Unified BHPS work-life histories: combining multiple sources into a user-friendly format

Brendan Halpin
ESRC Research Centre on Micro-social Change,
University of Essex

May 1997

Unified BHPS work-life histories: combining multiple sources into a user-friendly format

Abstract

Longitudinal data is often difficult to use, and continuous histories collected in a panel are a particularly unfriendly case. This paper reports an exercise to re-organise the British Household Panel Study's work-life history data into a format more convenient for analysis.

The British Household Panel Study collects extensive labour market history information from its respondents, both during the panel period and retrospectively from labour market entry. That this information is of necessity stored in multiple locations, and of varying levels of detail, has made use somewhat inconvenient. This paper describes an exercise to bring the labour market information together in a more convenient format. It also considers some of the problems of retrospective and panel longitudinal data, and discusses issues of recall error and measurement error.

The data files described are available through the UK Data Archive.

Acknowledgements

The work this paper reports draws on program code written by Nick Davey, J. Gershuny and Mark Taylor. Discussions among members of the Centre have also been extremely important, both in guiding the initial design and reformulating it in the light of users' experience: therefore thanks to Nick Buck, Carmel Hannan, Jonathan Scales and Adrian Birch, and outside the Centre, Sheila Jacobs and Richard Layte. Finally, the useful suggestions of participants of the BHPS User Group meeting, London December 1996, are gratefully acknowledged.

Unified BHPS work-life histories: combining multiple sources into a user-friendly format

1 Organising longitudinal data for easy use: a problem

Longitudinal data is more difficult to deal with than cross-sectional data, if only because it introduces an extra dimension of complexity, time. This is true of simple panel data, with observations at repeated time-points, but it is even more true of continuously represented data (*e.g.*, event histories) where these are collected repeatedly. While there is conceptual complexity involved in the analysis of temporal data, there is also organisational complexity: the way in which longitudinal data are organised has a large effect on the facility with which analyses can be carried out. Indeed, researchers often complain that they spend more time manipulating data than analysing it, and some researchers simply avoid the longitudinal aspect altogether. This paper reports an exercise to take a relatively complex longitudinal data set, the work-life history components of the British Household Panel Study, and to re-organise them in as simple a format as is compatible with minimal loss of information. That is, we take a data set with up to seven different event-history accounts of labour market activity, and five single time-point reports, and systematically generate unified representations of the data, which are far easier to analyse.

The British Household Panel Study collects extensive information on respondents' labour market status, (i) at the time of interview at each wave of the panel, (ii) through the period between 1 September a year before and the interview date, and (iii) retrospectively from first leaving full-time education. Because the retrospective information was collected in two tranches (one focusing on employment status, the other on occupational information) there are four different types of labour-market history information, located (at Wave 5) in twelve different files in the BHPS database.[1] This complexity is a necessary aspect of longitudinal information, but it has inhibited use of the work-life history data.

This paper describes the creation of a set of 'reconciled', user-friendly, files, examines their output and discusses some aspects of measurement error and recall bias relevant to the exercise. The first part of the exercise is to take the 'current status' information and combine it with the inter-wave history, for each wave, and then to combine the five waves thus creating a continuous record from September 1990 to the September 1995 (and later).

[1] At the time of writing, five waves, representing 1991 to 1995, of the BHPS are released, and available through the UK Data Archive.

The second stage is to take the life-time employment status history collected at Wave 2, and the life-time occupational history collected at Wave 3, and to combine each of them with analogous information drawn from the combined panel file, thus creating employment and occupational histories that stretch from labour-market entry to the latest wave. The third stage is to combine these two extended life-time histories into a single record which contains both employment-status information (with good information about non-employed spells) and occupational information (that is, details about the job held during each employed spell).

The paper describes the methods used, in terms of an initial specification and its detailed implementation, and goes on to consider the output produced. By including the retrospective data we have information stretching back many decades, though from the point of view of breadth of detail and quality of recall the panel-derived data (covering 1990–95) are much better. When we compare data from different sources, we find systematic differences, as would be expected, from different sources, but a reasonably good level of agreement between the two long-term retrospective files.

Finally we consider issues of measurement error and recall bias, both of which are particularly relevant to longitudinal data. The design of the BHPS is advantageous from this point of view as it allows us to assess the extent of recall bias, in that there are built-on overlaps in coverage. Under this rubric we also consider 'seam' effects, *i.e.*, the artefactual status changes that are created by combining different data sources, and suggest means of taking account of them.

1.1 Multiple sources of information

The British Household Panel Study is a panel survey of approximately 5,500 households in Great Britain. The survey collects information on a broad range of topics, one of which is labour market activity. At the time of writing fieldwork for the seventh wave of the panel is underway, the data for the sixth wave is nearly ready for release, and five waves of data, covering 1991 to 1995, are available to researchers. All adult members of the 5,500 households are interviewed, which amounts to between 9,000 and 10,000 individual full respondents each year. The fieldwork period starts on 1 September. A broad range of areas is covered in the questionnaire, which has a core which remains the same from year to year, as well as sections which are repeated on a lower frequency or are one-off. In the exercise this paper reports we focus exclusively on labour market status, with a view to bringing together information distributed across a number of data files into a more convenient package.

Information is recorded on labour market status at each interview, and for the period beginning on 1 September a year (or more) prior to the interview. This method ensures that a continuous record of labour market status is collected, at the expense of some overlap from wave to wave. Thus, for example, for respondents present at Waves 1 to 5 we have a complete and detailed record of their labour market status from 1 September 1990 (or before: the start-date of a job held at that date is known) to at least 1 September 1995. However, since it is highly desirable to have information on the respondent's entire career, retrospective data were also collected in Waves 2 and 3, to fill in the gap since leaving full-time education to the start of the panel-derived labour market history. In Wave 2 a complete employment status history was collected, recording non-employed states in detail, and in Wave 3 a complete job history was collected, with detailed information on every job held.

Thus the information exists to construct a complete employment/labour market status history for nearly every individual in the survey, from his/her first job to the latest wave of the panel. The problem is that this information is of necessity collected at different times, in somewhat different ways, and recorded in different locations. Below we describe the work done in reconciling the various sources and producing such a single continuous record.

Full information on what data were collected, and how and when they were collected, is available in the main documentation of the BHPS (Taylor, 1996).

1.1.1 Multiple records

Table 1 presents and describes all the files from the main BHPS database from which information is drawn in constructing the unified longitudinal files. Details of the current job (or other status) are recorded in *wINDRESP*. (By convention, all data files – and all variables therein – relating to a particular wave start with the same letter, 'A' for Wave 1, 'B' for Wave 2, *etc.* Thus individual level data for Wave 1 is stored in *AINDRESP*. We use *wINDRESP* and so on as a means of referring to files independently of wave. To refer to consecutive pairs of waves we talk of *qINDRESP* and *pINDRESP*, where q indicates the wave after p.) Because rather more information is collected for the current status than for prior spells during the wave's reference period, and because the data have a different structure (*i.e.*, the event history has zero or more records per individual), the inter-wave job history is stored separately in *wJOBHIST*. The two long-term retrospective life histories (LTRs) are recorded in *BLIFEMST* and *CLIFEJOB*, representing respectively a continuous history

Table 1: Files in the main BHPS database containing information on respondents' work-life history

Filename	Wave	Start of fieldwork	Description
'Panel' files			
AINDRESP	1	Sept 1991	The main 'individual respondent' file, containing *inter alia* detailed information on current status at the date of interview
AJOBHIST	1	Sept 1991	Information on all employment status spells between 1/9/90 and the date of interview
BINDRESP	2	Sept 1992	Wave 2 equivalent of *AINDRESP*
BJOBHIST	2	Sept 1992	Inter-wave history: details of all employment status spells between 1/9/91 and the date of interview
CINDRESP	3	Sept 1993	Wave 3 equivalent of *AINDRESP*
CJOBHIST	3	Sept 1993	Inter-wave history: details of all employment status spells between 1/9/92 and the date of interview
DINDRESP	4	Sept 1994	Wave 4 equivalent of *AINDRESP*
DJOBHIST	4	Sept 1994	Inter-wave history: details of all employment status spells between 1/9/93 and the date of interview
EINDRESP	5	Sept 1995	Wave 5 equivalent of *AINDRESP*
EJOBHIST	5	Sept 1995	Inter-wave history: details of all employment status spells between 1/9/94 and the date of interview
Long-term retrospective files (LTR)			
BLIFEMST	2	Sept 1992	Information on all employment status spells since first leaving full-time education until the date of interview
CLIFEJOB	3	Sept 1993	Information on all jobs held since first leaving full-time education until the beginning of data collection in the main panel

Note: Files up to and including Wave 5 are represented. Further waves will add further *wINDRESP*/*wJOBHIST* pairs.

of employment status since first leaving full-time education up to the Wave 2 interview, and a complete account of all jobs held between entering the labour market and the first job reported in the panel. Thus for the five waves currently released, there are twelve locations in which information on labour market history is recorded. Some parts of this information are easier to reconcile, for instance *wINDRESP* and *wJOBHIST* in any given wave (*i.e.*, the information on current job or status, and the information on status history since 1 September a year or more ago). Reconciling consecutive waves is only slightly less straightforward, given that we should have complete coverage of the interval between interviews, but there will usually be a small amount of overlap (in most cases last year's interview will have taken place after 1 September) giving us multiple (and, necessarily, sometimes conflicting) records of status in the period between 1 September and the p-interview date.

The two long-term retrospective histories, *BLIFEMST* and *CLIFEJOB*, present greater difficulties. While integrating them with the panel-derived histories is relatively straightforward, save for data conflicts, integrating them with each other is more challenging. They are both accounts of the same history, but were collected a year apart, and have substantially different emphases. In particular, in Wave 2 life histories for several domains were collected, including fertility and family formation. The employment status history was collected as part of this exercise, and can be expected to benefit from being part of a multi-domain recollection, where remembering salient events in one domain can refresh the memory of events in other domains. However, because of the time involved in collecting good occupational, as opposed to employment-status, information, the occupational history could not be collected in this exercise and was held over to Wave 3. There is a good level of agreement, all things considered, between the *BLIFEMST* and *CLIFEJOB* reports of broad employment-status categories, but it is not complete (see section 4.4.1).[2] Resolution of these conflicts requires decisions which must be driven by the substantive interest behind the analysis.

1.1.2 From real life to databases

Figure 1 illustrates the process by which an individual's labour market history is collected and then recorded in the database, and a little of how the present exercise combines this information into a reconciled format. We use a fictitious individual who experiences four labour-market status spells between the mid-1980s and the panel period. At the date of interview in Wave

[2] Agreement is no doubt enhanced by the practice of 'feed-forward', that is, letting respondents at Wave 3 see their Wave 2 life-time employment-status data.

1 (sometime a little after 1/9/91) the respondent is in his/her third spell, and information about this is stored in *AINDRESP*. The reference period for Wave 1 stretches back to 1/9/90, when this respondent was the second of the four spells represented. Information about this spell (including its start and end dates) is recorded in *AJOBHIST*. As the first line below the time line shows, these can be considered jointly as representing the Wave 1 work history.

In 1992, this respondent is interviewed again, but is still in the same spell as at the previous interview. Therefore there is no *BJOBHIST* information to record, and the entire Wave 2 work history consists of a description of current state in *BINDRESP*. As the second line below the time line shows, the Wave 2 work history overlaps the Wave 1 work history, in two ways: first, the Wave 2 reference period extends to 1/9/91, which is shortly before the Wave 1 interview date (for nearly all respondents); and second, because the same job was held at both interviews, thus creating overlap in the period between job-start and the Wave 1 interview.

We can create a single panel-derived work history by simply putting successive waves together, giving precedence to the earlier wave's information up to its interview date in the case of data disagreements (see the third line below the time line). All job spells prior to the spell current at 1/9/90 will be picked up in *CLIFEJOB*, and a continuous history of employment status is recorded in *BLIFEMST*.

1.2 Coding and recall error

There are two major sources of inconsistency between records: coding error and recall error.

Coding error is a serious problem with occupational data: coders may code a particular job description differently at different times because of inherent ambiguity (this problem is reduced by computer-aided coding procedures such as CASOC[3]) and (perhaps more seriously) individuals may describe a given job differently at different times. This will lead to spurious occupational mobility and is an endemic problem with data of this sort (see

[3] CASOC is an interactive program for coding occupational descriptions into (primarily) the Standard Occupational Classification. It 'knows' the classification and has a large database of job descriptions attached to categories in the classification, and it searches it dynamically as the coder inputs the textual job description. It is believed to have a large positive effect on the reliability of occupational coding, and a small positive effect on its speed.

CASOC is widely used in coding occupational data, especially in UK government agencies. It is described fully in Elias, Halstead and Prandy (1993).

CASOC is used for all occupational coding in the BHPS, and provides not only SOC but also automatically generates a variety of social classifications derived from SOC.

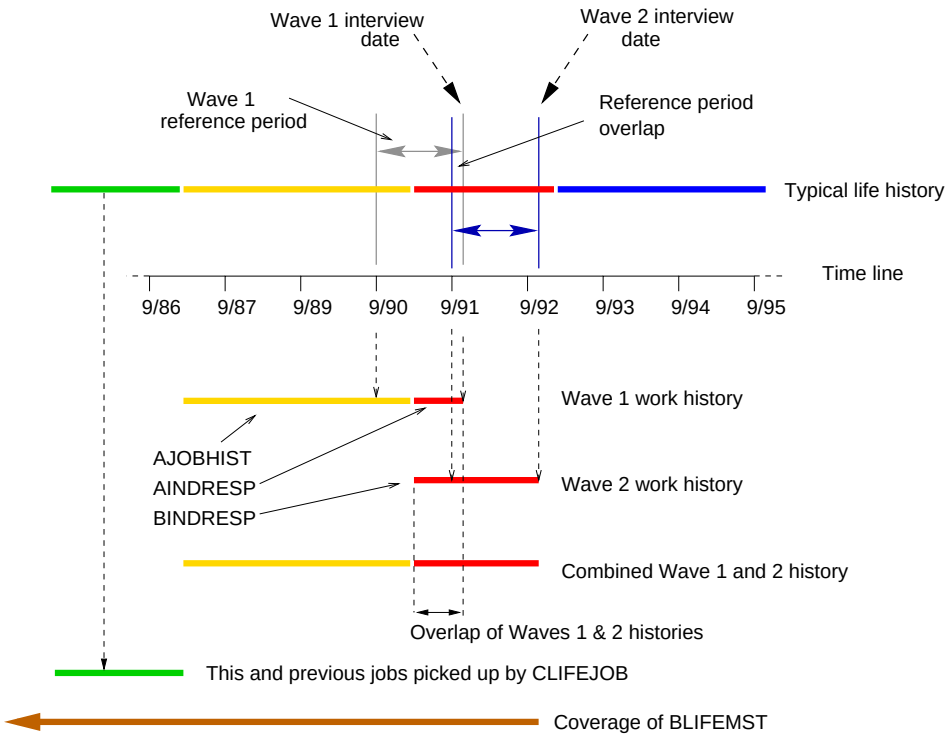

Figure 1: A typical life history as it would be recorded in the BHPS

section 4.4.3). Recall error is a product of the fallibility of human memory. In general terms, the longer since an event, the less accurately we remember it. However, recall reliability is not simply a function of elapsed time: type of event also affects recall. For instance, short episodes are more easily forgotten, as perhaps are unpleasant episodes such as unemployment. Also, certain transitions (such as starting a job) may be better remembered than others (such as moving from one non-employed state to another, for instance from unemployment to early retirement, or from home-working to unemployment). Dex (1995) summarises literature on attempts to assess the magnitude of recall effects, often by re-interview after an extended period; Paull (1996) uses the BHPS wave-on-wave overlap to assess the nature of recall bias over a year; Elias (1996) and Dex and McCulloch (1997) look at recall of unemployment in the BHPS ***BLIFEMST*** employment-status history and in the *Family and Working Lives* survey, and with reference to the Labour Force Survey.

Each of these types of data problem is present in the BHPS, and complicates any reconciliation project by raising the level of inconsistency. However, the multiplicity of data sources relating to work-life history presents us with a unique opportunity to analyse problems of measurement error in

both retrospective and panel recording of economic activity. Other longitudinal data sources often avoid data conflicts by eliminating overlap by design. In effect they hide the problem, while the BHPS has the means to assess its magnitude. In section 4 we discuss some aspects of the investigation of measurement error made possible by the design of the overlap. However, in general the algorithms used in the present exercise do not attempt to resolve errors.

1.3 The project: reconcile the data sources

The project this paper reports is the reconciliation of these several data sources into a single longitudinal record (in practice, several purpose-specific records) for ease of analysis by the end user. This involves combining those records designed to be directly combined (and resolving simple data errors therein) and resolving them with the two more-or-less free-standing long-term retrospective histories (which are more free to conflict in substantive terms).

We now go on to describe the outputs this project generates, in terms of specific files and their derivations. We then describe the rules and methods used to create the derived files, both in terms of a general specification and its actual implementation. We then consider the results of the exercise, in terms first of some simple overviews of the data in the reconciled files, in terms of where it came from and of its substantive content. We also address some statistical issues related to the bias introduced, or reproduced, by our procedures, and go on to consider the more general matter of measurement error.

2 Outputs

2.1 Output files generated

The outputs of this exercise are a number of files each representing a single reconciled work history. Several files are generated, partly to represent reconciliations of subsections of the data, and partly to serve different analytical needs. Their 'family tree' is presented in Figure 2, and Table 2.

The files are stored in two parallel forms, either 'calendar-' or 'episode-structured'. The 'calendar' files are status histories, presenting the respondent's status on a month-by-month basis; the 'episode' files report start and end

Demote this whole section to an appendix? Necessary roadmap for understanding what comes next.

dates (and state information) for all continuous periods in which the state is unchanging (see Appendices A and B below for more discussion).

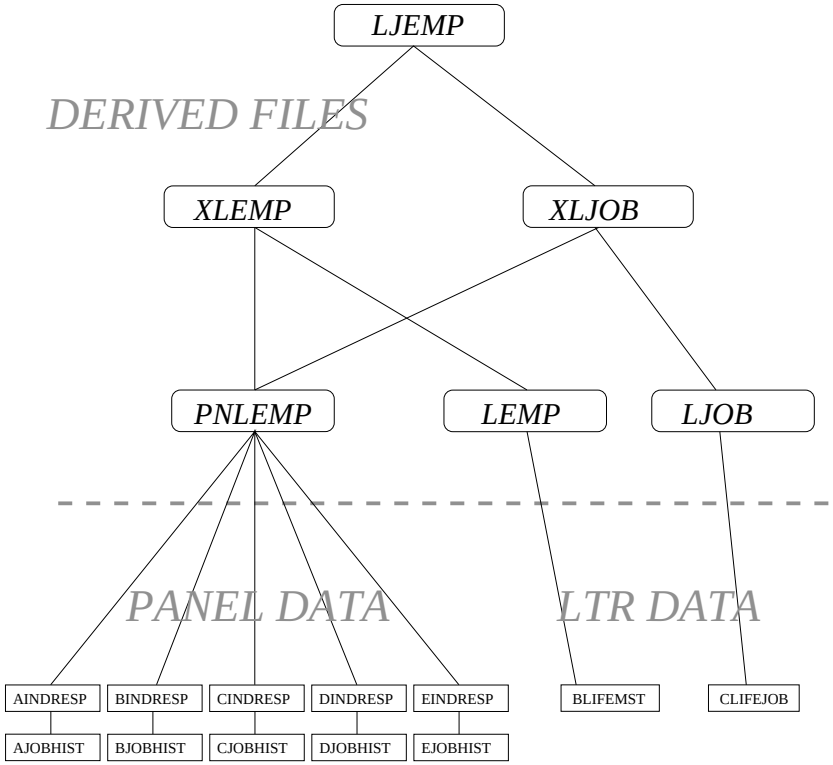

Figure 2: The six output files in relation to the main database files from which they are derived. *pnlemp* represents a combination of all *wJOBHIST*s and *wINDRESP*s into a single record, *lemp* and *ljob* are representations of *BLIFEMST* and *CLIFEJOB* respectively. *xlemp* and *xljob* are combinations of *pnlemp* and *lemp* and *ljob* respectively, extending the long-term retrospective (LTR) histories with analogous data from the panel. *ljemp* represents a particular combination of *xlemp* and *xljob*, giving precedence to *xlemp* and constituting a life-time work history with both employment-status and occupational data.

2.2 Derived files: Descriptions

Using the notation *file* to refer to both the calendar (*filec.sys*) and episode-structured version (*filee.sys*) of an output file, the six files generated are as described in the following sections (for more information on the relationship between the calendar and episode structured files, see Appendices A and B; for more information on variables, see Appendix C).

Table 2: Derived files generated by this project

		Source files	Output file names	
			Calendar	Episode
1:	Panel data	*wJOBHIST*, *wINDRESP*	*pnlempc*	*pnlempe*
2:	*CLIFEJOB*	*CLIFEJOB*	*ljobc*	*ljobe*
3:	Extended *CLIFEJOB*	*CLIFEJOB*, File 1	*xljobc*	*xljobe*
4:	*BLIFEMST*	*BLIFEMST*	*lempc*	*lempe*
5:	Extended *BLIFEMST*	*BLIFEMST*, File 1	*xlempc*	*xlempe*
6:	Combined and extended lifetime/panel employment status and job history	Files 3 and 5	*ljempc*	*ljempe*

In the following sections, brief descriptions of the files are given; details of the specific methods used to generate them are given in section 3.

2.2.1 Directly derived files

Five files are direct representations of files or groups of files in the main database.

panel-derived employment status and occupational histories

The *wJOBHIST* and *wINDRESP* files are here combined into *pnlemp*, a single longitudinal record covering the period of the panel study, extending backwards to the start date of the job held on Sept 1 1990 (a year before the start of Wave 1 fieldwork). Because this is *panel* data rather than *long-term retrospective* (LTR) data it is of a significantly higher quality – having better recall and substantially more detail – but is short-range and suffers from left-truncation (*i.e.*, it contains no information for experience before the job held on 1/9/90, which means that for most respondents it does not provide a complete career history).[4]

[4]The initial work on uniting *wJOBHIST* and *wINDRESP* was done by J. Gershuny.

Processed retrospective job history The *CLIFEJOB* file holds the data from the Wave 3 retrospective lifetime job history, but has some minor inconsistencies, especially with respect to dates. The majority of these inconsistencies have been dealt with (in a largely mechanical but sensible way: see section 3.2) to produce *ljob*.

Processed retrospective employment history *lempc.sys* is a calendar representation of *BLIFEMST*, with dates cleaned to the extent possible (*e.g.*, missing values replaced with imputed ones, using information from adjacent spells, and so on).[5]

Extended job history In principle *CLIFEJOB* covers every job since leaving full-time education until the job before the first employment-status spell described in *AJOBHIST* or *AINDRESP*. Thus by taking information from the main panel *CLIFEJOB* can be extended to cover the period up to the most recent interview date, creating *xljob*.

Extended employment status history *BLIFEMST*, the retrospective employment status history, contains information about employment status spells in the period since the respondent first left full time education, up to the Wave 2 interview date. This can be extended in the same manner as the job history, using panel information, to create *xljob*. It is worth making available both the basic and extended versions of both life-time histories, as they have a different logic at the latter end and this could conceivably be of interest to end users.

2.2.2 End-user specialised files

The end user's analytical interests will dictate which file to use, or possibly, which new sort of file can be generated. *ljemp*, described in the next paragraph, constitutes an example of a relatively specific combined file that unites the two long term retrospective files in a manner that suits certain analytical interests, giving priority to the employment status history. Other analytical interests would suggest other methods for combination, perhaps reversing the present mapping to give priority to the occupational history, using the employment status history to fill its gaps.

[5]Most of the work on converting *CLIFEJOB* into *ljobc* was done by Nick Davey; most of that on converting *BLIFEMST* into *lempc* by Mark Taylor.

Employment status history with mapped occupational data *ljemp* maps occupational data from the extended occupational history (*CLIFEJOB* plus panel-derived information from the panel) onto employed spells in the extended employment status history. This is intended to be useful where occupational information may enter an analysis of employment status as a secondary variable (*e.g.*, the analysis of the effect of prior occupational status on later moves between labour-market statuses).

3 Methods

The fundamental feature of the strategy adopted to carry out the reconciliation of the work-history data was to focus on *calendar* time rather than *events* or *episodes*. That is, while it is often better (in conceptual and data representation terms) to think in terms of episodes (*i.e.*, spells in a given state) or events (*i.e.*, transitions between states), we have chosen to think in terms of a month-by-month list of status information. This is a simplification, but one that pays off: an episode has at least two attributes (start-date and status), as does an event (date and outcome), whereas each entry in a status-calendar has only one value: the status. Thus it is more difficult to identify the same episode in multiple reports, as it has two degrees of freedom to vary.[6] It is simpler to ask the question, 'what was the respondent doing in this month?', than 'are these two episodes the same?'. With multiple sources for a given month, it is relatively simple to adjudicate between them.

Given that we can thus create an adjudicated status history, we can recreate an episode- or event-history by tracking changes in the status information. In order that successive distinct episodes with the same substantive state information can be distinguished (*e.g.*, a change from one employer to another, where SIC and SOC, and anything else measured, remain identical), we carry a spell-count indicator in the status information (see Appendix B).

Given this shift to a calendar perspective, we need two sets of rules: how to adjudicate between multiple accounts of the same period, and how to translate episode data into calendar format. For the former we apply one over-riding rule which deals with most situations: the earlier account, being

[6] Paull (1996) has examined the overlap period between consecutive waves in terms of spells, attempting to match reports of the same employment-status spell in each account. Even with a definition of a 'match' that was fairly permissive with respect to inconsistencies in starting date, a large proportion of spells did not match. The likelihood of being matched depended strongly on type of spell: approximately 90 per cent of employment spells matched, whereas only about 60 per cent of unemployment spells could be paired. Also, spells in the overlap period that persisted until the second interview date were much more likely to match than those ending in the interval.

nearer the time being described, is preferred.[7] In relation to the latter, we need various rules about dealing with successive dates (*e.g.*, where a start date is in the same month as the end date of the previous episode) and gaps.

Thus, the exercise started with a set of general rules, which were developed into a detailed specification. This was then implemented, mostly in SPSS, with varying degrees of faithfulness to the specification as practical issues began to impinge.

In the rest of this section we outline the general specification, and then describe the implementation. In the next section we attempt to assess the effects of this work on the data, in terms of fidelity to the original database, and in terms of the general problems of repeatedly collected longitudinal data.

3.1 The specification

The initial specification[8] had two tasks in mind: (i) to generate a cross-wave file representing the combined *wJOBHIST*s and *wINDRESP*s, and to use it to extend the two LTRs, *BLIFEMST* and *CLIFEJOB*, and (ii) to join the two extended LTRs into a single life-time record containing occupational and

[7] Because it avoids dealing with identifying different reports of the same spell in different records, this method will reproduce spurious state changes that are present in the data, typically occurring at the transition between *pINDRESP* and *qJOBHIST* or *qINDRESP*: when the first spell in *qJOBHIST* reports different state information (due to measurement error) but with the same (or an overlapping) start date as the spell reported in *pINDRESP*, there will appear to be a transition just after the *p* date of interview. It is not always impossible to resolve this data problem: first we have to decide whether the two accounts are referring to the same spell (*e.g.*, do their start dates agree sufficiently, and are the states closely enough related for the difference to be accounted for by re-description), and then, if they are, which account to believe with respect to the state information. But this is extremely difficult to do in an automated fashion, and raises problems of its own: for instance, can it be valid to accept *pINDRESP*'s state information as describing a period *after* the data was collected? Alternatively, can we justify overwriting Wave P's information about the respondent's current and recent status with information collected an extra year later? It is also extremely difficult to distinguish between an erroneous change in state information and an error in the start date of a genuinely new spell (really starting some time after the Wave P interview) such that it seems to overlap the *pINDRESP* spell.

But by choosing not to attempt to resolve this data problem, we choose to reproduce it in the combined work histories. This is a problem, particularly for analyses that focus on spells, durations or transitions, such as hazard models. In such contexts it is critically important to take account of the effect: one way of doing so is briefly discussed in section 4.3 below.

[8] The specification was drawn up by an *ad hoc* committee consisting of J. Gershuny, Mark Taylor, Nick Davey, Jonathan Scales and Brendan Halpin, with other members of the Research Centre on Micro-social Change being consulted.

employment-status information.

Because in some records dates are recorded to the nearest month, all dates are reduced to months.

3.1.1 Stage 1: production of *pnlemp*, *lemp* and *ljob* files

pnlemp: Cross-panel labour market history

1. Give priority to the first mention of a date.
2. Give priority to the end date of a spell over the start date of a subsequent spell, if necessary overwriting the beginning of the subsequent spell.
3. Always take *wINDRESP* as the last state in sequence (give it priority over *wJOBHIST*).
4. If the *qINDRESP* spell started before last September and differs from *pINDRESP* force the start of the *qINDRESP* spell to be on the month after the p interview.

ljob: Processed retrospective occupational history

1. Give priority to the first mention of a date.
2. Give priority to the end date of a spell over the start date of a subsequent spell, if necessary overwriting the beginning of the subsequent spell.
3. In gaps between jobs, code status by the reason the last job ended, but terminate the history at the end of the last job.

lemp: Processed retrospective employment-status history

Largely to be treated as *ljob*, but without the necessity of dealing with gaps, as *BLIFEMST* provides a continuous history.

Join *pnlemp* to *ljob*

1. If the start of *pnlemp* is before the end of the last *CLIFEJOB* spell, give priority to *pnlemp* account.
2. If the start of *pnlemp* is after the end of the last *CLIFEJOB* spell (*i.e.*, there is a gap) code as a non-employment period using reason the last job ended.

Join *pnlemp* to *lemp* Rules as for *ljob*/*pnlemp*.

3.1.2 Stage 2: merge wave B and wave C retrospective data

Here we made the operational decision to take *BLIFEMST* as the basic activity matrix and impute (add) *CLIFEJOB* occupational characteristics to employed spells:

Merge *ljob* and *lemp* (Imputing *CLIFEJOB* occupations onto the *BLIFEMST* calendar)

1. Define a *BLIFEMST* episode to be any continuous series of months in full-, part- or self-employment, including a mix of these states.
2. Wherever there is a valid *CLIFEJOB* occupation in an employed month according to *BLIFEMST*, transfer it.
3. Where there is a *BLIFEMST* employed spell where some or all months do not have corresponding *CLIFEJOB* occupational information, check through the *BLIFEMST* spell for *any* months with occupational data.
   - If any such months are found, expand them to fill the *BLIFEMST* spell, in proportion to the relative length of the job spells they represent,[9] recording the ratio between the months with data and the total spell length in an imputation flag vector.
   - If no such months are found, set the imputation flag vector to 0.

3.2 The implementation

The implementation was carried out almost exclusively in SPSS (with some use of SPSS's macro language, some SIR, and some automatic generation of wave-specific program files from a common template, using Awk). Insofar as possible it was driven by the specification as laid out above, but in practice some deviation from the specification occurred, and some situations not covered in the specification were dealt with (in particular in resolving conflicts in series of event dates within an individual's record).

3.2.1 Programs and assumptions

The structure of Figure 2 replicates the structure of the programming: At the first level we take files from the main database, and transform them into calendar format. In the case of *wJOBHIST* and *wINDRESP* this involves the extra steps of first combining files within each wave, and then joining the resulting wave-specific files into a single record, *pnlemp*. The second level is constituted by taking the outputs of the first level to generate extended versions of the long-term retrospective data files by combining them with

[9] That is, in proportion to their length within the bounds of the *BLIFEMST* spell.

panel information from *pnlemp*. The third level is constituted by combining the two extended long-term retrospective files into a long-term file with both occupational and employment-status information.

pnlemp The two steps involved in generating the cross-panel file are (i) to join *wJOBHIST* and *wINDRESP* for each wave, in the process putting them into calendar format (*i.e.*, each original variable is represented by a vector of variables with one value for each month from January 1900), and (ii) then to combine them into a single record for all five waves. Step 1 generates a calendar with data in the vectors from the date of the start of the first spell recorded in *wJOBHIST* until the date of interview, with the *wINDRESP* information on current status over-riding *wJOBHIST* information in the case where there is a conflict.

Employment status is used as the primary state variable, and this is mainly derived from *wJBSTAT*, supplemented with information from *wJBSEMP* and *wJBFT* in *wINDRESP*, corresponding to *wJHSEMP* *etc.* in *wJOBHIST*. That is to say, how the respondent describes his/her general employment-status situation is supplemented by answers to explicit questions about self-employment and full-time/part-time status.[10] A small amount of imputation of missing date information is carried out. Certain ambiguities in the main database can be resolved at this stage. For instance, *wJOBHIST* distinguishes between new jobs with a new employer and new jobs with the same employer, while *wINDRESP* does not. However, in general we know the reason the previous job was left (from *wJOBHIST*) and where this was a promotion we can identify the *wINDRESP* spell to be a new job with the same employer.

In combining the job history with the current-status information, the date of the start of the current status overrides the end date of the last spell in the job history.

In combining the resulting wave-specific calendars, a simple overlay is used: the earlier calendar is used up to its final month, and then the next calendar.

[10] It is important to note that *wINDRESP* provides other methods for determining current status, including the variable *wJBHAS* which records whether the respondent actually has a job, and variables which help to distinguish between non- and unemployment. A case can be made that such variables should be used in the present excerise, as they are more accurate than self-description alone; alternatively, all reports other than those in *wINDRESP* are self-description and it is in this sense more consistent. In the questionnaire, routing is dependent more on *wJBHAS* and its friends than on self-described status, which is why we see a certain amount of occupational information in the output files for people recorded as not working (see for instance section 4.1.2 and Figure 4).

lemp and *ljob* The conversion of *BLIFEMST* and *CLIFEJOB* into *lemp* and *ljob* is a mechanical process in principle. However, a substantial amount of imputation of missing dates occurs in both cases■. Imputation of missing dates is guided by dates of preceding and subsequent spells, where possible. Month values given to the nearest season are a case in point: where the season is winter, the respondent may mean the start of the year (Jan/Feb) or the end (Dec): where possible this is resolved by reference to adjacent spells (in the case of *BLIFEMST* only: this is Mark Taylor's code). The processing of *CLIFEJOB* involves a lot of manipulation of dates (switching suspicious start/end pairs, eliminating spells completely overlapped by other spells and so on).

■ The CLIFE-JOB work needs to be redone later

xlemp and *xljob* Extension of the calendar forms of the long-term retrospective files is straightforward: in the earlier period where the only information comes from the LTR, that is used; later where panel information also exists, the temporal priority rule means that *wJOBHIST* and *wINDRESP* will predominate if they contain non-missing information.[11] In practice this means that the ends of the long-term calendars are simply overwritten with information from the panel, with the level of detail in the panel information reduced to that in the retrospective files. In particular, *CLIFEJOB* does not record job shifts within employers, so such shifts reported in the panel files are ignored. Similarly, since the panel and LTR versions of certain variables do not have the same sets of categories, they have to be reduced to their largest common set.

ljemp The most problematic aspect to the creation of *ljemp* is deciding on the method for combining *xlemp* and *xljob*. By the time we get to *xlemp* and *xljob* the main problems of data oddities and incompatible definitions and variables have been largely dealt with. Thus the implementation of the specification described in section 3.1.2 above is fairly straightforward.

3.2.2 The specification and the implementation

The specification is skeletal, but its main elements and the focus on status calendars rather than event histories serve as the core of the implementation. The implementation does a lot more than the specification demands, and alters some of its requirements: for instance, the priority rule is adjusted

[11] Purely 'temporal' priority is in practice supplemented with a rule that *wINDRESP* is believed over *wJOBHIST* and they both are believed in preference to an LTR collected in the same interview. Temporal priority suffices for *e.g.*, *AINDRESP* to dominate over *BLIFEMST*.

to over-write missing data in a 'prior' source, and a substantial amount of manipulation of problematic dates goes on in the two LTRs. Similarly the specification does not address issues of compatibility of definition between data from different sources: variables do not have the same sets of categories in different source files, and spells do not have the same definition. In terms of the latter, *wJOBHIST* records job changes within employers, while *CLIFE-JOB* does not, and *wINDRESP* leaves it ambiguous: in extending *ljob* to create *xljob* we have to drop such job changes in the panel data, and in creating *pnlemp* we have to use information from *wJOBHIST* to resolve the status in *wINDRESP*.

Some variables turned out to be more important than expected: in particular the reason a job ended. It was anticipated that this would be important in *ljob* in characterising inter-job gaps by the reason the gap began, but as mentioned this information is also exploited in deciding whether the *wIND-RESP* job is with a new employer or not. Because it is useful, it is included wherever proper in the output files.

A detail the specification did not deal with is the determination of full-time/part-time status in the Wave 1 job history: the question was not asked in the interview. Our strategy was to look to *BLIFEMST* and to use its information for each month for which this information was missing. This works for a lot of cases, but a substantial number of person months remain in a special 'full-time/part-time undetermined' category. It should be noted that a side-effect is that the individual's first spell in *AJOBHIST* may be broken into sub-spells (*e.g.*, where *BLIFEMST* indicates a move from part-time to fill-time, or where it provides information for part of a spell but not all of it), with the result that *pnlemp* contains in a small number of spells which end before the Wave 1 reference period.

In the case of spells remain which in the 'full-time/part-time undetermined' category, it is up to the end user to treat them appropriately. In particular, in the episode-structured files, a spell in the unresolved category is not necessarily a real employment-status spell, but may represent only a portion of a spell for which this information is missing. A general strategy would be to look to adjacent employed spells, where full-time/part-time status should be known. A strategy to collapse artificial sub-spells into single spells would be to look to the data source information (*e.g.*, in *pnlemps* or the relevant vector in *pnlempc*) and to combine spells coming from the same *AJOBHIST* record.

3.3 Who's included

The specification says nothing about appropriate subsamples in the different files: *BLIFEMST* contains a subset of all wave 2 respondents, in turn a subset of those ever interviewed. *CLIFEJOB* contains a subset of those interviewed in *BLIFEMST*. Therefore the implementation includes all respondents present in any of the main database files in the combined output files. Thus, while *lemp* and *ljob* contain only respondents represented in *BLIFEMST* and *CLIFEJOB* respectively, respondents represented in any of the other output files may feature in as little as a single *wINDRESP*. Table 3 shows the participation patterns of the approximately 12,500 respondents represented in the reconciled files (*i.e.*, in *pnlemp*, *xlemp* and *xljob*; the files *lemp* and *ljob* contain 9,435 and 7,074 individuals). As the table makes clear, only (but nearly all of) those present at Wave 2 are present in *BLIFEMST*, whereas only those present in both Waves 2 and 3 are present in *CLIFEJOB*. When we look more generally at longitudinal participation in the panel, we find that somewhat more than half those who ever participate are present at all five waves. Approximately 1,300 miss only one wave, and there are about 2,600 new entrants at waves after the first.

4 Examining the output

In this section we consider the output from the reconciliation exercise, first taking a general overview of the information contained in the new files in terms of historical coverage and how the main database files contribute information to the combined files, and second looking at some summaries of substantive information. Thirdly, we look at the problem of seam effects, implicit in the data and reproduced by our methods, and discuss ways of dealing with them. Finally we consider issues of measurement error.

4.1 Some sample results

Four of the output files contain information from more than one source (*xljob*, *xlemp*, *pnlemp* and *ljemp*). To get an overview of the coverage of these files this section presents summaries of the information contained therein, in graphical form. Being longitudinal, the data in the files can be represented as time-series, and the summaries below consist of monthly series stretching back over the 20th century, showing the numbers of respondents in each file in particular statuses in each month.

Table 3: Individuals in reconciled files, broken down by interview status at each wave, and participation in the life-history components

Individuals present in at least one of the five waves are represented in the reconciled files. Only individuals present at Wave 2 contribute a *BLIFEMST* record and only those present at both Waves 2 and 3 contribute a *CLIFEJOB* record.

Full Interview in Wave					N	With LTR record	
1	2	3	4	5		*BLIFEMST*	*CLIFEJOB*
•	•	•	•	•	6658	6646	5871
•	•	•	•	–	473	472	412
•	•	•	–	•	105	105	81
•	•	•	–	–	386	385	337
•	•	–	•	•	164	162	
•	•	–	•	–	43	43	
•	•	–	–	•	37	37	
•	•	–	–	–	702	696	
•	–	•	•	•	148		
•	–	•	•	–	27		
•	–	•	–	•	7		
•	–	•	–	–	35		
•	–	–	•	•	47		
•	–	–	•	–	17		
•	–	–	–	•	17		
•	–	–	–	–	1042		
–	•	•	•	•	403	402	245
–	•	•	•	–	67	67	43
–	•	•	–	•	14	14	6
–	•	•	–	–	103	103	79
–	•	–	•	•	17	16	
–	•	–	•	–	13	13	
–	•	–	–	•	7	7	
–	•	–	–	–	267	267	
–	–	•	•	•	316		
–	–	•	•	–	86		
–	–	•	–	•	10		
–	–	•	–	–	186		
–	–	–	•	•	372		
–	–	–	•	–	208		
–	–	–	–	•	505		

Totals

In each wave					All[a]	Life-histories	
9908	9459	9024	9059	8827	12482	9435	7074

Note: (a) This number is the total for Waves 1 to 5 of persons with at least one full interview, and is thus the number of respondents represented in *pnlemp*, *xljob* or *xlemp*.

4.1.1 Comparing *xlemp* and *xljob*

We begin by comparing the two extended long-term retrospective files, *xljob* and *xlemp*. Figure 3 presents for each of these files the monthly totals of respondents for whom (i) we have any employment status information (*i.e.*, the number of respondents whose retrospective account starts on or before that month) and (ii) we have information indicating they were in employment.

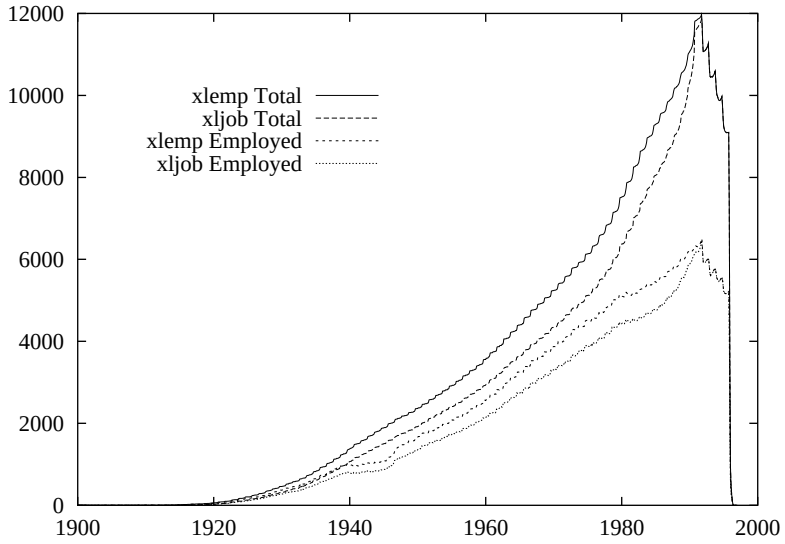

Figure 3: Comparing *xlemp* and *xljob*: Monthly numbers reporting any employment status, and reporting employment

The first thing to note is that we have information stretching back to 1920 and before: a small number of respondents report their first job as long ago as that (however, it is not until much later that we have significant numbers). At the other end of the century we see that for the period around the first wave of the survey (September 1991) we have employment status information for approximately 12,000 people. Indeed, this is higher than the number of respondents to Wave 1, because it also contains retrospective information for people entering at later waves. Up to this point, the job history and the employment-status history report different numbers, but after this point they coincide exactly. This is because from this point on, both retrospective data sources are replaced by analogous information from the panel (*i.e.*, *wJOBHIST* and *wINDRESP*). Before the panel period, numbers from the employment-status history are always higher (for all states and for employment) than the job history: this is primarily because there are more respondents to *BLIFE-MST* than *CLIFEJOB*. Despite this discrepancy, the shapes of the series are

substantially similar. The all-states curves are relatively featureless, and rise steadily as a function of the distribution of the starting-dates of the histories, while the employed curves rise more slowly. A notable feature of the latter is the 'bite' that World War II takes out of employment, as people enter the armed forces. Both employed curves also show a kink around 1980, as their slope suddenly lessens.▪

▪ This is a puzzle! J says improved recall/less missing data

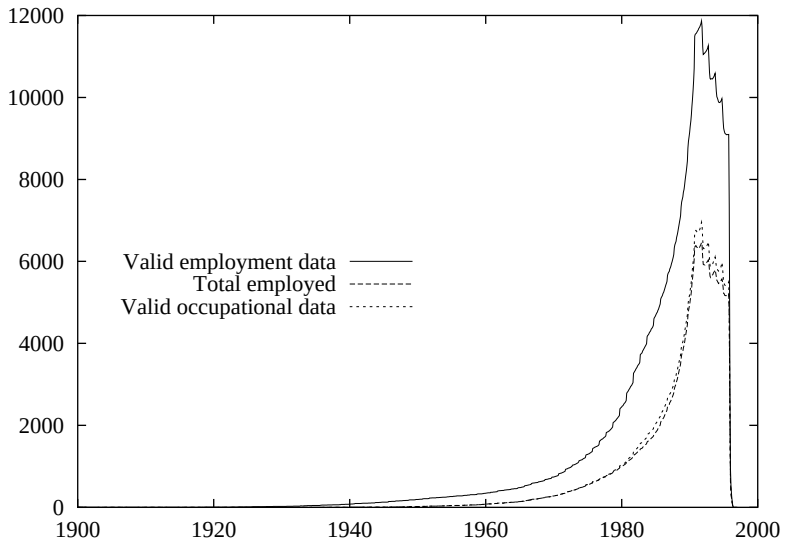

Figure 4: Information from *pnlemp*: Total, in employment, and with occupational information

4.1.2 Information in the cross-wave file, *pnlemp*

Next we turn our attention to what is in the combined panel files, *pnlemp*. Figure 4 shows overall information from this file, in terms of monthly series of total number, those reporting employment, and those reporting occupational data. This is on the same scale as Figure 3, which allows us to see how much more restricted to recent times this information is: not surprisingly, as the earliest episodes reported in this data set lasted until 1 September 1990. Nevertheless, we see that a portion of these earliest episodes stretches back a substantial period of time, to 1940 and before. From approximately September 1990, these curves are effectively the same as those in Figure 3, as they represent almost exactly the same data.

An interesting point to note is that more people report occupational characteristics than are coded as in employment. This is a feature of the main

database, where certain respondents are coded as not in employment but nevertheless report valid occupational information. This is partly due to our depending on self-described status, while the questionnaire routing exploits additional status questions (see section 3.2.1, especially footnote 10).

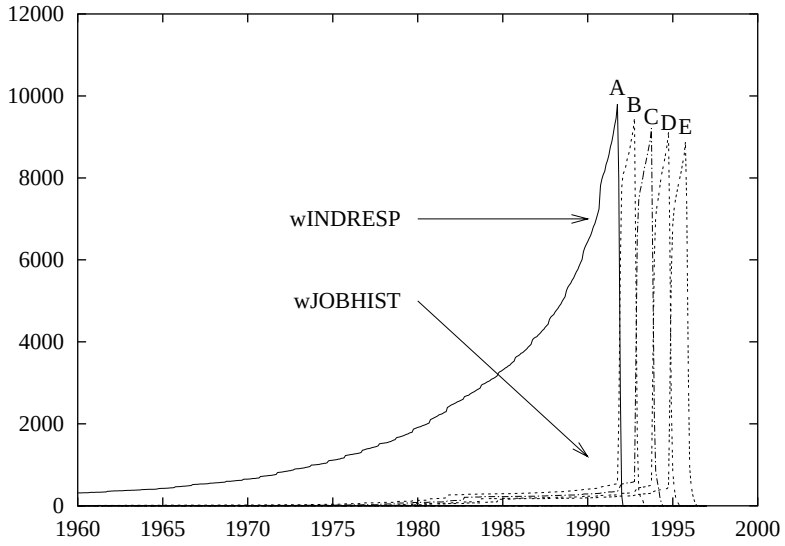

Figure 5: Data sources in *pnlemp*

Figure 5 traces the origin of each month's data in *pnlemp*. Each wave provides two potential sources (*wJOBHIST* and *wINDRESP*) and these are interleaved in a strict priority sequence, where earlier references dominate, on the grounds that they are closer to the period described. From the figure this pattern is clearly visible: the main source up to the end of 1991 is *AINDRESP*, with extremely small numbers of person–months being attributable to any of the other *wINDRESP*s until the end of Wave 1 fieldwork. Thereafter the later waves take over in turn, more or less twelve months at a time.

While each *wINDRESP* has a sharp peak at the date of start of fieldwork, its corresponding *wJOBHIST* peaks a year earlier, at 1 September the year before. It is notable how little person–time is accounted for by the *wJOBHIST*s, compared with the *wINDRESP*s. This should not be surprising, of course, as the inter-wave history will only cover episodes which have ended in the twelve or fifteen months leading up to the interview, which will constitute a relatively small proportion of spells.

As an aside, the falling – but flattening – curve implied by the peaks of the *wINDRESP* curves presents a nice picture of the combined pattern of inter-wave attrition and new sample entrants.

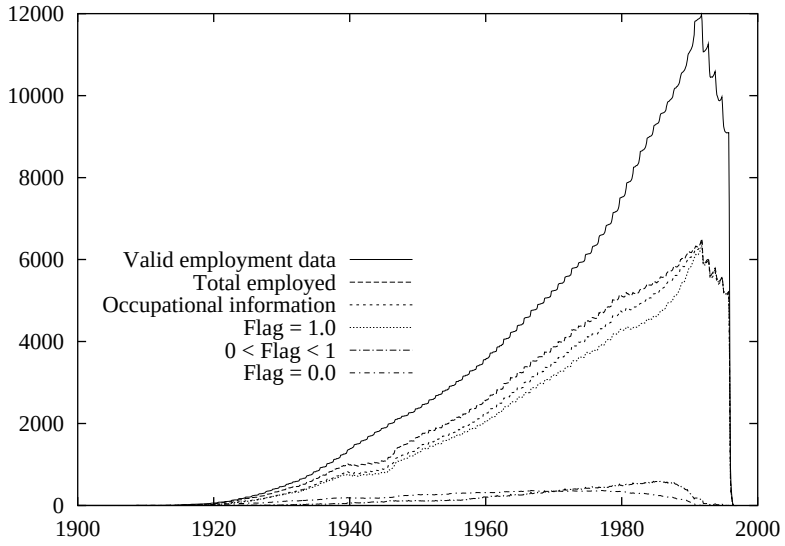

Figure 6: *ljemp* : combining disparate series

4.1.3 *ljemp* : combining disparate series

Figure 6 presents several series from *ljemp*, the file combining information from both extended retrospective data sets. The employment-status data is the same as that in *xlemp*, and the occupational data is mapped from *xljob*, onto employed spells in *xlemp* on a month-by-month basis. Where there is a simple mapping (*i.e.*, both series agree that the individual was employed on the month in question), the mapping is direct, and this is marked by setting an imputation flag to the value of 1.0. Where there is no occupational data for the month, the entire duration of the employed spell is searched, and any occupational data found is mapped to the whole spell, with the imputation flag having the value of the ratio of the number of months of occupational data to the full length of the spell. Where there is no occupational data found during the spell the flag defaults to 0.0.

Looking at Figure 6, we first see that the curves for the overall total ('valid employment data') and total employed correspond exactly to the curves from *xlemp* in Figure 3: they represent the same information. The most interesting comparison is between the curve for total employed and 'occupational information': these are the cases for whom we have successfully mapped job-history information. In the main it shadows the total-employed curve fairly closely.

In turn, we can break down the 'occupational information' figures by reference to the imputation flag. The graph shows series for 1.0 (*i.e.*, a

'perfect' match), 0.0 (a complete failure) and values between 0 and 1 (cases where there is month-by-month disagreement but with adjacent occupational information). Reassuringly, the curve for value 1.0 is quite close to the total, throughout. We can get a more detailed picture of the deviation from the other two curves: they are both fairly low and flat, with more complete failures than imperfect matches up to about 1970, and a preponderance of imperfect matches thereafter, up to the panel period.

4.1.4 Long coverage

From the foregoing we can see that the historical coverage of this data set is extensive. By using the retrospective life histories, we have information of substantial numbers of people back to 1960 or before. Of course, as a sample of the population in 1960 this is very age-skewed and suffers from survival bias, but it can nonetheless be useful. It must also be stressed that long-term retrospective data is strongly affected by recall problems, given the long recall period, and can by no means be considered equivalent in reliability to panel information (Dex, 1995; Paull, 1996; Elias, 1996).

4.2 Comparison of file outputs

Two comparisons of the data in the generated files are presented in this section, first, a comparison of durations in employment statuses in *lemp*, the transformation of *CLIFEJOB*, and *pnlemp*, the combined panel-derived file, and second, a comparison of information in the extended lifetime job history, *xljob*, with what the individual panel files have to say about the same time points.

4.2.1 Distribution of time

If we compare what *lemp* and *pnlemp* have to say about the lengths of spells in various employment statuses we find that the different data sources have substantially different patterns of duration. This is partly due to their reporting different data – the LTRs refer to spells not covered in the panel files – but it is also due to differences in reporting and recollection.

What is remarkable about the comparison between *lemp* and *pnlemp* reported in Table 4 is how much the difference for spell length varies across employment-status category. Employed spells are well more than twice as long in the LTR, but unemployment, retirement, maternity leave and family care are similar. Education is much less in the LTR, but that is probably due to the panel picking up people before leaving full-time education while

Table 4: Mean spell length in months: *lemp* and *pnlemp*, for selected employment-status categories

	lemp		*pnlemp*	
	Length	*N*	Length	*N*
Self-employed	105.68	1533	47.16	2642
F/t paid employment	124.81	14281	31.91	16015
P/t paid employment	65.91	4372	26.16	4419
Unemployed	13.34	4157	12.72	3468
Retired	120.21	1549	109.91	2935
Maternity leave	12.73	668	11.34	239
Family care	107.77	5015	91.24	2758
FT studt, school	24.00	788	66.93	2726
LT sick, disabld	67.67	497	46.53	820

the LTR excludes them by definition. Where LTR spells are longer, this is in part due to the omission of some (short) spells, increasing the length of adjacent spells.

4.2.2 Comparing outputs with *wINDRESP*

Table 5 compares *xljob* (*i.e.*, the extended occupational history) with the panel variables, *wJBSTAT*, *wJBSTATL* AND *wJBSTATT* from *wINDRESP*, which report the respondent's situation at respectively the date of interview, 1 September immediately before the interview, and 1 September a year (and more) before the interview. Since *xljob* is constructed from both the combined panel-derived file *pnlemp* and the LTR file *ljob*, this test serves as a check both on how the panel information is put together and on how it is integrated with the LTR data.

For the purposes of the comparison, the variables are reduced to an employed/non-employed distinction, and missing value cases are dropped. In most cases our temporal priority rule will mean that panel information rather than LTR information will be represented in *xljob* and therefore we should expect a good match with *wJBSTAT*, the day-of-interview status.[12] As the

[12]In particular, our temporal priority rule means that *wJBSTAT* is highly likely to be used for the information for the month of interview in the combined file, and will not be overwritten by later *wINDRESP*s. Also, in what could be considered a minor exception to this rule, *wINDRESP* always dominates over its contemporaneous *wJOBHIST*, with the important effect that if the last spell in *wJOBHIST* ends in the month of interview, it does *not* force the start date of the current spell to the following month, which is the practice within the *wJOBHIST*s and LTR files (see section 3.1.1, especially rule 2 for *pnlemp* / *ljob*).

Table 5: Agreement between *xljob* and *wINDRESP* for date of interview, 1 September a year ago, and 1 September at the start of fieldwork

	Date of interview (*wJBSTAT*)		Start of reference period (*wJBSTATL*)		Start of fieldwork period (*wJBSTATT*)	
	N	Percent	*N*	Percent	*N*	Percent
Wave 1						
Mismatch	3	.0	420	4.3	161	1.6
Match	9904	100.0	9418	95.7	9731	98.4
Total	9907		9838		9892	
Wave 2						
Mismatch	2	.0	494	5.3	143	1.5
Match	9644	100.0	8831	94.7	9307	98.5
Total	9646		9325		9450	
Wave 3						
Mismatch	1	.0	365	4.0	123	1.4
Match	9438	100.0	8650	96.0	8879	98.6
Total	9439		9015		9002	
Wave 4						
Mismatch	3	.0	431	4.8	136	1.5
Match	9333	100.0	8589	95.2	8906	98.5
Total	9336		9020		9042	
Wave 5						
Mismatch	0	.0	372	4.2	103	1.2
Match	9100	100.0	8391	95.8	8698	98.8
Total	9100		8763		8801	

Note: 'Agreement' is in terms of broad employment status; both sources must record either employment or non-employment. Missing values are excluded from consideration.

table shows, this is indeed the case with only a few stray cases in disagreement in any wave. However, the reliability of the other two *wINDRESP* measures is not so great: they both refer to the past, albeit *wJBSTATT* to the quite recent past. In the case of the 'year-ago' status, we find between 4 and 5 per cent of cases disagreeing, and in the case of 'this-September' status, around 1.5. The data in *xljob* will in nearly all these cases come from that wave's *wJOBHIST* or a prior wave's panel information, unless the prior information is missing in which case the data will be drawn from *CLIFEJOB*. In the case of conflicts with *wJBSTATT* (status this September) the *xljob* information is most likely drawn from that wave's *wJOBHIST* as there should be no conflict unless the person has changed state since then. A discrepancy of 1.5 per cent is small, but substantial enough when you consider that this is recall error over a period of about one to six months. Recall over the longer period (12 to 18 months) is noticeably worse: here it is much more likely the case the information we are comparing with *wJBSTATL* comes from the previous year's interview, rather than from the simultaneously collected *wJOBHIST* information.

4.3 Seam effects

Seam effects are an inevitable consequence of repeated collection of longitudinal data. Data collected at a later time point will either overlap with or abut data collected earlier and thus we are likely to see changes in values at the meeting point of the two data sources, the 'seam'. This may be because respondents report different substantive values for what might be the same episode, or report the starting date for a subsequent episode as lying before the date of the previous measurement point. Either way, there is a tendency for substantive transitions to appear to occur where there is a change in data source. It is necessary to have some means of taking account of these so-called seam effects.▪

▪ Is there a literature on seam effects?

4.3.1 Hazard model approach

Either sort of error will produce an apparent transition between two states exactly at the point where the source of the data changes. Thus one way of thinking about the seam effect is as a sharply bounded time-dependent covariate in a hazard modelling context: at the 'moment' of the data-source change, the hazard of transition can be expected to be raised. We can exploit this by fitting a hazard model with a time-dependent covariate which is zero nearly all the time, but for the 'moment' at which the data source changes (*i.e.*, for a month, given all our dates are collected or reduced to monthly),

has the value 1. On its own this variable can be expected to have a strong effect, simply because a proportion of apparent substantive transitions do take place at data-source transitions, but this is not interesting: instead, we consider what effect including this variable has on parameters estimates for other variables.

To this end we have estimated two Cox proportional hazard models on *pnlempe*▪, which has data from up to 10 different sources. Not all transitions between data sources should be associated with spurious transitions: the switch from *wJOBHIST* to *wINDRESP* is designed to occur at a real transition. All other transitions are registered in the model. All spells for all individuals are included (in this case, all periods within which employment status is constant, with job shifts between but not within employers constituting transitions).

▪ Analysis carried out on old version, which dropped within-employer job shifts.

Table 6: Proportional hazard model estimates with and without the control for the seam effect

	β	SE	p	β	SE	p
Age at start of spell	-0.015	0.000	0.000	-0.017	0.000	0.000
Sex (female)	-0.023	0.013	0.072	-0.027	0.013	0.035
Employment status						
– Self-employed (null)						
– Full-time employee	-0.024	0.023	0.316	0.012	0.023	0.617
– Part-time employee	0.278	0.027	0.000	0.379	0.027	0.000
– Unemployed	0.201	0.029	0.000	0.458	0.030	0.000
– Non-employed	-0.397	0.026	0.000	-0.204	0.026	0.000
Transition effect				1.563	0.013	0.000

Table 6 reports estimates for a simple model with and without the seam control. Age at start of spell, sex and a five-way categorisation of employment status constitute the substantive variables. The parameter estimates report the effect on the log of the hazard of a transition occurring: a positive parameter means the variable increases the likelihood of a transition. First we see that the transition effect itself is strong: the largest β by far, and clearly significant. (It must be borne in mind that this effect is 'instantaneous': it bears only on the month in which the data-source transition occurs and not thereafter, and thus its 'average' value is very low.) Thus, we see strong evidence that seam points are associated with transitions. Secondly, we see that the all the substantive parameter estimates change: age and sex become slightly larger, and the pattern for employment status shifts.

While the difference between self-employment (the reference category) and full-time employee status remains insignificant, the effects for part-time work and unemployment become substantially larger. Correspondingly, the large negative effect of non-employment becomes smaller.

We can conclude that seam effects are present in the data, and that they will affect model results if not taken into account.

4.3.2 Randomising dates

An alternative strategy for coping with seam effects, particularly when we believe the error is one of a too-early start date being reported for a genuine later spell, is to randomise the start date. That is, since we believe the the respondent's account for both interview dates, but disbelieve the second account's report of the start of the later spell, we infer that (i) a real transition took place, (ii) some time between the two interview dates. We can therefore impute a start date as a random date between the two dates in question. From the point of view of modelling durations this practice has good properties.▪

▪ This cries out for references

However, where the seam effect is due to a re-description of a state that has not in reality changed, the strategy is less attractive. Furthermore, since is constitutes an extra alteration of the data – one step further away from the main database – it is not implemented in the data for release. However, the end user can identify seam points from the information provided[13] and implement the randomisation independently, if the analysis requires it.[14]

4.4 Measurement error issues

Measurement error is a major source of inconsistency in all survey data, but it is especially obvious in panel data, where measurement is repeated. It is even more obvious where an attempt is made to construct longitudinal records, especially where overlapping information is collected, as in the BHPS. While it presents difficulties in adjudicating a single 'best' form of the data, this multiple overlapping measurement is a very important resource for analysing the nature, extent and effect of measurement error. Other designs of panel study which are strictly non-overlapping simply brush the conflict under the carpet, and lose the opportunity to assess the measurement error present.

Measurement error is a complex phenomenon, and a proper treatment is outside the present scope. However, it is nonetheless important to have

[13] Using the source vector in the calendar files or the source-information files that parallel the episode-structured files.

[14] SPSS code to implement this is available from the author.

some idea of the extent of measurement error – and the conflicts it creates – in the data we are processing. Therefore in the following sections we look at three analyses of its extent: first a brief overview of the extent of agreement in terms of broad employment status of the two long-term retrospective exercises, *BLIFEMST* and *CLIFEJOB*; second, of agreement between *BLIFEMST* and *AJOBHIST* in terms of recall of unemployment in the year for which they overlap; and third, an examination of spurious occupational transitions occurring because of poor occupational description and coding.

4.4.1 Agreement between *BLIFEMST* vs *CLIFEJOB*

While the domains addressed by *BLIFEMST* and *CLIFEJOB* are not precisely the same, there is a lot of common information. In particular, we can reduce each LTR to the following categories:

- full-time employment
- part-time employment
- self-employment
- non-employment

The latter category collapses unemployment and the various out-of-labour-force statuses in *BLIFEMST* and represents the gaps between jobs in *CLIFEJOB*. Viewing the data in this way we can address the issue of how much inconsistency there is between the two records, inconsistency which can be attributed to the differing recall period and the differing data collection method (in particular, collection of *BLIFEMST* was in terms of creating a continuous list of all statuses since first leaving full-time education, whereas was in terms of all jobs held, allowing gaps). This inconsistency is an estimate of measurement error (and especially recall bias) in the data, and serves as a benchmark against which to assess how well we would expect a merging of the two LTRs to function (as in *ljemp*). We have already seen how well the projection of *CLIFEJOB* data onto *BLIFEMST* employment spells fits, in terms of the imputation flag variable graphed in Figure 6 and discussed in section 4.1.3; this is a more general view of the same relationship.

By putting the two calendar versions of the data side by side (*lempc* and *ljobc*) we can create a vector of agreement in these terms for each individual for all months for which both records have valid data. We can then convert this data into a monthly time-series of the amount of agreement as a proportion of the total number of cases with data from both sources.[15] Figure 7 presents this time-series, using a strict and a loose definition of agreement.

[15]Thus, we implicitly disregard disagreement that takes the form of one record having information for a time that the other does not cover.

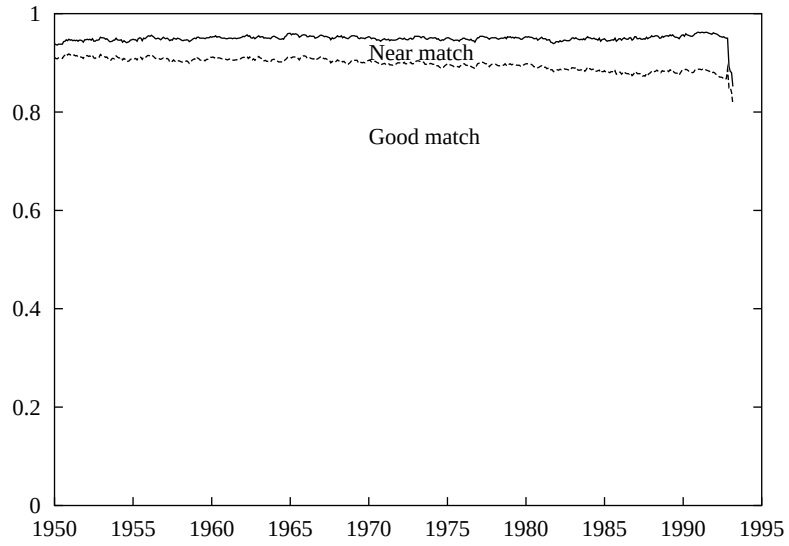

Figure 7: *BLIFEMST* and *CLIFEJOB* agreement in broad terms: proportion of cases for which there is monthly data where the two sources agree about the state

Strict agreement means that each LTR has the same value in terms of the four categories listed above; loose agreement collapses the three employed categories into one. Over most of the series presented, loose agreement rarely varies much from 95 per cent, which is a reassuringly high general level, though the absence of an increasing disagreement with increasing elapsed time may be puzzling. The level of strict agreement is also stable, but with some decline. Indeed, if we focus on the difference between the two series, we see a rise from about 3 per cent to 8 per cent of cases with valid data having imperfect matches, between 1950 and the early 1990s. Part of the good level of agreement between the two data collection exercises is undoubtedly due to the fact that at Wave 3 information from *BLIFEMST* was 'fed forward' through the interviewers: that is, interviewees were able to consult a printout of their Wave 2 lifetime employment status history when recalling their occupational history.

If we expect error to increase simply with recall period, the increasing gap between loose and strict agreement is very puzzling: why should disagreement (on whether one is self-employed, or full- or part-time employed, given one is working) increase as the recall period falls? Perhaps what matters is the relative difference in the recall periods: elapsed time since 1950 is practically the same for each LTR, but September 1991 is twice as long ago for *CLIFEJOB*

as *BLIFEMST*. It may also be that for periods well before the interview date, the respondent refers to a stylised internal account of his/her past, and that this account is relatively stable (over periods of the order of year) while not necessarily being particularly accurate.

4.4.2 Comparing of *BLIFEMST* and *AJOBHIST*: recall of unemployment

The period between September 1990 and September 1991 is covered by two sets of data: that collected in Wave 1 and represented in *AJOBHIST* and *AINDRESP*, and that collected in Wave 2, represented in *BLIFEMST*. In what follows we examine the twelve months from 9/90 to 8/91, for individuals who have both Wave 1 panel information and Wave 2 *BLIFEMST* information on all twelve months, from the point of view of the reporting of unemployment.[16] This is an interesting comparison, consisting of two retrospective accounts of the same period, but one being approximately a year later than the other (it also differs in that it is part of a life-long account, unlike *AJOBHIST*'s short period of reference). We might expect the recall of unemployment to deteriorate, given other evidence, but we would have no strong *a priori* views on the size of the deterioration.

Table 7 reports the level of agreement in recall of unemployment: it is rather low, with only approximately half of the unemployed months in either the panel or the LTR appearing in both. For instance, in September 1990, we find the panel information suggesting 287 (= 157 + 130) persons to be in unemployment, and *BLIFEMST* suggesting 254, but only 130 of these respondents register as unemployed in both databases. However, as we move forward in time, the agreement increases dramatically, with August 1991 showing a rather higher proportion of unemployed persons registering in both series.

Table 8 cross-tabulates time in unemployment in this twelve-month period in the two databases, to give another perspective. Respondents are concentrated in cells indicating zero unemployment in one or other source, and in the 10–12/10–12 cell (in the latter case, this is due largely to people in spells of long-term unemployment which span the entire period). The panel measure gives 242 people as having at least one month of unemployment for whom there is no unemployment in the *BLIFEMST* record, and correspondingly *BLIFEMST* indicates 130 people as experiencing unemployment that the panel does not report. On the other hand, 373 people are indicated as experiencing unemployment in both records, and the bulk of these are indicated as having shorter periods of unemployment in both records.

[16] Elias (1997) has conducted a similar analysis, comparing the two years of overlap between the first two waves of panel data and *BLIFEMST* in greater detail.

Table 7: Agreement on unemployment

	Employed or	Unemployed			Total
	non-employed	Panel only	LTR only	Both	
Sep 1990	7674	157	124	130	8085
Oct 1990	7674	153	119	139	8085
Nov 1990	7677	149	114	145	8085
Dec 1990	7671	155	109	150	8085
Jan 1991	7655	157	109	164	8085
Feb 1991	7642	160	100	183	8085
Mar 1991	7613	169	102	201	8085
Apr 1991	7603	168	101	213	8085
May 1991	7608	154	92	231	8085
Jun 1991	7600	155	100	230	8085
Jul 1991	7581	173	95	236	8085
Aug 1991	7580	161	100	244	8085

Table 8: Number of months unemployed, panel by *BLIFEMST*

Panel	*BLIFEMST*					
	Zero	1–3	4–6	7–9	10–12	Total
Zero	7340	49	15	19	47	7470
1–3	103	65	16	-	11	195
4–6	36	25	38	7	12	118
7–9	24	4	16	30	16	90
10–12	79	7	6	10	110	212
Total	7582	150	91	66	196	8085

When we look more generally at the agreement between these two records (*i.e.*, in terms of the categories: self-employed; full-time employee; part-time employee; unemployed; not in labour force; and missing) we see a more reassuring picture, largely because recollection of other categories, and in particular employment, is better than recall of unemployment. Depending on how we define a match (see Table 9) agreement is in the range 87.5 per cent to 93.3 per cent. What is interesting to note is that agreement between the records rises (almost monotonically) with the months: even over as short a period as this, we can see deteriorating recall.

Table 9: Overall agreement between Wave 1 panel information and *BLIFE-MST*, monthly from 9/90 to 8/91

Month	Precise match[a]	Looser match[b]	Loosest match[c]
Sep 1990	87.5%	90.2%	91.4%
Oct 1990	88.2%	91.0%	92.1%
Nov 1990	88.6%	91.4%	92.5%
Dec 1990	88.6%	91.5%	92.5%
Jan 1991	88.7%	91.6%	92.7%
Feb 1991	88.9%	91.8%	92.8%
Mar 1991	88.9%	91.8%	93.1%
Apr 1991	88.9%	91.8%	93.0%
May 1991	89.3%	92.2%	93.2%
Jun 1991	89.2%	92.0%	93.0%
Jul 1991	89.1%	91.9%	93.2%
Aug 1991	89.3%	92.1%	93.3%
Number of cases: 8085			

Notes: (a) match in terms of self-employed/full-time/part-time/non-employed/missing; (b) match in terms of self-employed/employee/non-employed/missing; (c) match in terms of employed/non-employed/missing. Only individuals with 12 non-missing values on both series are included.

4.4.3 Measurement error in job description: some evidence

Unpublished work has been carried out within the ESRC Research Centre on Micro-social Change at the University of Essex, looking at the issue of volatility of occupation coding between Waves 1 and 2 (Rose, Laurie and Perrin, 1994). For respondents who were employees in both waves, and had not changed their jobs, 32 per cent registered a change in their 3-digit SOC code, and almost 18 per cent changed their SOC major group (in similar work on the US Survey of Income and Program Participation, Kalton and McMillen (1986) found 40 per cent change at the three digit level).[17] A sample of the relevant questionnaires was examined and while there were various explanations for the discrepancies, the bulk of cases were caused by

[17]This comparison is complicated by the fact that at the time of the research, the Wave 1 occupational codes had been determined manually, whereas the Wave 2 codes were generated using the CASOC program (see footnote 3). This might be understood to introduce an extra source of inter-wave inconsistency. Since then, the Wave 1 codes have been re-generated using CASOC.

either different codings of what was essentially the same job description, or different descriptions of the same job.

5 Conclusion

The richness of longitudinal data goes hand in hand with an extra cost: it is more complex and raises problems of data manipulation and data conflict. This exercise with the British Household Panel Study represents an attempt to take a moderately complex longitudinal data set, containing panel, inter-panel and retrospective elements, and to reshape it in a form that is more conducive to analysis, while at the same time retaining as much as possible of the information in the 'raw' data (including the sprurious!). The main advantage of the exercise is that all data relating to a particular longitudinal framework (*e.g.*, all data relating to status during the panel period) is brought into one location, despite being recorded at multiple occasions and stored initially in multiple locations. This has undeniable advantages, in removing from users of the data set the onus of designing and carrying out a relatively difficult and error-prone data-organisation task, and providing them with a standard version, with the additional advantage of greater comparability with other researchers working on the same data.

While there is a certain amount of reconciliation of data conflicts involved in the exercise, this is mainly done mechanically, according to clear rules (such as giving priority to the report nearest to the event). That is, little data cleaning takes place: this is important in that it means the differences between this data set and the main data set minimised. The downside of this is that inconsistencies and measurement error in the data, such as seam effects and recall errors, are reproduced (and, given the greater ease of use, may be overlooked).

Appendices

A Calendars, Events and Episodes

All files produced are available in two forms, calendar and episode structured. The 'calendar' form is a state history, and the 'episode' form is an event history with some extra information about the spell following the event. Some terminology:

Calendar In 'calendar' files the life-history is represented as vectors of variables. For each substantive variable there exists a sequence of numbered ad-

jacent SPSS variables (*e.g.*, *SIC1* to *SIC1152*) where each SPSS variable represents the state of the substantive variable in each month from January 1900 (*i.e.*, *SIC1* contains the value of SIC in January 1900, *SIC1000* that in April 1983, *SIC1120* April 1993, *SIC1152* December 1995, and so on). These are easy to use and especially to conceptualise, but are extremely big (but compress very well using programs like pkzip or Unix compress) and are very slow to process.

Event An event is defined as a change of state, located at a particular moment, for instance the start of a job. If we know the date of every change of state of interest, we can find out the state at any time. Thus an 'event history' or a list of all state changes and their dates can be turned into a calendar with no loss of information, and vice versa (with some provisos; see below). Pure event histories are in general much smaller than calendars, and are much faster to process (but are less intuitive to manipulate). Pure event histories are more general and are more suitable for further processing (for instance, event histories relating to separate domains can be interleaved to generate a single history relating to the combined domain; alternatively certain types of state-change may be defined as not interesting and simply dropped). They are also in a form more suitable for certain types of analysis, *e.g.*, duration or hazard modelling.

Episode If an event is an instantaneous change of state, an episode (or spell) can be defined as a period of time bracketed by two consecutive events (where observation is also considered an event, generating a censored episode, that is, one whose true end date and outcome state is not known). Equivalently, an episode is the whole of a period of time in which the variables of interest remain constant.

In terms of data records, what differentiates an episode from an event is that the episode record also contains information about the duration and end-state (and end-date) of the spell following the event, whereas the event record contains only the date of the entry to the spell, and the value of the state variables. In practice it can be very convenient to know the duration and outcome (censoring, or transition to a new state) of a spell. However, if further processing is to occur (interleaving of another history, or deletion of certain classes of transition) then the end-date, duration and outcome information becomes meaningless and must be re-computed.

In principle these three file formats contain exactly the same information, and in practice each can be derived from either of the others (differences between the latter two are slight, but it takes a moderate amount of processing

time to move between calendar and event/episode formats).

B Calendar to episode conversion

The episode-structured files are generated in an entirely mechanical manner from the calendar files. Essentially, we define an episode as a period during which all variables of interest remain constant. Thus a change in the value of any one of the variables constitutes an event marking the start of a new spell (and the end of the old). However, unless the set of variables of interest is chosen with care, it is possible for two consecutive spells to be collapsed into one if all the monitored variables remain the same: for instance, two consecutive jobs with identical attributes (and no gap between) will look like one. In order to catch all transitions it is necessary to monitor a variable which tracks the spell sequence number: this is incorporated in the vector tracking the data source, for all sources where it is relevant (*i.e.*, *wJOBHIST*, *CLIFEJOB*▪ and *BLIFEMST*). By monitoring sequence number, and transitions between *wJOBHIST* and *wINDRESP*, we can recover all episodes encoded in the main database (transitions from *pINDRESP* to *qJOBHIST* or from the long-term retrospective records to the panel do not indicate necessary transitions, but may do so, often due to seam effects; see section 4.3).

▪ Not implemented yet

C Files and variables

In calendar files variables are represented as vectors of the form *ABCD1* to *ABCD1164* representing months from January 1900 (= 1) to December 1996 (= 1164). The corresponding representation in the episode-structured file is *STABCD*. Episode files additionally have start date, end date, duration and spell sequence number (not the same as the sequence number from any of the main database files). They also have *FIRST* and *LAST*, indicating a respondent's first and last episodes.

References

Dex, S. (1995), 'The reliability of recall data: A literature review', *Bulletin de Méthodologie Sociologique* (49).

Dex, S. and A. McCulloch (1997), 'The reliability of retrospective unemployment history data'. Unpublished paper, Judge Institute for Management Studies, Cambridge.

Elias, P. (1996), 'Who forgot they were unemployed?'. Unpublished paper, Institute for Employment Research, University of Warwick.

Elias, P. (1997), 'Reduncancy and reskilling'. Seminar paper presented at ESRC Research Centre on Micro-social Change, University of Essex.

Elias, P., K. Halstead and K. Prandy (1993), *Computer Assisted Standard Occupational Coding*, HMSO, London.

Kalton, G. and D. B. McMillen (1986), 'Nonsampling error issues in the Survey of Income and Program Participation', SIPP Working Paper 8602, US Bureau of the Census, Washington, DC.

Paull, G. (1996), 'Dynamic labour market behaviour in the British Household Panel Survey: The effects of recall bias and panel attrition'. Unpublished paper, Institute for Fiscal Studies.

Rose, D., H. Laurie and K. Perrin (1994), 'Wave 1/wave 2 cross-wave report'. Internal report, ESRC Research Centre on Micro-social Change.

Taylor, M. (1996), 'British Household Panel Survey User Manual', User documentation, ESRC Research Centre on Micro-social Change.