MEMORY BIAS IN RETROSPECTIVELY COLLECTED EMPLOYMENT CAREERS: A MODEL-BASED APPROACH TO CORRECT FOR MEASUREMENT ERROR

Anna Manzoni*
Jeroen K. Vermunt†
Ruud Luijkx†
Ruud Muffels†

Event history data constitute a valuable source to analyze life courses, although the reliance of such data on autobiographical memory raises many concerns over their reliability. In this paper, we use Swedish survey data to investigate bias in retrospective reports of employment biographies, applying a novel model-based latent Markov method. A descriptive comparison of the biographies as reconstructed by the same respondents at two interviews carried out about 10 years apart reveals that careers appear simpler and less heterogeneous and have fewer elements and episodes.

This research was partially funded by NWO project 400-04-552 at Tilburg University. We are grateful to the Swedish Institute for Social Research for providing the data and advice on their use. We thank Cees Elzinga, Richard Breen, the Editor, and the reviewers for useful comments and suggestions. Direct correspondence to Anna Manzoni, Center for Research on Inequalities and the Life Course, Yale University, 140 Prospect St., New Haven, CT 06511; e-mail: anna.manzoni@gmail.com.*

*Center for Research on Inequalities and the Life Course (CIQLE), Yale University
†Tilburg University, The Netherlands
when reported at a point long after their occurrence, with a particularly high underreport of unemployment. Using matching techniques, the dissimilarity between the two reconstructions turns out to be unaffected by respondents’ sociodemographic characteristics but particularly affected by the occurrence of unemployment spells and career complexity. Using latent Markov models, we assume correlated errors across occasions to determine the measurement error and to obtain a more reliable estimate of the (true) latent state occupied at a particular time point. The results confirm that (correlated) measurement errors lead to simplification and conventionalism. Career complexity makes recall particularly problematic at longer recall distances, whereas unemployment underreporting also happens very close to the interview. However, only a small fraction of respondents make consistent errors over time, while the great majority makes no errors at all.

1. INTRODUCTION

The study of the patterns and determinants of employment careers, their heterogeneity across sociodemographic groups, their changes over time and across cohorts, as well as the impact of labor market and welfare contexts requires the use of longitudinal data sets (Mayer and Huinink 1990; Solga 2001). One possible way to collect employment career data is through prospective panel studies in which respondents are followed during a certain time period and in which the same set of questions about the respondent’s current employment status is asked at each of the panel waves. An alternative is to use a retrospective study design, in which respondents report their complete previous employment career in a single survey (Taris 2000). Retrospective data collection designs are very valuable for studying change over time in that they allow for tracing, in a relatively fast and in expensive way, entire individual careers of respondents belonging to different cohorts that cover long periods in historical time. Such retrospective surveys usually collect information in the form of event histories, where respondents are asked to report the full sequence of episodes that they experienced up to the moment of the interview (Scott and Alwin 1998).

The quality of life-history data obtained via retrospective surveys depends on whether respondents are able and willing to recall and date each of the relevant episodes (or, equivalently, all relevant transitions). That is, retrospective surveys rely on the respondents’ autobiographical
memories (Rubin and Baddeley 1989), in the course of which information may be voluntarily or involuntarily distorted. Completeness and accuracy might be severely at risk when using memory-based reports, especially when the recall covers a longer period (Reimer 2004). As a consequence, reported life-histories may contain serious errors, which may invalidate the substantive findings obtained when analyzing such data. Insight into the quality of reconstructed histories is therefore of great value for researchers using such data sets, since it will prevent them from drawing wrong inferences—for example, about the alleged rising turbulence on labor markets associated with globalization forces. Moreover, once the type and amount of measurement error is known, it might allow substantive researchers to correct for it during the actual data analysis.

In this paper, we examine the quality of reconstructed employment biographies and how that might depend on factors such as the time lag between the events of interest and the interview, using a unique Swedish data set containing two reports of the same period. Descriptive methods are used to analyze the two reports, and matching techniques are used for quantifying similarities between event sequences (Brzinsky-Fay, Kohler, and Luniak 2006; Elzinga 2007). Latent Markov models are used as well (Magidson, Vermunt, and Tran 2009; Van de Pol and Langeheine 1990). The main contribution of the paper is that it demonstrates how to use the latter model-based approach to determine measurement error in life-history data, which, as will be shown, provides an improvement over purely descriptive approaches and matching techniques. The latent Markov model we use expands on the models proposed among others by Van de Pol and Langeheine (1990) in that it relaxes the assumption that errors are independent across time points. The remainder of this paper is organized as follows. We first explain the data used for our analyses. Subsequently, possible sources of errors in the survey and in retrospective employment history data in general are discussed. We then report on the data analyses, followed by discussion and conclusions.

2. THE SWEDISH LEVEL OF LIVING SURVEY

A major source for studying labor market dynamics in Sweden is the Swedish Level of Living Survey (LNU), in which retrospective
information on work histories is collected. Conducted by researchers at the Swedish Institute for Social Research (SOFI) at Stockholm University, the LNU is one of the longest running longitudinal social science surveys. It was conducted for the first time in 1968 and then repeated in 1974, 1981, 1991, and 2000. Using a panel approach, these surveys were carried out as face-to-face interviews with national probability samples of individuals between the ages of 18 and 75 (15 to 75 until 1981). The methodology of the survey has gradually developed and in the 1991 wave a retrospective section was added to the questionnaire, recording information on cohabitation, children, education, and economic activities with a precision of one month duration (Jonsson and Mills 2001). The 2000 wave contained the same retrospective section.

In the working life biography section, data were collected in the form of event histories, where respondents had to recall and date one by one the full sequence of subsequent episodes in their life-course, in a temporal order, starting with the first job that lasted at least six months. For each episode, respondents had to provide information on the labor force status and end date of the episode,\(^1\) with monthly precision in dates.\(^2\) Consequently, a respondent’s employment history consists of a continuous sequence of employment episodes interrupted by episodes of unemployment and labor market inactivity; that is, respondents move chronologically through mutually exclusive states. For the employment spells, information was gathered on occupation, industry, firm size, and sector (both in 1991 and 2000), as well as on part-time versus full-time work, permanent versus temporary employment, and whether the temporary employment forms part of a labor market program (in 2000 only).

New respondents in 2000 as well as those who participated in 1991 but did not answer the retrospective part had to report the whole employment history (from the first job lasting at least 6 months),

\(^1\)The design assumes no gaps, since no information on start was collected, and parallel activities were not registered.

\(^2\)In 1991 this information was collected only for respondents born between 1925 and 1965. Furthermore, those who claimed to have had 15 jobs or more were diverted in the questionnaire to a shorter section on occupations. We wanted to focus on employment (and not job) histories, and we consequently decided not to consider those who completed only the occupational history in our analysis, which means that we excluded respondents who declared that they had more than 15 jobs.
whereas for those who participated in the 1991 survey, the data collection in 2000 starts asking the activity in January 1990. This allows us to reconstruct the whole career history of each respondent, until the last interview, offering a rich source of information to analyze the entire career patterns of several birth cohorts over a long time period.

The Swedish Level of Living Survey provides a unique opportunity to study memory bias in retrospective reports. As previously mentioned, respondents who participated in the 1991 survey were reinterviewed in 2000 (when possible) and asked to report their employment biography starting from the activity they held in January 1990. As a side effect, respondents provided a second report on the period from January 1990 to the date of the interview in 1991. More specifically, for respondents who participated in both waves, we have the information from the 1991 survey starting in January 1990 up to the interview date in 1991 and the information from the 2000 survey starting in January 1990 up to the interview date in 2000. This yields overlapping information for a certain number of months during the years 1990 and 1991. It should be noted that for the 1991 interview the overlap concerns a recall on what happened in the work life in the past year whereas in the 2000 interview it concerns a reconstruction of about 10 years ago. Figure 1 provides a graphical illustration of the information collected in the two interviews and how these overlap.

Having employment biographies from two retrospective interviews with an overlapping period may generate inconsistencies between reports. Such inconsistencies could be attributed to inconsistent

---

**TABLE 1.** Representation of the data structure.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Restrospective Distance</td>
<td>6</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

3Although the period reconstructed is relatively short, the levels of mobility on the Swedish labor market allow us to obtain a sufficiently large number of transitions for which memory error can be observed (Blomskog 1997). The labor market is characterized by intermediate levels of job mobility, predominantly upward in direction (DiPrete 2002), weak occupational boundaries and organization-based labor market boundaries causing jobs to be less stable compared to other countries, and firms to resort more quickly to layoffs—even of experienced workers—as their way of adjustment (Björklund and Holmlund 1987; Büchtemann 1993; Grubb and Wells 1993; OECD 1994).
autobiographical reconstructions and to the greater recall difficulties in the second interview since about 10 years had elapsed in 2000 for the recall of events in 1990/1991. This might, on the one hand, be somewhat problematic because merging the histories to obtain the complete history should deal in some way with the inconsistencies between the two sources. On the other hand, it also offers a unique opportunity to study the quality of retrospective reports, allowing us to investigate whether and how recall patterns vary according to the length of the recall period.

3. POSSIBLE SOURCES OF MEASUREMENT ERROR IN RETROSPECTIVE REPORTS

Recall bias might endanger the reliability of retrospective data, due to involuntary memory lapses but also to a voluntary modification of events (Powers et al. 1978). Respondents’ lack of memory accuracy may consequently produce an artificial sequence of episodes, which is different from the real life-course sequence. The subjective reconstruction by the respondent might be affected by (and adapted to) normative considerations about the logic of a life-course and the individual’s self perception of that at the time of recall. Respondents may completely forget episodes but also erroneously add some, artificially increasing the number of transitions. Memory’s tendency toward reduction in complexity is usually expected to lead to transitions being more often forgotten than added. Furthermore, respondents may report their employment biographies differently when asked at different moments in their lives and at different distances from the occurrence of the events (Babbie 1973; Sudman and Bradburn 1973).

A variety of reasons might lead respondents to intentionally or unintentionally give an inaccurate account of their work histories, and such reasons might act in opposite directions (i.e., overreport or under-report of certain states) and follow different mechanisms at different recall distances. Overall, we could think of different reasons leading to misreports in both the directions at both occasions, and in no systematic way. For this reason, we find it reasonable to assume that intentional misreport does not lead to errors in one specific direction.
If the same respondents are re-interviewed and no changes have occurred in the survey format, which is the case in our data, the sample as well as the questionnaire format cannot be an explanation for the differences observed between the two interviews. What is, however, different in the two sources is the length of the recall period (about 1 year as opposed to about 10 years). Given the assumption that memory tends to decay over time, recall difficulties were much larger at the second interview, in which respondents reported on employment histories that occurred about 10 years before. Therefore, it seems reasonable to be more confident about the 1991 report. However, both interviews may contain errors, and it is thus not correct to treat the 1991 report as the gold standard without errors.

4. ANALYSIS OF THE LEVEL OF LIVING DATA SET

All our analyses are based on the overlapping period of retrospective work histories described in Section 2. The analyses are based on the reports of 1973 respondents, 47% of whom are men and 53% women. For these respondents, the 1991 and 2000 measurements contained between 3 and 22 months of overlap, on average 15 months, during the period starting in January 1990 (the earliest possible date) and November 1991 (the latest recorded interview date in 1991). The average length of the recall distance is 8 months for the 1991 survey and almost 10 years for the 2000 survey. All our analyses are based on a distinction of four employment states: employed (E), self-employed (S), not employed (N), or unemployed (U).

4Note that 83% of the months reported in 1991 are part of the current spell at the interview date. This makes it more difficult to skip reporting that very spell and may also be seen as a reason for more accuracy in the 1991 report. Looking at the posterior estimates, we indeed find that months that are not part of the current spell at the interview date in 1991 contain more errors than months that are part of the spell held at the interview date (23% against 6% of error at the person-month level).

5We checked whether attrition affects the representativeness of our sample, but this does not seem to be the case: Our sample composition does not differ from the original sample in terms of age, gender, education, and career complexity.

6A few respondents reported their activity only from a date later than January 1990. Consequently, for these respondents we have information on fewer than 13 months.
TABLE 1

Number of Reported Episodes and Transitions, by Interview

<table>
<thead>
<tr>
<th></th>
<th>1991</th>
<th>2000</th>
<th>Reduction in 2000 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Episodes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Employment</td>
<td>1732</td>
<td>1652</td>
<td>4.62</td>
</tr>
<tr>
<td>Self-employment</td>
<td>155</td>
<td>127</td>
<td>18.06</td>
</tr>
<tr>
<td>Nonemployment</td>
<td>534</td>
<td>474</td>
<td>11.24</td>
</tr>
<tr>
<td>Unemployment</td>
<td>96</td>
<td>55</td>
<td>42.71</td>
</tr>
<tr>
<td>Total</td>
<td>2517</td>
<td>2308</td>
<td>8.30</td>
</tr>
<tr>
<td>Transitions</td>
<td>544</td>
<td>335</td>
<td>38.42</td>
</tr>
</tbody>
</table>

In Section 4.1, we provide a descriptive analysis of the two event sequences. In Section 4.2, we quantify the difference between the reported individual career sequences at the two interviews using matching techniques, and link the dissimilarity measures to the career and personal characteristics. In Section 4.3, we move to a model-based approach using latent Markov (LM) models.

4.1. Description and Comparison of the 1991 and 2000 Sequences

Table 1 provides information on the number of reported episodes per state for the two interviews. An episode is defined as a period in which a person maintains the same employment status (either E, S, N, or U), and a change in the employment status defines the end of an episode. It can be seen that the number of episodes reported is much smaller in the 2000 interview than in the 1991 interview. The difference is largest (43%) for unemployment episodes. However, rather than looking at episodes, it is more informative to look at descriptive characteristics of the full career trajectories—that is, at the full sequence of episodes in the biography of each respondent during the period of interest.

7Huber and Schmucker (2009) recently adopted the same approach to identify types and determinants of inconsistencies between administrative and survey data on employment cycles.

8Recently, various new approaches have between proposed for the analysis of sequences such as the graphical tool for sequence exploration by Piccarreta and Lior (2010). Here, we do not use this new tool because it is not suited for our purpose—namely, sequence comparison.
Table 2 provides such a descriptive comparison of the sequences in the two reports. Employment careers appear less heterogeneous according to the report in the second interview (2000) than in the first (1991): Fewer different elements (i.e., states) and fewer episodes seem to be reported on average in a career sequence when the reconstruction took place at a later point after the occurrence of the events (in 2000).

Table 3 presents a two-way cross-tabulation of the type of sequence reported at the two occasions. The least complex sequences consisting of one episode appear to be the ones most likely to be in agreement across the two interviews, showing also the similarity to be largest for employment (E). Given that the sequences are ordered from the least to the most complex, the larger frequencies under the main diagonal indicate that the reported sequences in the 2000 interview tend to be simpler than those in the 1991 interview. The 2000 reports contain fewer changes and simpler careers.

These findings are particularly relevant given the current debate about the alleged rising volatility on the labor market. Our results suggest that the earlier part of the career history tends to be simplified (reporting fewer transitions in earlier life than there were in reality) in retrospective reports, which relaxes the conclusions with respect to rising turbulence, since part of it is simply due to measurement error.10

9The order in which sequences are presented is based on the number of different episodes in the trajectories and their frequency.

10If increasing length of the recall period leads to more simplified careers, career volatility is underestimated, especially for the transitions farthest away in the career history and less for the most recent, those closer to the interview, which would overestimate the trend toward increasing volatility over time.
<table>
<thead>
<tr>
<th>Sequence in 1991</th>
<th>E</th>
<th>N</th>
<th>S</th>
<th>EN</th>
<th>NE</th>
<th>EU</th>
<th>ES</th>
<th>ENE</th>
<th>EUE</th>
<th>NEN</th>
<th>Other</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>E</td>
<td>95.5</td>
<td>0.8</td>
<td>0.5</td>
<td>1.2</td>
<td>0.6</td>
<td>0.3</td>
<td>0.2</td>
<td>0.3</td>
<td>0.2</td>
<td></td>
<td>—</td>
<td>0.4</td>
</tr>
<tr>
<td>N</td>
<td>5.3</td>
<td>85.6</td>
<td>—</td>
<td>5.3</td>
<td>2.3</td>
<td>—</td>
<td>2.7</td>
<td>—</td>
<td>—</td>
<td></td>
<td>—</td>
<td>1.5</td>
</tr>
<tr>
<td>S</td>
<td>5.4</td>
<td>9.0</td>
<td>79.3</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td></td>
<td>—</td>
<td>3.6</td>
</tr>
<tr>
<td>EN</td>
<td>14.2</td>
<td>18.1</td>
<td>—</td>
<td>59.1</td>
<td>1.6</td>
<td>—</td>
<td>—</td>
<td>1.6</td>
<td>—</td>
<td></td>
<td>—</td>
<td>3.1</td>
</tr>
<tr>
<td>NE</td>
<td>29.3</td>
<td>7.3</td>
<td>1.2</td>
<td>1.2</td>
<td>48.8</td>
<td>—</td>
<td>—</td>
<td>4.9</td>
<td>—</td>
<td></td>
<td>3.7</td>
<td>3.6</td>
</tr>
<tr>
<td>EU</td>
<td>52.4</td>
<td>—</td>
<td>—</td>
<td>9.5</td>
<td>—</td>
<td>19.1</td>
<td>—</td>
<td>4.8</td>
<td>—</td>
<td></td>
<td>—</td>
<td>14.2</td>
</tr>
<tr>
<td>ES</td>
<td>52.9</td>
<td>5.9</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>5.9</td>
<td>35.9</td>
<td>—</td>
<td>—</td>
<td></td>
<td>—</td>
<td>17</td>
</tr>
<tr>
<td>ENE</td>
<td>54.6</td>
<td>—</td>
<td>3.0</td>
<td>3.0</td>
<td>6.1</td>
<td>3.0</td>
<td>—</td>
<td>24.2</td>
<td>3.0</td>
<td>—</td>
<td>—</td>
<td>3.1</td>
</tr>
<tr>
<td>EUE</td>
<td>64.3</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>14.3</td>
<td>—</td>
<td>—</td>
<td>7.1</td>
<td>—</td>
<td></td>
<td>14.3</td>
<td>14</td>
</tr>
<tr>
<td>NEN</td>
<td>—</td>
<td>35.7</td>
<td>—</td>
<td>35.7</td>
<td>14.3</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td></td>
<td>14.3</td>
<td>14</td>
</tr>
<tr>
<td>Other</td>
<td>5.3</td>
<td>85.64</td>
<td>—</td>
<td>5.3</td>
<td>2.3</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td></td>
<td>—</td>
<td>1.5</td>
</tr>
<tr>
<td>N</td>
<td>1356</td>
<td>230</td>
<td>103</td>
<td>111</td>
<td>69</td>
<td>11</td>
<td>11</td>
<td>21</td>
<td>6</td>
<td>9</td>
<td></td>
<td>46</td>
</tr>
</tbody>
</table>
To further address the issue of career instability, we use a recently developed measure of (career) “complexity” (Elzinga forthcoming), also referred to as “turbulence” (Elzinga 2007; Elzinga and Liefbroer 2007), which quantifies the number of employment changes, the predictability of the order of jobs, and the variability of the durations spent in different employment states. Elzinga’s measure takes into account the order of career states as well as their duration variation in different career states.\(^{11}\)

Table 4 reports the mean turbulence for the two reports,\(^{12}\) as well as their bootstrap\(^{13}\) confidence intervals (Efron and Tibshirani 1993). As can be seen, the average turbulence is larger for the 1991 report than for the 2000, and the difference between these two means is significant since the bootstrap confidence intervals show no overlap. This indicates again that the 2000 report may underestimate the actual career complexity. This finding is relevant because it shows that the

\[ T(x, t) = \log_2(\phi(x) \frac{s^2_{t, \text{max}} + 1}{s^2_t}) \]

with \( \phi(x) \) denoting the number of distinct subsequences, \( s^2_t \) denoting the variance of state durations, and \( s^2_{t, \text{max}} \) the maximum of that variance given the total duration of the sequence. \( T(x, t) \) is therefore a sequence property quantified so that the sequencing of the states and the variance of their durations are taken into account; it is not sensitive to the specific time scale used, and it increases with the decreasing variance of the durations of the states (Elzinga and Liefbroer 2007).

\(^{11}\)As a measure of turbulence of sequence \( x \) of duration \( t \), Elzinga (2007:33)

\(^{12}\)Whereas in Table 4 we present the results based on this turbulence measure, we also computed it without accounting for duration (Elzinga 2007:33). Although the absolute values of the measure differ, the substantive results on differences in the two reports remain the same.

\(^{13}\)Since the average T is just a sample-average that has a t-distribution under very mild conditions, it is not necessary to bootstrap. However, the bootstrapped confidence intervals provide us with a robust test.
length of the recall period may affect the observed turbulence, yielding biased results on changes in employment statuses over the life-course.

4.2. Quantifying Sequence Similarities Using Optimal Matching Techniques

So far, we looked at aggregated sequence patterns and aggregated differences between the 1991 and 2000 sequences. In the next step, we want to quantify the amount of measurement error in the sequences at the level of the individual. For the moment, we assume that the dissimilarity between the two sequences is a good indication of the amount of measurement error they contain. The obtained dissimilarity measure can subsequently be used as the dependent variable in a regression analysis to investigate whether differences are associated with respondent and career characteristics.

Various methods have been developed for quantifying the dissimilarity between pairs of sequences, of which the most widely used and best known in the social sciences is optimal matching (OM) (Abbott 1995; Abbott and Forrest 1986). This method compares sequences using the Levenshtein distance, a metric based on the minimal cost (i.e., the most efficient set of operations) to transform one string into another using a series of “elementary operations” of insertion, deletion, and substitution (Abbott 1995; Abbott and Forrest 1986; Abbott and Tsay 2000; Sankoff and Kruskal 1983).14

In OM, the costs associated with each of the transformation operations have to be defined, implying assumptions about the relative cost of substituting, inserting, and deleting. The specification of the substitution costs between pairs of elements (which in our application are the four labor market states) may be especially complicated because it involves quantifying the dissimilarity between the elements (states).15

14The only difference between the Needleman-Wunsch (1970) algorithm and Levenshtein’s is that in the former algorithm, the operations are weighed by a cost-function of which the arguments are the characters involved in the edit. Because a thorough explanation of the algorithm goes beyond the scope of this paper, interested readers are referred to Sankoff and Kruskal (1983).
15This step in OM is always arbitrary to a certain extent, and in the social sciences a theory is not always sufficiently precise to inform us about the
We used the default settings of the Stata OM routine that we applied in our analysis, which implied using a substitution cost matrix treating all the states as equally dissimilar and assuming substitution costs to be twice as large as the cost associated with inserting or deleting elements.

We first employed optimal matching to quantify the dissimilarity between the pair of sequences for each respondent and subsequently used regression analysis to estimate the effect of respondent and career characteristics on the dissimilarity. The first column of Table 5 contains the results of the regression analysis, where the dependent variable is the dissimilarity measure; we include, as explanatory variables, gender, birth year, episodes of unemployment in the career (as reconstructed in 1991), number of episodes the sequence consisted of (according to the 1991 reconstruction), and a constant standardized to the number of months that are observed calculating the degree to which a respondent’s number

<table>
<thead>
<tr>
<th></th>
<th>OM Distancea</th>
<th>Elzinga’s Distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>−0.001</td>
<td>0.006</td>
</tr>
<tr>
<td>Birth year</td>
<td>−0.002</td>
<td>−0.001</td>
</tr>
<tr>
<td>Episodes of unemployment</td>
<td>0.644**</td>
<td>0.473**</td>
</tr>
<tr>
<td>Number of episodes</td>
<td>0.575**</td>
<td>0.811**</td>
</tr>
<tr>
<td>Constant</td>
<td>2.943</td>
<td>1.865</td>
</tr>
<tr>
<td>N</td>
<td>1973</td>
<td>1973</td>
</tr>
<tr>
<td>R squared</td>
<td>0.20</td>
<td>0.32</td>
</tr>
</tbody>
</table>

*Default subcost matrix.

**P < 0.01.

appropriate cost specification assumptions, which may cast concerns about the lack of a theoretical basis to determine the substitution costs matrix that may justify the analysis. Theoretical considerations about the similarity between elements may guide the definition and differentiation of substitution costs. Alternatively, it has been proposed that they are fully derived from the data themselves based on the frequency of transitions from one state to another, with lower costs for high transition frequency, based on the intuition that two statuses are less different when there are more transitions between them. We do not use such a specification of the substitution cost matrix because we consider, for example, employment more similar to self-employment than to nonemployment, although we observe fewer transitions in the first case.
of episodes deviates from the average number of episodes reported by the respondent observed for the same period of time.\textsuperscript{16}

As can be seen from the results in Table 5, gender and age\textsuperscript{17} are not significantly related to sequence dissimilarity. The number of episodes in the sequence and the occurrence of episodes of unemployment, instead, have strong positive effects on sequence dissimilarity, indicating that recall accuracy is lower for respondents with more transitions and with episodes of unemployment. The significance of both terms indicates that each has an effect controlling for the other, which means that episodes of unemployment decrease recall accuracy independent of the effect of unemployment on raising the career complexity. Trusting the 1991 report more, these results also suggest that in such circumstances recall is more problematic when the recall period is longer.\textsuperscript{18}

We acknowledge that OM has been subject to criticism (Wu 2000), especially for the way it (fails to) handle duration and because it is not able to account for the order of events and the direction of change over time. Some aspects in OM might be particularly problematic. The substitution cost matrix in OM has to be symmetric, which in our case means that reporting employment when actually being unemployed is implicitly considered to be equally wrong to reporting unemployment when actually being employed. Moreover, OM does not handle right censoring, and it neglects nonlinear time dependencies and lacks concern with covariates (Halpin 2003).

\textsuperscript{16}We are not interested here in the deeper substantive definition of career complexity, but simply in assessing whether increasing complexity of the career leads to more mistakes in reporting. Therefore, we believe that it is more appropriate to use a measure indicating the (standardized) number of episodes instead of Elzinga’s turbulence, the interpretation of which might not be so straightforward to the reader who would like to understand the meaning of an increase of one unit in the value of the turbulence measure. However, estimations using Elzinga’s measure lead to similar results.

\textsuperscript{17}The time of the interview being fixed, birth cohort also indicates the age at interview, where later cohorts—i.e., increasing birth year—are younger at the time of interview.

\textsuperscript{18}We also included education in the model (the results of which are available from the first author upon request), but its effect was found insignificant. Due to a slight loss of power resulting from a few cases with missing information on education, we chose to present the result of the model without education. However, the effects of the included variables remained very similar.
Several alternatives to the standard OM have been proposed.\textsuperscript{19} Yet, our results turned out to be robust not only to different specifications of the substitution cost matrix but also to the specific technique used to compare the two sets of sequences.\textsuperscript{20} To demonstrate this robustness, the second column of Table 5 provides the results of a regression analysis using a recently developed distance measure by Elzinga (2005), which is based on a nonalignment technique that accounts for the order of events within sequences and avoids arbitrary decisions on transformation costs since it does not use transformation operations.\textsuperscript{21}

The algorithmic matching approaches differ from formal statistical models in that they do not require specifying the process that generated the data. As a result, they can be used only for description and exploration, not for prediction and explanation. Though stochastic variants for testing hypotheses may be developed in the future, there is still a long way to get there (Biemann et al. 2009). In contrast, most sequences of interest develop stochastically over time, where, for example, earlier events affect the probability of occurrence of later events (path dependency). That is, for most sequences we will be more interested in the processes that created them than their actual occurrence over time (Halpin 2003:9). Next, we describe a model-based approach that allows specification of the underlying process that generated the observed sequence data.

4.3. \textit{Modeling Sequences Using Latent Markov Models}

To overcome several of the shortcomings of the algorithmic approach presented above, we propose using a model-based approach to

\textsuperscript{19}These include OMAV, an optimal matching technique adjusting for duration; \textit{Degenne}, a method focusing on comparing vectors of cumulated duration at each time point; \textit{Hamming} distance, which compares sequences element by element and equals the number of positions in which two equally long sequences have different characters; \textit{Elzinga}'s technique, which calculates the similarity between the sequences based on identical subsequences.

\textsuperscript{20}The results of such analyses are not shown here but are available from the first author upon request.

\textsuperscript{21}This technique is based on a different metric and handles durations in a more consistent way: It calculates the similarity between the sequences based on identical subsequences. Distance is defined as the number of noncommon subsequences, weighted by their occurrence frequency in either of the sequences and their duration. It has advantages concerning duration and the importance of the order of nonadjacent states. The algorithm consists of efficiently counting the number of common $n$-tuples in pairs of sequences.
investigate the recall bias problem. The two key requirements are that
the stochastic model for modeling the two sets of reports should be
suited (1) for modeling event sequences or discrete-time longitudinal
data on transitions across a finite number of states and (2) for dealing
with measurement error in the recorded states, which in our application
also involves combining the possibly inconsistent information in the
1991 and the 2000 reports of the month-specific labor market status.

A suitable stochastic model is the Markov model, which yields
estimates for the probability of making a transition between each pair
of states of the outcome variable of interest. Although a simple Markov
model solves the problem associated with the OM approach that adjac-
ent time points are treated as somewhat disconnected, it does not
tackle the measurement error issue discussed above. However, there is
a more extended variant of the Markov model called the latent (or hid-
den) Markov model that can be used for this purpose. Here, “latent”
or “hidden” refers to the fact that a person’s true state at time point
\( t \) is unknown, and it should thus be treated as a latent variable. In a
latent Markov model, the true states are the categories (classes) of a
dynamic latent (unobservable) variable. The latent states are connected
to the observed states using a model similar to a latent class model,
which defines the measurement error component of the latent Markov
model.

There are various examples of applications of latent Markov
models for the analysis of discrete-time longitudinal data subject to mea-
surement error (Collins and Wugalter 1992; Hagenaars 1990; Poulsen
1982; Van de Pol and Langeheine 1990; Breen and Moisio 2004;
Vermunt, Tran, and Magidson 2008; Magidson et al. 2009). In our
application, we use the latent Markov model to estimate the error in the
recorded employment status at the two interviews, which we believe will
yield more interesting information on the pattern and amount of recall
bias in the two retrospective reports than was obtained with the de-
scriptive and algorithmic approaches. It should be noted that the latent
Markov model not only yields estimates of transition and measurement
error probabilities but it can also be used to obtain (posterior mode)
estimates of the true state occupied at each time point. Hence, whereas
in OM we could only compare the distance of the 2000 report to the
1991 report, which was in fact used as a kind of perfect measure, the
latent Markov approach allows us to compare each of these two reports
with the estimated true sequences, and thus to relax the assumption that
the information from the 1991 interview is free from errors. Typically, latent or hidden Markov models are applied with a single report per respondent per time point. However, our data consist of two (imperfect) reports of a respondent’s employment status per time point (per person-month). As explained in more detail below, this makes it possible to relax certain key assumptions of the basic latent Markov model that are unrealistic in our application.

We denote the 1991 and 2000 report of the employment status of person \( i \) at month \( t \) by \( y_{1991}^{i t} \) and \( y_{2000}^{i t} \), respectively. Note that \( y_{1991}^{i t} \) and \( y_{2000}^{i t} \) are categorical variables with \( M = 4 \) categories (1 = employed (E), 2 = self-employed (S), 3 = not employed (N), 4 = unemployed (U)); that is, \( 1 \leq y_{it} \leq M \). The total number of time points (months) for which person \( i \) provides information is \( T_i + 1 \), with \( 0 \leq t \leq T_i \).\(^{22}\) The observed data for a particular respondent consist of two response vectors of length \( T_i + 1 \) denoted by \( y_{1991}^{i t} \) and \( y_{2000}^{i t} \), respectively. The probability of having the observed set of responses is denoted by \( P(y_{1991}^{i t}, y_{2000}^{i t}) \). Note that \( P(y_{1991}^{i t}, y_{2000}^{i t}) \) is the probability for which we are going to define a statistical model.

In addition to the observed responses, the latent Markov model for \( P(y_{1991}^{i t}, y_{2000}^{i t}) \) contains \( T_i + 1 \) discrete latent variables, each having \( K \) categories or latent states. We denote the true (latent) state at time point \( t \) by \( x_t \), where \( 1 \leq x_t \leq K \). In the current study, \( K = M = 4 \); that is, the number of latent labor force states is assumed to be equal to the number of observed labor force states. The latent Markov model has the following form:

\[
P(y_{1991}^{i t}, y_{2000}^{i t}) = \sum_{x_0 = 1}^{K} \sum_{x_1 = 1}^{K} \ldots \sum_{x_{T_i} = 1}^{K} \prod_{t=1}^{T_i} P(x_{0t}) \prod_{t=1}^{T_i} P(x_{tt} | x_{(t-1)}) \times \prod_{t=0}^{T_i} P(y_{1991}^{i t} | x_{tt}) \prod_{t=0}^{T_i} P(y_{2000}^{i t} | x_{tt}). \quad (1)
\]

The unknown model probabilities to be estimated are the initial latent state probabilities \( P(x_0 = s) \), the latent transition probabilities \( P(x_t = r | x_{t-1} = s) \), and the classification error probabilities \( P(y_{1991}^{i t} =

\(^{22}\)The number of months for which we have employment information varies across respondents, which is why we use the index \( i \).
\[ \ell | x_t = s ) \text{ and } P(y_{it}^{2000} = \ell | x_t = s ), \text{ where } s \text{ and } r \text{ refer to a particular latent state and } \ell \text{ to a particular observed state. In the Latent GOLD software (Vermunt and Magidson 2008) that we used to estimate our models, the model probabilities are parameterized using the following multinomial logistic equations:} \]

\[ P(x_0 = s) = \frac{\exp(\alpha_s)}{\sum_{k=1}^{K} \exp(\alpha_k)} \]  

\[ P(x_t = r | x_{t-1} = s) = \frac{\exp(\gamma_{rs})}{\sum_{k=1}^{K} \exp(\gamma_{ks})} \]  

\[ P(y_{it}^{1991} = \ell | x_t = s) = \frac{\exp(\beta_{1991}^{1991})}{\sum_{m=1}^{M} \exp(\beta_{ms}^{1991})} \]  

\[ P(y_{it}^{2000} = \ell | x_t = s) = \frac{\exp(\beta_{2000}^{2000})}{\sum_{m=1}^{M} \exp(\beta_{ms}^{2000})} \]

Under the identifying restrictions \( \alpha_1 = 0 \text{ and } \gamma_{11} = \gamma_{22} = \gamma_{33} = \gamma_{44} = 0 \text{ and } \beta_{11} = \beta_{22} = \beta_{33} = \beta_{44} = 0, \)

we get the following rather straightforward interpretations of the logit parameters:

\[ \alpha_s = \log \frac{P(x_0 = s)}{P(x_0 = 1)} \]

\[ ^{23} \text{Because these are multinomial logit models, we need one constraint on } \alpha, \text{ four on } \gamma \text{ (one for each value of } x_{t-1} ), \text{ and four on } \beta \text{ (one for each value of } x ); \text{ that is, one category of the dependent variable should be treated as the reference category. In our “transition coding” and “error coding” schemes, the reference category changes with the value of the independent variable in the equation, giving the parameters an interpretation that makes sense in latent Markov models.} \]
\[ \gamma_{rs} = \log \frac{P(x_t = r | x_{t-1} = s)}{P(x_t = s | x_{t-1} = s)} \] (7)

\[ \beta_{\ell s}^{1991} = \log \frac{P(y_{1991}^{\ell} = \ell | x_t = s)}{P(y_{1991}^{s} = s | x_t = s)} \] (8)

\[ \beta_{\ell s}^{2000} = \log \frac{P(y_{2000}^{\ell} = \ell | x_t = s)}{P(y_{2000}^{s} = s | x_t = s)} \] (9)

That is, as the log-odds of having initial state \( s \) rather than 1 (equation 6), the log-odds of making a transition from state \( s \) to state \( r \) rather than staying in state \( s \) (“transition coding,” equation 7), and the log odds of misclassifying someone with latent state \( s \) by assigning observed state \( \ell \) rather than the correct state \( s \) (“error coding,” equation [8] for the 1991 response and equation [9] for the 2000 response). Note that equations (7), (8), and (9) follow from the fact that the “no transition” and the “no error” categories, respectively, serve as the reference categories; that is, they follow from the fact that the reference category changes with the origin state (true state).

The model described in equation (1) is the basic latent Markov model. It makes certain assumptions about (1) the way people move between latent states (the transitions) and (2) the way the observed states are reported given the true states (the measurement error). The main assumption about the (latent) transitions is that these are in agreement with a first-order Markov process, which implies that the state occupied at time point \( t \) depends only on the state occupied at time point \( t-1 \) (the previous state), and that it is thus independent of the states occupied at all other time points. In addition, it is assumed that transition probabilities are time homogeneous and that there are no observed or unobserved factors affecting the transition probabilities.24 Although it would be possible to relax each set of assumptions, we keep the model for the transitions as simple as possible since our main interest is not in the parameters of a transition model but in the measurement error.

24Another assumption about the transitions is that the 1991 and 2000 responses do not affect the latent state occupied at a later time point. This assumption is very unlikely to be violated.
parameters as well as in the posterior estimates of the respondents’ true states. It turns out that this information of main interest is not (or very weakly) affected by the model we use for the transitions. That is because we have two measurements of a person’s true state (not just one) and because we have a rather long time series. We obtain very good (almost perfect) estimates for the true event sequences, which cannot be improved by a more sophisticated transition model. Relaxing the first-order Markov assumption, adding covariates (including time) in the model, and modeling unobserved heterogeneity using a time constant latent variable will improve the prior estimates of the transition probabilities for a specific respondent, but it will not affect either the estimates of the measurement error or the posterior estimates of a person’s true state. For this reason, we decided to keep our models as simple as possible as far as the transition part of the model is concerned.

The two key assumptions in the measurement part of the basic latent Markov model are that, conditional on a person’s true states, the 1991 and 2000 responses at time point $t$ are (1) independent of one another and (2) independent of the responses at the other time points. In latent class analysis, the former assumption—which is connected to the fact that we have two measurements—is referred to as “local independence” (Lazarsfeld and Henry 1968; Goodman 1974). The second assumption is sometimes referred to as “independent classification errors” (ICE; Bassi et al. 2000; Biemer and Bushery 2000). It is the ICE assumption in particular that turns out to be problematic in our application, which is why we propose various modifications of the basic latent Markov model for relaxing this assumption. Similar kinds of latent Markov models with correlated measurement errors were proposed by Hagenaars (1988) for the analysis of two-wave panel data and by Bassi and colleagues (2000) for the analysis of a single employment history collected with a rotation panel design. Elzinga, Hoogendoorn, and Dijkstra (2007) proposed similar models, which they called linked Markov sources, for modeling outcome-dependent social processes, such as conversations arising in structured interviews.

Two similar assumptions are that the 1991 response does not affect the 2000 response at later time points and that the 2000 response does not affect the 1991 response at later time points. These assumptions are very unlikely to be violated, and the latter is even logically impossible.
For $t=0, 1, 2, \ldots, T$

**Models 1 and 2**

<table>
<thead>
<tr>
<th>Manifest</th>
<th>$y_{it}^{1991}$</th>
<th>$y_{it}^{2000}$</th>
<th>$y_{it}^{1991}$</th>
<th>$y_{it}^{2000}$</th>
<th>$y_{it}^{1991}$</th>
<th>$y_{it}^{2000}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Latent</td>
<td>$X_{it0}$</td>
<td></td>
<td>$X_{it1}$</td>
<td></td>
<td>$X_{it2}$</td>
<td></td>
</tr>
</tbody>
</table>

Markov process

**Model 3**

<table>
<thead>
<tr>
<th>Manifest</th>
<th>$y_{it}^{1991}$</th>
<th>$y_{it}^{2000}$</th>
<th>$y_{it}^{1991}$</th>
<th>$y_{it}^{2000}$</th>
<th>$y_{it}^{1991}$</th>
<th>$y_{it}^{2000}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Latent</td>
<td>$X_{it0}$</td>
<td></td>
<td>$X_{it1}$</td>
<td></td>
<td>$X_{it2}$</td>
<td></td>
</tr>
</tbody>
</table>

Markov process

**FIGURE 2.** Graphical representation models of three.

In our application, we used three kinds of specifications for the measurement error component of the latent Markov model. Figure 2 provides a graphical representation of the models. Table 6 presents their fit measures (log-likelihood value and BIC), as well as their measurement errors estimates.

Model 1 is a standard latent Markov model assuming ICE. In addition, it assumes that the response at the 1991 interview is perfect.\(^{26}\) Note that a specification of measurement errors for the 1991 response is conceptually similar to what was done with the OM approach, which quantified how well the 2000 sequences match the 1991 sequences. The first set of probabilities reported in Table 6 represents the estimates of the probability of being in a particular latent state averaged across

---

\(^{26}\)Note that the probability of making a classification error can be fixed to zero by equating each of the logit parameters $\beta_{1991}^{\ell}s$ appearing in equations (4) and (8) to a large negative value (e.g., $-100$) for $\ell \neq s$. 

---
### Table 6
Results for the Markov Models

<table>
<thead>
<tr>
<th>Model 1: ICE with $y^{1991}$ no error</th>
<th>Model 2: ICE</th>
<th>Model 3: non-ICE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Log-likelihood = $-14813$;</td>
<td>Log-likelihood = $-11850$;</td>
<td>Log-likelihood = $-9277$;</td>
</tr>
<tr>
<td>Parameters = 27;</td>
<td>Parameters = 39;</td>
<td>Parameters = 51;</td>
</tr>
<tr>
<td>BIC = 29831</td>
<td>BIC = 23995</td>
<td>BIC = 18940</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>True State</th>
<th>True State</th>
<th>True State</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y^{1991}$</td>
<td>$y^{1991}$</td>
<td>$y^{1991}$</td>
</tr>
<tr>
<td>Size</td>
<td>0.75</td>
<td>0.71</td>
</tr>
<tr>
<td>E</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>S</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>N</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>U</td>
<td>0.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$y^{2000}$</th>
<th>$y^{2000}$</th>
<th>$y^{2000}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>E</td>
<td>0.96</td>
<td>0.96</td>
</tr>
<tr>
<td>S</td>
<td>0.01</td>
<td>0.00</td>
</tr>
<tr>
<td>N</td>
<td>0.03</td>
<td>0.00</td>
</tr>
<tr>
<td>U</td>
<td>0.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$y^{2000}$ with error at $t-1$</th>
<th>$y^{2000}$ with error at $t-1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>E</td>
<td>E</td>
</tr>
<tr>
<td>S</td>
<td>S</td>
</tr>
<tr>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>U</td>
<td>U</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$y^{2000}$ with error at $t-1$</th>
<th>$y^{2000}$ with error at $t-1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>E</td>
<td>E</td>
</tr>
<tr>
<td>S</td>
<td>S</td>
</tr>
<tr>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>U</td>
<td>U</td>
</tr>
</tbody>
</table>
time points. The other ones are the response probabilities, which sum to 1 within columns. The results for Model 1 show that the estimated probability of being truly employed is .75. For this state the estimated measurement error in the 2000 report is rather small (less than 5%). The self-employed and nonemployed states (containing 7% and 17% of the person periods) have much larger measurement errors (22% and 19%). For the unemployment state (estimated to contain 1.5% of the person months), the measurement error is extremely large with a misclassification rate of 69%.

Model 2 is also an ICE model, but it relaxes the assumption that the 1991 response is perfect. As indicated by its BIC value, this model should be preferred over Model 1. However, as can be seen from the measurement error probabilities, it yields the unrealistic result that the 2000 report is much better than the 1991 report. This is due to an artifact of the basic latent Markov model, which sees the larger stability of the 2000 series as an indication that there are fewer measurement errors. This problem, which is caused by the ICE assumption, becomes visible when both the 1991 and 2000 measurement error probabilities are estimated freely. It can be solved by relaxing the ICE assumption, which is feasible in our application because we have two measurements of a person’s state at each occasion.

The reason why the ICE assumption is unrealistic in our application relates to the fact that in a retrospective survey, during the recall process, a respondent could forget or misplace episodes, which is much more likely to happen with increasing recall time. The consequence of such mistakes is that the same mismatch between true and observed state (the same type of error) occurs at consecutive months. As a result, errors at consecutive time points are not independent of one another, and the ICE assumption is thus clearly invalid. Although it is unlikely that such correlated errors are fully absent in the 1991 report, it is reasonable to assume that these are much weaker when reporting over the previous year than when reporting about approximately 10 years before.27 If we were to allow dependent classification errors for both responses, we would run into the risk of not only fully distorting the

27Here we would like to note that comparing survey data to administrative data, although offering additional insight into the measurement error problem, would not solve it, since administrative data do not necessarily reveal the “truth” because they also contain error, though of a different nature.
structure of the latent Markov model but also running into identification problems. Therefore, we relaxed the ICE assumption only for the 2000 report. If the ICE assumption for 1991 does not hold exactly, the amount of measurement error in the 1991 responses may be slightly underestimated and in the 2000 responses slightly overestimated.

Model 3 relaxes the ICE assumption by including dependencies between errors occurring across months in the 2000 report. More specifically, we allow the probability of a particular type of error at month \( t \) to depend on the type of error made at month \( t - 1 \). This requires that \( y_{it}^{2000} \) depends on both the latent and the observed state at \( t - 1 \). This is achieved by replacing the term \( P(y_{it}^{2000} = \ell| x_t = s) \) appearing in equation (5) with the term \( P(y_{it}^{2000} = \ell|x_t = s, x_{t-1} = r, y_{i,t-1}^{2000} = k) \).

Given that \( K = M = 4 \), introducing both the lagged latent state and the lagged 2000 reported state \( (x_{i,t-1} \text{ and } y_{i,t-1}^{2000}) \) into the measurement error part of the latent Markov model implies that there are 16 (4 times 4) matrices of error probabilities, one for each possible combination of lagged observed and latent state. On the one hand, this shows that this model allows considerable flexibility. On the other hand, while it is possible to estimate all 16 sets of error probabilities freely, this may yield many parameters possibly difficult to interpret. We therefore opt for imposing more structure on the measurement error parameters for the 2000 report. More specifically, we assume that there are two sets of error probabilities (instead of 16), one for the situation in which the report at \( t - 1 \) was correct and one for the situation in which it was incorrect. Such a more restricted structure can be easily defined using a logit parameterization of the error probabilities; that is,

\[
\log \frac{P(y_{it}^{2000} = \ell|x_t = s, x_{t-1} = r, y_{i,t-1}^{2000} = k)}{P(y_{it}^{2000} = s|x_t = s, x_{t-1} = r, y_{i,t-1}^{2000} = k)} = \beta_{\ell s}^{2000} \delta_{rk} + \lambda_{\ell s}(1 - \delta_{rk}).
\]

Here, \( \delta_{rk} \) is an indicator variable (a dummy) taking on the value 1 if \( k = r \) and 0 otherwise. Hence, compared to the ICE model, we have two sets of parameters, one for the situation in which the report at \( t - 1 \) was correct \( (\delta_{rk} = 1) \) and one for the situation in which it was incorrect \( (\delta_{rk} = 0) \). As in the ICE model, for identification, we set \( \beta_{\ell s}^{2000} = \lambda_{\ell s} = 0 \) for \( \ell = s \). This yields two sets of 12 measurement error coefficients.

Model 3 has a much lower BIC value than Model 2, showing that including dependencies between classification errors improves the
model fit. The estimated measurement errors for the 1991 response show that there is basically no error for the employment and self-employment states, very small error for the not-in-employment state, and much larger error (51%) for the unemployment state. The much larger measurement error in the unemployment state is well documented in the literature (Magidson et al. 2009). The estimated measurement error probabilities show that the error at the second interview is very small if a respondent correctly reported the state at the previous time point. For unemployment, though, the misreport is larger than for the other states (about .19), even when no error in the lag occurred. When any type of classification error occurred at the previous time point instead, many more errors occur at the current time point. This applies to all four true states. In this situation, self-employment is mainly confused with employment and nonemployment and unemployment with employment and self-employment, while employment and nonemployment are often reversed.

An interesting result is that when no error occurred at \( t - 1 \) (which is the case for 92% of the person-months, with more than 82% of the respondents who never made a mistake at all),\(^{28}\) the quality of the response at the second, more distant, interview (2000) is similar to the one of the first interview (1991), and sometimes even better (because the error for unemployment is smaller). In contrast, when error occurred at a previous time point, the quality of the response greatly deteriorates (in more than 91% of the person-months with an error at \( t - 1 \) errors occur at \( t \) as well; in terms of respondents, more than 99% of those who ever made a mistake at \( t - 1 \) made a mistake at \( t \)). These results confirm that this model specification makes sense: After filtering out the bad cases, the quality of the 1991 and 2000 reports are similar, which means for the shorter and longer recall periods.

The specific constraints we impose on \( P(y_{it}^{2000} = \ell|\mathbf{x}_t = s, \mathbf{x}_{t-1} = r, y_{i,t-1}^{2000} = \tilde{\ell}) \) represent one of the many possible restrictions that can be used. However, it turned out that the results of main interest are robust to the exact specification that is used. We also tried a different specification that instead of distinguishing between reports on and off the main diagonal at \( t - 1 \) (no error versus any error) looks at the type of mistake and in particular whether the respondent made the

\(^{28}\)We are able to identify errors comparing the observed responses with the posterior mode estimates at each time point \( t \).
same mistake at the previous time point. The underlying idea is that a particular type of mistake is more likely to occur if the same mistake occurred at the previous time point: the log-odds of a mistake at $t$ depend on having had the same mistake at $t - 1$. This specification is obtained from a logit model of the form $\beta^{2000}_s + \ell_s \delta_{srk}$, where $\delta_{srk} = 1$ if $\ell \neq s & \ell = k & s = r$, and 0 otherwise.

### 4.4. Latent State Estimates and Sequences

After estimating the parameters of the latent Markov model of interest, we can obtain estimates of the true state occupied by individual $i$ at time point $t$ using the parameter estimates and the observed responses $y^{1991}_i$ and $y^{2000}_i$. These estimates are based on the posterior class membership probabilities $P(x_t = s | y^{1991}_i, y^{2000}_i)$, where, as in a standard latent class analysis, a subject is usually assigned to the class with the largest probability (Goodman 1974). The estimated proportion of correct classification is obtained as the average of $\max_s P(x_t = s | y^{1991}_i, y^{2000}_i)$ across person-months, a number that turned out to be extremely large in our application. That is, we obtain an estimated proportion of correct true state predictions of 99.8%.

We used the true state predictions for a further analysis in which this estimated sequence of true states is compared with the two observed sequences. Note that this represents an improvement compared to the OM approach in which we quantified how well the 2000 sequence matches the 1991 sequence. The OM approach yields a symmetric distance between the two sequences, which means that it does not matter which sequence is treated as the (error free) reference sequence. However, because the 1991 report can be assumed to be the more reliable one, it is most logical to treat that as the reference.

<table>
<thead>
<tr>
<th>OM Distance</th>
<th>Mean</th>
<th>Bootstrap C.I.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1991–latent</td>
<td>0.035</td>
<td>(0.0268-0.0435)</td>
</tr>
<tr>
<td>2000–latent</td>
<td>0.165</td>
<td>(0.144-0.185)</td>
</tr>
<tr>
<td>1991–2000</td>
<td>0.195</td>
<td>(0.172-0.217)</td>
</tr>
</tbody>
</table>
TABLE 8
OLS Regression on the Distance Between Latent and Observed Response, by Interview

<table>
<thead>
<tr>
<th></th>
<th>Latent-Observed</th>
<th>(a) Latent-1991</th>
<th>(b) Latent-2000</th>
<th>Sig. of (a)-(b)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>0.011</td>
<td>−0.003</td>
<td>0.025</td>
<td></td>
</tr>
<tr>
<td>Birth year</td>
<td>−0.0006</td>
<td>0.0006*</td>
<td>0.0007</td>
<td></td>
</tr>
<tr>
<td>Episodes of unemployment</td>
<td>0.394**</td>
<td>0.492**</td>
<td>0.296**</td>
<td>*</td>
</tr>
<tr>
<td>Number of episodes</td>
<td>0.078**</td>
<td>−0.067**</td>
<td>0.224**</td>
<td>**</td>
</tr>
<tr>
<td>Later report (2000)</td>
<td>0.130**</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>0.127</td>
<td>−1.147*</td>
<td>1.525</td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>3946</td>
<td>1973</td>
<td>1973</td>
<td></td>
</tr>
<tr>
<td>R²</td>
<td>0.17</td>
<td>0.33</td>
<td>0.14</td>
<td></td>
</tr>
</tbody>
</table>

**P < 0.01.

obtained by comparing the estimated true states according to Model 3 with both the 1991 and 2000 sequences.

In Table 7,30 we present an OM-based dissimilarity measure between the observed and the estimated true sequences, obtained in the same way that we compared the two observed sequences (see results in Table 5). While in the latter case the observed response in 1991 was used as a reference, we now take the true states as reference, and we compare both the 1991 and 2000 responses to it. As shown in Table 7, the average distance to the true sequence is larger in the 2000 than in the 1991 report, and as indicated by the bootstrap confidence interval this difference is statistically significant. Moreover, comparison of these numbers with the estimated distance between the two observed sequences (in 1991 and 2000) shows that the mismatch of 2000 sequences with the true sequence is smaller than its mismatch with the 1991 sequence. This indicates that the amount of recall bias in the 2000 interview is overestimated when its assessment is based on a comparison with a better though still imperfect observed sequence.

We used regression analysis, similar to the approach used in Section 4.2 (Table 5), to assess the effects of the respondent and career characteristics on the quality of reporting. Table 8 shows the results.

30The distance measures presented in Table 7 are based on the “default” substitution cost matrix as defined in Section 4.2.
The first column refers to a regression analysis on the full set of sequence distances, where, compared to our approach in Table 5, we also add an independent variable for survey (using 1991 as reference), while experience of unemployment episode as well as the number of episodes in the sequence both refer to the reconstruction based on the latent state. Here, we look at the overall effect of the covariates of interest, controlling for the data source of observed response—that is, whether the dissimilarity with the latent sequence refers to the 1991 or 2000 report. We find that the number of episodes indicating the complexity of the career and the presence of episodes of unemployment have a significant effect, increasing the number of errors. However, their effects seem different at different recall times, as can be seen in the second and third columns, where we show the results of the regression analysis on each survey separately; the last column, obtained after adding the interaction terms between the survey and each of the explanatory variables, shows whether the effects of the covariates are significantly different in the two equations.\textsuperscript{31} Controlling for other differences, episodes of unemployment increase report errors to a significantly greater extent for particularly short recall distances (1991 interview), although their effect is positive and strongly significant also for longer recall distances (2000). The number of episodes in the sequence seems instead to make the quality of the report worse only for longer recall times.

Table 9 provides the last results we would like to discuss. For both the 1991 and 2000 interview, it presents the two-way cross-tabulation of the type of observed sequence with estimated true sequence. This table shows that the sequences reported at the first interview in 1991 are closer to the latent sequences than the sequences reported at the interview in 2000. However, this seems to be especially the case for careers involving (more) transitions. These results show again that comparing the two observed responses with one another, instead of with the true sequence, overestimates the amount of recall bias.

\textsuperscript{31}The Stata cluster option was used to correct the standard errors for the dependency of the 1991-latent and 2000-latent distances.
5. CONCLUSION

This article discussed methods for studying memory bias in retrospective reports when multiple reports are available for the period of interest. In addition to descriptive and exploratory matching techniques, we proposed to use a model-based approach that makes use of latent Markov
models. The main advantage of the latter approach is that it is more flexible while relaxing the requirement that one of the responses is perfect; hence, all reports may contain measurement errors. A novel element in the proposed latent Markov modeling approach is that it allows for dependent classification errors across adjacent time points. Moreover, after the estimation of a latent Markov model, matching techniques can be used to quantify the dissimilarity between the true and observed sequences. These dissimilarities can be used as dependent variables in a regression analysis to detect factors related to the occurrence of recall bias.

Our analyses of the Swedish Level of Living Survey showed that recall errors lead to simplified employment careers, especially for longer recall times, resulting in recorded careers with fewer episodes and transitions than they actually had. One of the consequences of this simplification is that career stability may be overestimated with retrospective data. Furthermore, we found that recall accuracy depends on career complexity and the occurrence of unemployment spells: Errors are more likely to occur in more complex careers, especially for longer recall periods, as well as in careers containing unemployment spells, also for shorter recall periods. It should be noted that unemployment is more likely to be misreported than employment (and other states) independent of spell duration. In fact, a short spell duration does not make the report of unemployment any worse, while it increases the measurement error for the other states, which seem to be mistakenly reported mainly for short(er) spell durations. This shows that although unemployment spells are shorter and shorter spells contain more errors, unemployment and duration effects are not confounded.

The proposed latent Markov modeling approach requires making assumptions about the transition and recall error processes, which means that the results may be dependent on these assumptions. In our application, assumptions about the transition process turned out to

32Results are not shown here but can be obtained from the first author upon request.
33The posterior mode estimates of the true state show that average length was 9 months for unemployment spells and 16 months for employment spells. The average duration of spells containing errors was about 4 months shorter than spells without errors.
have little impact on the measurement error estimates. As far as the recall error is concerned, the two most critical assumptions are the ICE assumption for the 1991 responses and the non-ICE assumption for the 2000 responses. Our main results turned out to be quite robust in the exact non-ICE model used for the 2000 responses. It was, however, not possible to relax the ICE assumption for the 1991 responses without distorting the complete model. If this assumption does not hold exactly, the amount of measurement error in the 1991 responses may be slightly underestimated and in the 2000 responses slightly overestimated. Other assumptions about the measurement error, such as local independence between the two responses, are very likely to hold in our application.

The approach developed here combining latent Markov modeling with matching techniques permits us to quantify recall bias and to identify its determinants. Although this yields highly relevant information for substantive research based on retrospective life-history data, a logical next step would be to develop a method for performing substantive analysis while correcting for measurement errors. The latent Markov modeling framework offers various possibilities to set this step as well. The most straightforward approach would be to build the substantive model of interest for the transitions across the true states from the latent Markov model with covariates as proposed by Vermunt, Langeheine, and Böckenholt (1999), and recently applied by Bartolucci, Lupparelli, and Montanari (2009) and Paas, Vermunt, and Bijmolt (2007). Another option would be to extend the transition model by allowing for unobserved heterogeneity—for example, through the specification of a mover-stayer model (Goodman 1961; Yamaguchi 2008). Both approaches would allow sociological researchers to draw more reliable inferences about the alleged rising labor market turbulence and their impact on stratification as a consequence of globalization.

REFERENCES


