocher, J., L. Hedges, and V. Prohaska. 1988. "Hierarchical Organization in Ordered Domains: Estimating the Dates of Events." *Psychological Review* 95: 471–484.

Ibrahim, J.G. and P.W. Laud. 1991. "On Bayesian Analysis of Generalized Linear Models using Jeffreys's Prior." *Journal of the American Statistical Association* 86(416): 981–986. Doi: https://doi.org/10.1037/0033-295X.95.4.471.

Jäckle, A. 2008. Measurement error and data collection methods: Effects on estimates from event history data. Technical report, Institute for Social and Economic Research, ISER. Available at: https://www.iser.essex.ac.uk/research/publications/working-papers/iser/2008-13.pdf (accessed February 2019).

Jeffreys, H. 1946. "An Invariant Form for the Prior Probability in Estimation Problems." *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences* 24: 453–461. Doi: https://doi.org/10.1098/rspa.1946.0056.

Jenkins, S.P. and P. Lynn. 2005. *Improving Survey Measurement of Income and Employment, 2001–2003* (2nd ed.). UK Data Service. Doi: https://doi.org/10.5255/UKDA-SN-5157-1.

Johnson, E.O. and L. Schultz. 2005. "Forward Telescoping Bias in Reported Age of Onset: An Example From Cigarette Smoking." *International Journal of Methods in Psychiatric Research* 14: 119–129. Doi: https://doi.org/10.1002/mpr.2.

Jürges, H. 2007. "Unemployment, Life Satisfaction and Retrospective Error." *Journal of the Royal Statistical Society, Series A* 170(1): 43–61. Doi: https://doi.org/10.1111/j.1467-985X.2006.00441.x.

Kapteyn, A. and J.Y. Ypma. 2007. "Measurement Error and Misclassification: A Comparison of Survey and Administrative Data." *Journal of Labour Economics* 25(3): 513–551. Doi: https://doi.org/10.1086/513298.

Kettunen, J. 1997. "Education and Unemployment Duration." *Economics of Education Review* 16(2): 163–170. Doi: https://doi.org/10.1016/S0272-7757(96)00057-X.

Kreuter, F., G. Mller, and M. Trappman. 2010. "Nonresponse and Measurement Error in Employment Research: Making Use of Administrative Data." *Public Opinion Quarterly* 74(5): 880–906. Doi: https://doi.org/10.1093/poq/nfq060.

Lancaster, T. 1979. "Econometric Methods for the Duration of Unemployment." *Econometrica* 47(4): 939–956. Doi: https://doi.org/10.2307/1914140.

Levine, P. 1993. "CPS Contemporaneous and Retrospective Unemployment Compared." *Monthly Labor Review* 116: 33–39. Available at: https://heinonline.org/HOL/Page?handle=hein.journals/month116&div=89&g_sent=1&casa_token=pTR6IZj22XsAAAAA:xq9wIH0hhVt7hMJgw6ViXuW_gWKx8-EARBvTPW32LcaWEKxYad-v0O53OauyAW25tklO5TD6&collection=journals (accessed February 2019).

Lunn, D.J., A. Thomas, N. Best, and D. Spiegelhalter. 2000. "Winbugs a Bayesian Modelling Framework: Concepts, Structure, and Extensibility." *Statistics and Computing* 10: 325–337. Doi: https://doi.org/10.1023/A:1008929526011.

Manzoni, A., R. Luijkx, and R. Muffels. 2011. "Explaining Differences in Labour Market Transitions between Panel and Life-course Data in West-Germany." *Quality and Quantity* 45: 241–261. Doi: https://doi.org/10.1007/s11135-009-9292-1.

Manzoni, A., J.K. Vermunt, R. Luijkx, and R. Muffels. 2010. "Memory Bias in Retrospectively Collected Employment Careers: A Model-based Approach to Correct for Measurement Error." *Sociological Methodology* 40: 39–73. Doi: https://doi.org/10.1111/j.1467-9531.2010.01230.x.

Mathiowetz, N. and G. Duncan. 1988. "Out of Work, Out of Mind: Response Errors in Retrospective Reports of Unemployment." *Journal of Business and Economic Statistics* 6(2): 221–229. Doi: https://doi.org/10.1080/07350015.1988.10509656.

Messer, K. and L. Natarajan. 2008. "Maximum Likelihood, Multiple Imputation and Regression Calibration for Measurement Error Adjustment." *Statistics in Medicine* 27(30): 6332–6350. Doi: https://doi.org/10.1002/sim.3458.

Morgenstern, R. and N. Barrett. 1974. "The Retrospective Bias in Unemployment Reporting By Sex, Race and Age." *Journal of the American Statistical Association* 69(346): 355–357. Doi: https://doi.org/10.2307/2285657.

Neter, J. and J. Waksberg. 1964. "A Study of Response Errors in Expenditures Data From Household Interviews." *Journal of the American Statistical Association* 59: 18–55. Doi: https://doi.org/10.1080/01621459.1964.10480699.

Neuhaus, J.M. 1999. "Bias and Efficiency Loss Due to Misclassified Responses in Binary Regression." *Biometrika* 86(4): 843–855. Doi: https://doi.org/10.1093/biomet/86.4.843.

Novick, M.R. 1966. "The Axioms and Principal Results of Classical Test Theory." *Journal of Mathematical Psychology* 3: 1–18. Doi: https://doi.org/10.1016/0022-2496(66)90002-2.

Office for National Statistics. Social and Vital Statistics Division. 2006. General Household Survey, 2003–2004. [data collection]. 2nd Edition. UK Data Service. SN: 5150. Available at: http://doi.org/10.5255/UKDA-SN-5150-1.

Office for National Statistics. 2008. Social and Vital Statistics Division. ONS Omnibus Survey, April 2006. [data collection]. UK Data Service. SN: 5997. Available at: http://doi.org/10.5255/UKDA-SN-5997-1.

Paull, G. 2002. Biases in the reporting of labour market dynamics. Technical report, Institute for Fiscal Studies. Doi: https://doi.org/10.1920/wp.ifs.2002.0210.

Pavlopoulos, D. and J.K. Vermunt. 2015. Measuring temporary employment. do survey or register data tell the truth? Technical report, Vrije Universiteit Amsterdam. Available at: https://www150.statcan.gc.ca/n1/pub/12-001-x/2015001/article/14151-eng.htm (accessed February 2019).

Peytchev, A. 2012. "Multiple Imputation for Unit Nonresponse and Measurement Error." *Public Opinion Quarterly* 76(2): 214–237. Doi: https://doi.org/10.1093/poq/nfr065.

Pickles, A., K. Pickering, E. Simonoff, J. Silberg, J. Meyer, and H. Maes. 1998. "Genetic Clocks and Soft Events: A Twin Model for Pubertal Development and Other Recalled Sequences of Developmental Milestones, Transitions, or Ages At Onset." *Behavior Genetics* 28: 243–253. Doi: https://doi.org/10.1023/A:102161522.

Pickles, A., K. Pickering, and C. Taylor. 1996. "Reconciling Recalled Dates of Developmental Milestones, Events and Transitions: A Mixed Generalized Linear Model with Random Mean and Variance Functions." *Journal of the Royal Statistical Society. Series* A1: 225–234. Doi: https://doi.org/10.2307/2983170.

Pina-Sánchez, J. 2016. "Adjustment of Recall Errors in Duration Data using Simex." *Advances in Methodology and Statistics* 12(1): 27–58. Available at: http://ibmi.mf.uni-lj.si/mz/2016/no-1/p3.pdf (accessed February 2019).

Pina-Sánchez, J., J. Koskinen, and I. Plewis. 2013. "Implications of Retrospective Measurement Error in Event History Analysis." *Metodología de Encuestas* 15: 5–25. Available at: http://casus.usal.es/pkp/index.php/MdE/article/view/1032/pdf_2 (accessed February 2019).

Pina-Sánchez, J., J. Koskinen, and I. Plewis. 2014. "Measurement Error in Retrospective Work Histories." *Survey Research Methods* 8: 43–55. Doi: https://doi.org/10.18148/srm/2014.v8i1.5144.

Plummer, M., N. Best, K. Cowles, and K. Vines. 2006. "Coda: Convergence Diagnosis and Output Analysis." *R News* 6: 7–11. Available at: http://oro.open.ac.uk/22547/ (accessed February 2019).

Poterba, J. and L. Summers. 1995. "Unemployment Benefits and Labor Market Transitions: A Multinomial Logit Model with Errors in Classification." *Review of Economics and Statistics* 77: 207–216. Doi: https://doi.org/10.2307/2109860.

Poterba, J.M. and L.H. Summers. 1984. "Response Variation in the CPS: Caveats for the Unemployment Analyst." *Monthly Labor Review* 107: 37–43. Available at: https://stats.bls.gov/opub/mlr/1984/03/rpt1full.pdf (accessed February 2019).

Pyy-Martikainen, M. and U. Rendtel. 2009. "Measurement Errors in Retrospective Reports of Event Histories. A Validation Study with Finnish Register Data." *Survey Research Methods* 3(3): 139–155. Doi: https://doi.org/10.1002/sim.4780121806.

Richardson, S. and W.R. Gilks. 1993. "Conditional Independence Models for Epidemiological Studies with Covariate Measurement Error." *Statistics in Medicine* 12(18): 1703–1722. Doi: https://doi.org/10.1002/sim.4780121806.

Rubin, D.B. 1996. "Multiple Imputation After 18+ Years." *Journal of the American statistical Association* 91(434): 473–489. Doi: https://doi.org/10.2307/2291635.

Rubin, D.C. 1987. *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley and Sons.

Rubin, D.C. and A.D. Baddeley. 1989. "Telescoping is Not Time Compression: A Model." *Memory & Cognition* 17: 653–661. Doi: https://doi.org/10.3758/BF03202626.

Shiffrin, R.M. and J.R. Cook. 1978. "Short-term Forgetting of Item and Order Information." *Journal of Verbal Learning and Verbal Behavior* 17(2): 189–218. Doi: https://doi.org/10.1016/S0022-5371(78)90146-9.

Skinner, C. and K. Humphreys. 1999. "Weibull Regression for Lifetimes Measured with Error." *Lifetime Data Analysis* 5: 23–37. Doi: https://doi.org/10.1023/A:1009674915476.

Solga, H. 2001. "Longitudinal Survey and the Study of Occupational Mobility: Panel and Retrospective Design in Comparison." *Quality and Quantity* 35: 291–309. Doi: https://doi.org/10.1023/A:1010387414959.

Veronesi, G., M.M. Ferrario, and L.E. Chambless. 2011. "Comparing Measurement Error Correction Methods for Rate-of-change Exposure Variables in Survival Analysis." *Statistical Methods in Medical Research* 22(6): 583–597. Doi: https://doi.org/10.1177/0962280210395742.

Wang, C.Y., L. Hsu, R.L. Feng, and Z.D. Prentice. 1997. "Regression Calibration in Failure Time Regression." *Biometrics* 53: 131–145. Doi: https://doi.org/10.2307/2533103.

Received July 2017
Revised July 2018
Accepted August 2018