Does data editing really matter? The relevance of within and post-interview data editing in the IAB-ALWA study¹

Katrin Drasch²

Abstract

Retrospective life course are extremely valuable for analysis. Unfortunately, quality requirements for this kind of data are high. Therefore, data have to be edited often with a time consuming process. Notwithstanding such enormous efforts, survey methodology has up until now paid little attention to the subject of data editing as well as to other post-data-collection procedures. I aim at filling this gap. In the following, I use the IAB-ALWA study that collected data on life courses of a nationally representative sample of 10,000 respondents. It will be examined whether such a design is beneficial from the perspective of data quality and efficiency. The results show that despite corrections the results are robust concerning different specifications of the dependent variable stemming from different stages of the editing process.

Keywords: Retrospective data collection; editing

¹ A similar version of this paper appeared as chapter of my dissertation Drasch, K. (2013). The re-entry of mothers in Germany into employment after family-related interruptions: empirical evidence and methodological aspects from a life course perspective. Bielefeld: Bertelsmann. I gratefully acknowledge permission of the Institute of Employment Research (IAB).

² Dr. Katrin Drasch, Friedrich Alexander University Erlangen-Nuremberg, Institute for Sociology, Chair for Methods of Empirical Social Research, Kochstr. 4, 91054 Erlangen, Germany; email: katrin.drasch@fau.de

1. Introduction

Retrospective life course data contain information on the characteristics and the dating of past episodes. Thus, such data are extremely valuable for the analysis of event histories, life courses, and long-term social change (Blossfeld, Golsch and Rohwer 2007). Unfortunately, quality requirements for this kind of data are high: they should be as complete and consistent as possible in order to reduce total survey error (Biemer 2010) and ensure the quality of the research process (Hillmert 2002). Therefore, in previous studies such as several projects related to the German Life History Studies (GLHS), life course data were edited extensively after the interview was conducted. A team of data editors performed extensive checks concerning temporal and content-related consistencies. This resulted in a process which often lasted for several years. The editors had to go back to the tape-recorded interviews or had to phone the respondents again, in order to obtain consistent life history data. For examples compare Brückner (1995), Brückner, Hoffmeyer-Zlotnik and Tölke (1983), Tölke (1982), or Hillmert (2002). These processes were often way more time consuming than the entire survey process itself had been, but were nevertheless considered worthwhile for the sake of the resulting higher quality of the data (Hillmert 2002).

Notwithstanding such enormous efforts, survey methodology has up until now paid little attention to the subject of data editing as well as to other post-data-collection procedures. Groves et al. (2009), for example, dedicate merely three pages to data editing in general (pp. 345-347). Their definition of editing refers to the ‘alteration of data recorded by the interviewer (...) to improve the quality of the data’ (p. 345). However, their explanation mostly concentrates on edition procedures for cross-sectional studies such as range, ratio, balance, or consistency edits. Comparable standards for the temporal and content-related consistency of longitudinal data are only mentioned briefly – which is in stark contrast to the fact that Grooves and his colleagues generally describe longitudinal data to be the kind of survey-data that have the highest requirements for data editing. The summary of Groves et al. (2009) refers to the following: ‘Editing systems will change as computer assistances moves to earlier points in the survey and becomes integrated with other steps in the survey’ (p. 347). They conclude that then eventually editing after data collection itself will decline.

In the following, a survey which makes use of a design that uses computer assisted editing procedures will be described. Furthermore, it will be examined whether such a design is really beneficial from the perspective of data quality and efficiency. The IAB-ALWA study collected

data on life courses of a nationally representative sample of the German resident population. In total, 10,000 respondents were interviewed on behalf of the Institute for Employment Research (IAB) between August 2007 and May 2008. The survey was administered by telephone (CATI) and systematically collected retrospective information in major life domains such as schooling, vocational training, employment or unemployment (Antoni et al. 2010; 2011; Gilberg et al. 2011; Kleinert, Matthes and Jacob 2008).

This chapter is structured as follows: first, the different steps of data collection and editing used in the IAB-ALWA survey will be described. Then, the results of the second step of data editing will be examined empirically. For both steps a specific example is chosen: the editing of parental leave episodes and periods of labour market inactivity due to childrearing. Collecting data on these episode types was especially error-prone. Reasons for this are, for example, recall-errors or mistakes made by interviewers while recording the information provided during the interview process (Matthes et al. 2012). Picking a specific example has been suggested as a strategy how to address the importance of the impact of data editing by Hillmert (2002). Thus, I will illustrate the effects of data editing procedures using the analysis described in Drasch (2011a) as an example. These analyses will be repeated with an unedited version of the dependent variable – the duration of family related employment interruptions of women. Then, the results using edited and unedited data will be compared in order to get an impression of the effects of data editing.

2. Data collection and editing in ALWA

2.1 First step– within-interview data collection and editing

Life history data in IAB-ALWA survey were collected in a modularised mode for each life domain separately. This means that during the interview each respondent had to go through his life repeatedly, during each round providing information on events in one of the various domains of interest. In a first step, after data collection itself was completed, the interviewer guided respondents through a computerised and standardised data revision module. This module combined data previously collected in different life domains and checked them for temporal inconsistencies such as gaps and illogical overlaps. Previous research has shown that applying such a module helps to improve completeness and consistency of the retrospective data (Drasch and Matthes 2009).

In the ‘children and parental leave module’ for example, information on all children who ever lived in common household with the respondent was covered. Within this module, personal information on these children was collected as well as information on episodes of parental leave. For the latter, respondents had to answer whether they were eligible to a period of parental leave and following that, whether they actually went on parental leave for the respective child. When doing so, the respondents also had to report the starting and ending dates of the respective leave-episodes. However, this process involved several problems: First, respondents also reported the statutory two-month period of maternity protection as a period of parental leave, although interviewer instructions explicitly forbid doing so. Second, and more important, respondents wrongly reported the entire period of labour market inactivity that followed childbirth, even though they had been asked only for the time of parental leave, which is defined as the period during which they are guaranteed the right to return to their previous workplace. Third, respondents with several children had problems to assign a particular episode of parental leave to a specific child and instead often reported the entire leave-period (consisting of several episodes) for the first child and then additional episodes for the second child and for subsequent children. This resulted in several overlapping or parallel episodes of parental leave and in many cases in more episodes of parental leave than children reported living in the household (Matthes et al. 2012).

In the ‘gap module’, which was part of the data revision module, temporal gaps containing no activity at all were completed with information. Here respondents could report periods of labour market inactivity due to child care obligations and times as homemaker. This also included start and end dates of the respective episode. For a more extensive description of the modules and the within-interview editing compare the ALWA code book (Matthes and Trahms 2010) or the report on data editing (Matthes et al. 2012).

2.2 Second step– post-interview data collection and editing

In a second step, after the data editing within the interview itself, the data were checked again when the interview was finished. This was done by trained editors in a time consuming and costly data editing process comparable to the one used in the GLHS. However, in contrast to the former, data edition procedures did not include contacting respondents again or listening to audio recordings of the interviews (Matthes et al. 2012).

A team of ten editors was hired to perform consistency checks in order to improve data quality even further. The editors were undergraduates most of them studying sociology / social sciences

or a related subject. The data editors received two days of training and were supervised by a team of two full-time employed supervisors at IAB. All editing rules imparted during the training as well as additional information were written down in an editing manual. In addition, the editors were trained to use the computer-tool ‘ALWAEdi/Patch Tales’ (Künster 2008), assisting them in the editing process. This tool graphically displayed the life courses of individuals, which made it easier to identify temporal as well as substantial inconsistencies in the data. All in all, this editing procedure lasted almost one year and included the following steps: First, cases that only included unproblematic episodes in all different modules were identified and subsequently dropped for they required no further editing. In a second step, the remaining cases were then usually resolved applying the standardised editing rules which were laid down in the editing manual. In rare cases, the editing manual had to be extended when new problems were identified that could not be resolved with available standardised routines. Cases that remained unresolved after this step were finally resolved in a team of two editors and discussed with the editing supervisors. For a more detailed description of the editing process and the rules applied compare Matthes et al. (2012). Table 1 describes the cases processed in each of these editing steps.

Table 1 Descriptive statistics of the post-interview editing process

	Number of cases (N)	Per cent (%)
No editing	6,311	62.0
Standardised editing	2,651	26.1
Single case editing	1,215	11.9
	10,177	

In 62 per cent of the cases it was not necessary to perform any post-interview editing. This shows that using the data edition module for data editing within the interview (cf. Drasch and Matthes 2009) is relevant. However, around 38 per cent of the cases had to go through a subsequent editing step, 26 per cent through standardised editing and 12 per cent through single case editing. This also points out to the importance of these later editing steps.

During the post-interview editing process both modules - the ‘children’ and the ‘gap module’ - were examined simultaneously. According to the editing manual, episodes of parental leave were shortened to the maximum duration of parental leave of 38 months (36 months plus two months maternity protection). This is an example for ‘implicit editing’ (Groves et al. 2009: p. 347), meaning that logically deduced rules were applied. In this case the rule was deduced from

the fact that the maximum duration of parental leave in Germany did at no point in time exceed 36 months. The shortened periods were either added to a gap module or a new gap episode was created. Furthermore, parallel and overlapping episodes were edited. This procedure led to only one inactivity episode per month with a new episode or parental leave starting with the birth of a new child. This is an example for an ‘explicit editing rule’ (Groves et al. 2009: p. 347) meaning that the data must look like that for each respondent.

Due to these changes which mainly harmonised information over different respondents, I expect the following consequences in terms of overall data quality: first, the results of empirical analyses should be robust with regard to the different editing steps. This is because the major aim of the editing procedures was to correct the misunderstandings of the respondents. However, I expect that the size of the coefficients becomes more stable with each subsequent editing step. Moreover, I would also assume that standard errors are smaller for the final data, because less ‘noise’ – meaning less reporting errors – is left.

3. Empirical Part

3.1 The second step editing of parental leave episodes and periods of labour market inactivity – descriptive findings

The descriptive findings compare the data at three different stages of the edition process: first, I will examine starting and ending dates of the data before they have been checked in the data revision module. However, even at this stage prior editing steps had to be applied before using the data: first, where respondents could not name the particular month but only the season an event took place, one particular month was chosen; and second, when respondents could remember only the year of an event, the missing information on the month was set to the middle of a year. This was necessary in order to lose not too many episodes due to missing information regarding the month an event took place. Afterwards, the data were transferred to a linear time format, indicating the number of months that had elapsed since January 1969.³

In a second step, the data were examined after the within-interview editing and the data revision module had been completed. Also in this case, information on the month and year an episode started and ended was transferred into a linear time format thus that dating of an episode only

³ This corresponds to the Stata time format.

involved two variables – start and end date – instead of four (start month, start year, end year, end month).

Third, the data were examined after post-interview editing (also including a linear transformation of the dates). Table 2 compares the frequencies of the parental leave and gap episodes for the different editing steps.

Table 2 Comparison of original, data revision and edited data frequencies

	Comparison original data and data revision module data		Comparison data revision module data and post-interview edited data		Comparison original data and post-interview edited data	
	N	%	N	%	N	%
Same number of episodes in both datasets	6,439	81.6 %	5,843	66.7 %	5,834 (4,658 parental leave + 1,176 gap episodes)	68.2 %
More episodes in the better edited dataset	197	2.5 %	793	9.1 %	1,909 (1,609 parental leave + 300 gap episodes)	22.3 %
	6,636		6,636		7,743	
More episodes in the less edited dataset	1,250	15.9 %	2103	24.1 %	809 (316 parental leave + 493 gap episodes)	9.5 %
	7,886		8,739		8,552	

When comparing the original raw interview data (i.e. before the data revision module) to the data obtained after application of the data revision module, one can see that the number of episodes for one person was identical at both stages in as much as around 82 per cent of the cases. This number drops to around two third of the episodes being the same when comparing the data of the second to those of the third step. This indicates that the post-interview editing process had a much higher influence on the number of episodes than data revision during the interview. Due to a better assignment of the parental leave episodes to individual children, however, around one quarter of the episodes were additionally created during the post-interview editing process. The deletion of episodes obviously played only a minor role in the post-interview editing process but a more important role in the data revision module. The total number of episodes increased because gap episodes were added in the data revision module. The number diminished again during post-interview editing due to the deletion of parallel and overlapping episodes.

However, in total this editing process should not have affected to overall analyses of periods of family related periods of labour market inactivity because only leave episodes and homemaking episodes were re-assigned. The duration of parental leave episodes and their frequency should in turn have been affected more seriously. In the following section this will be examined in more detail.

Table 3 Comparison of start and ending dates of the original, data revision and edited data

	Identical start date Parental leave episodes	Identical end date Parental leave episodes	Identical start date Homemaker episodes	Identical end date Homemaker episodes
Original data and data revision data compared	97.6 %	98.6 %	90.7 %	90.4 %
Data revision data and edited data compared	79.5 %	77.4 %	75.7 %	83.6 %
Original data and edited data compared	81.9 %	68.2 %	75.3 %	62.9 %

Table 3 compares the start and ending dates of the original episodes, the episodes after the data revision module and after post-interview editing of matched episodes. The episodes were matched using a generated identifier variable that sorted the episodes with respect to the unique person identifier and the start and end dates. When comparing data from post-interview editing with those from within-interview editing, one can see that an overwhelming majority of the episodes was not changed with respect to their starting and ending dates. The number was higher again for parental leave episodes than for homemaker episodes. When comparing the original data to the final edited data one arrives at the conclusion that around 80 per cent of the parental leave periods and 75 per cent of the homemaker episodes had already identical start dates at this editing stage. The numbers are somewhat lower for the end dates, 68 and 63 per cent respectively. So, a maximum of 25 or 37 per cent of the episodes were changed with respect to the dating of the episode. Due to almost identical dates the numbers are similar when comparing data from the data revision module with the final data.

Parental leave episodes	Homemaker episodes
--------------------------------	---------------------------

Mean duration original data	23.3 months	61.8 months
Mean duration data revision data	19.6 months	60.3 months
Mean duration edited data	20.1 months	68.4 months

Table 4 Mean durations of parental leave and homemaker episodes at different stages of the editing process

Table 4 displays the average durations of the different episodes. The parental leave episodes were longer in the original data than they were in the data revision data and the post interview edited data. In turn, the average duration of homemaker episodes was shorter in the original data than at the two other stages of the editing process. This means that the original durations significantly differ from the durations of the final data. This finding can be attributed to the editing process and the clearance of overlapping and parallel episodes in this editing stage. In contrast to that, there are no significant duration differences as tested with t-test when comparing the original data and the data from the data revision module.

3.2 How does data editing affect multivariate analyses?

The previous results have shown that the changes, especially the changes from the raw data to the data from the data revision module, had a striking impact on the frequencies as well as on durations of episodes. This violates one of the rules presented in Groves et al. (2009: p. 346) which claims that the frequencies should be maintained when editing data. But it is questionable whether this is also valid when editing longitudinal data and temporal information on episode dating in particular. Therefore, I replicate the analyses presented in Drasch (2011a) with the raw data including the season and missing value data correction and compare the results to those obtained using the final data after post-interview editing. A comparison of the data after within-interview editing with the final data or the raw data is difficult to establish because the data structure is very different compared to the other two datasets. More precisely, the original children and parental leave module was split up into two different modules which was not the case for the data editing module. However, the analyses presented here compare the two stages of the data edition process that should display the maximum difference. First, I present Kaplan-Meier estimators distinguishing the dependent variable by cohort.

Figure 1 Kaplan-Meier failure curves edited and unedited data

The results in Figure 1 show that there are similarities as well as differences. Similarities for all three cohorts (until 1986, 1992-2000 and 2001-2006) can be found regarding the shape of the curves. Furthermore, their overall probabilities to have returned to employment after 36 months are similar, as well. The major difference is that the 1986 to 1991 cohort is more strongly affected by the data editing process than the others and the pattern is very different for the two editing steps. However, due to the limited number of cases these results should not be over-interpreted, especially for longer interruptions.

Second, I estimate an identical model specification with an unedited but date corrected version of the dependent variable (the duration of a family related employment interruption generated by adding parental leave and homemaker episodes) and the final linearised start and ending dates which are available to the public in the ALWA scientific use file (cf. Antoni et al. 2010; 2011).

Third, I estimate the coefficients using seemingly unrelated estimation and test the coefficients as well as the overall model for significant differences using a Wald (F-) test. This can be done because the samples of both analyses are overlapping.

Table 5 Discrete-time event history models on the likelihood to re-enter the labour market including results from seemingly unrelated estimation

	Original data	Edited data	Significant differences (F-tests)
parental leave stage (ref. around expiration of parental leave)			
before expiration of parental leave	0.150*** (0.013)	0.162*** (0.013)	**
after expiration of parental leave	0.444*** (0.060)	0.510*** (0.061)	
educational attainment (ref. university degree or similar)			
no vocational degree / no schooling degree and apprenticeship	0.409*** (0.070)	0.344*** (0.067)	
lower schooling degree and apprenticeship	0.732* (0.107)	0.631** (0.107)	
higher secondary schooling degree and apprenticeship / higher vocational training degree	0.695** (0.093)	0.647* (0.100)*	
educational attainment in partnership (ref. homogamy)			
hypergamy (woman lower than partner)	1.114 (0.147)	1.237 (0.186)	*
hypogamy (woman higher than partner)	0.873 (0.101)	0.986 (0.131)	
no partner	1.703*** (0.220)	2.631*** (0.354)	
child birth cohort (ref. 1979-1985)			
1986-1991	0.803 (0.123)	0.625** (0.092)	
1992-2000	0.727 (0.128)	0.627** (0.111)	
2001-2006	0.828 (0.169)	0.778 (0.165)	
Age (in years)	1.055*** (0.014)	1.049*** (0.014)	*
Age of youngest child (in years)	0.977 (0.053)	0.524*** (0.039)	***
Age of youngest child squared	0.999 (0.003)	1.034*** (0.004)	***
maximum number of children in household (ref. one child)			
two children	0.671*** (0.053)	0.694*** (0.053)	
three and more children	0.475*** (0.070)	0.489*** (0.069)	
labour force experience (in years)	1.019* (0.009)	1.011 (0.009)	
unemployment experience (in years)	0.837** (0.056)	0.695*** (0.050)	

regional type (ref. urban/agglomeration)			
rural	0.864 (0.090)	0.917 (0.109)	
not defined	1.027 (0.145)	1.125 (0.179)	
regional unemployment rate	1.004 (0.025)	1.034 (0.024)	
Observations	90,887	118,368	Joint test of significant Differences (excluding time-dependency) chi2(20) = 93.15
AIC	12376.9	15076.0	

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

The results in Table 5 show the following: The odds ratios obtained from both estimations are similar with only some minor exceptions. These apply to the cohort variable, for example. The significance levels are similar, too, with only some minor deviations of one significance level. In essence, irrespective of the edition status of the data used, one would arrive at similar conclusions to be drawn and also the interpretations of the coefficients do not differ substantially. When looking at the results from the seemingly unrelated estimations⁴ one does not find significant differences for almost all independent variables, the age of the youngest child (linear and squared) being the sole exception. This finding is most likely a result of the editing process, because in many cases one leave period was split up into several ones. The same explanation holds for the results regarding parental leave. The test of the overall model reveals a highly significant difference between the different versions of the data. This result is not surprising, because the dependent variable is different for both models while the independent variables are derived from the same edited data. In sum, I would still argue that there are more similarities than differences in the models compared, which indicates that the editing process has generated more complete and coherent data but the results and the interpretations remain the same.

⁴ Seemingly unrelated estimation is not possible with models for recurrent events used in the original version of the analyses. Therefore I examined standard event history models clustered by person-identifier. Significance tests were estimated using coefficients and not the exponentiated coefficients displayed, which does not influence the overall results.

4. Discussion and conclusion

The results show that although quite some changes were made during the editing processes, the results and implications concerning a specific research question substantially remain the same, independent of which version of the data are used. Whether this holds with respect to other research questions as well, cannot be answered conclusively at this state of research. However, because I have chosen a research question for which many of the editions applied are relevant, I am confident that the results might at least to some extent be generalised. In sum, my conclusion is that it might indeed be doubted that the financial means required for the post-interview editing are really spent wisely. This step of the editing surely produces ‘better’ data with respect to temporal and content-related consistency on the individual level. However, overall the errors seem to be more or less ‘at random’ meaning that the editing procedures reduce noise but substantially do not change the results.

So all in all, with only this short examination of a single example for post-survey data editing, the discussion on whether post-survey data editing is indeed necessary will most likely not be answered once and for all. Nevertheless, post-survey data editing is without doubt a very costly step and putting more effort in within-survey editing done in collaboration with the respondent might offer a promising and more cost efficient alternative. Due to increasing technical possibilities, we have seen rapid progress regarding such in-survey methods in recent years. However, which method will work best does still depend on the specific circumstances of the survey. In the adult stage (start cohort 6) of the NEPS (National Educational Panel Study), for example, no post-interview editing has been applied so far. Among the main advantages of this strategy is that total survey quality – meaning the quality from producer and user perspective (Biemer 2010) – is increasing due to faster availability of the data.

The results presented here could be extended in several directions: first, at least one different research question should be picked and a similar procedure should be applied. Especially analyses of transition events that are less prone to memory-error than is parental leave could be of interest. One example for such an event is employment episodes. This would give us at least some idea about whether the above stated assumption holds and the results presented here can indeed be generalised. Second, in the analyses presented only the dependent variable

was changed and an edited and unedited version of this variable was used to perform the data analyses. However, it might also be possible to edit the independent variables applying similar procedures. This certainly involves enormous data preparation efforts, which are beyond the scope of this chapter.

References

Antoni, Manfred; Drasch, Katrin; Kleinert, Corinna; Matthes, Britta; Ruland, Michael; Trahms, Annette (2010): Working and learning in a changing world. Part I: Overview of the study. FDZ Methodenreport 05/2010.

Biemer, Paul P. (2010): Total survey error: design, implementation and evaluation. In: Public Opinion Quarterly, Vol. 24, No. 5, pp. 817-848.

Blossfeld, Hans-Peter; Golsch, Katrin; Rohwer, Götz (2007): Techniques of event history modeling using Stata: New approaches to causal analysis. Mahwah, NJ:

Erlbaum.

Brückner, Hannah (1995): Surveys don't lie, people do? An analysis of data quality in a retrospective life course study. In: Materialien aus der Bildungsforschung Nr. 50. Berlin: Max-Planck-Institut für Bildungsforschung.

Brückner, Erika; Hoffmeyer-Zlotnik, Jürgen; Tölke, Angelika (1983): Die Datenedition als notwendige Ergänzung der Datenerhebung bei retrospektiven Langzeitstudien. In: ZUMA-Nachrichten 13(17), pp. 73-83.

Drasch, Katrin (2011): Do changing institutional settings matter? Educational attainment and family related employment interruptions in Germany. IAB Discussion Paper 13.

Drasch, Katrin; Matthes, Britta (2009): Improving retrospective life course data by combining modularized self-reports and event history calendars. Experiences from a large scale survey. IAB Discussion Paper 21.

Gilberg, Reiner; Hess, Doris; Prussog-Wagner, Angela; Steinwede, Angelika (2011): Arbeiten und Lernen im Wandel. Teil III: Methodenbericht. FDZ Methodenreport 10/2011.

Groves, Robert M.; Fowler, Floyd J. Jr.; Couper, Mick. P., Lepkowski, James M.; Singer, Eleanor; Tourangeau (2009): Survey methodology. 2nd edition. Hoboken, NJ: John Wiley & Sons

Hillmert, Steffen (2002): Edition von Lebensverlaufsdaten: Zur Relevanz einer systematischen Einzelfallbearbeitung bei standardisierten Befragungen. In: ZUMA-Nachrichten 51(26), pp.120-140.

Kleinert, Corinna; Matthes, Britta; Jacob, Marita (2008): Die Befragung "Arbeiten und Lernen im Wandel". Theoretischer Hintergrund und Konzeption. IAB-Forschungsbericht 5/2008.

Künster, Ralf (2008): Dokumentation zum Editionstool „ALWAEdi“. Unpublished manuscript.

Mathes, Britta; Trahms, Annette: 2010): Arbeiten und Lernen im Wandel. Teil II: Codebuch. FDZDatenreport 02/2010.

Mathes, Britta; Valentin, Margot-Anna (forthcoming): Arbeiten und Lernen im Wandel. Dokumentation der Edition.

Tölke, Angelika (1982): Möglichkeiten und Grenzen einer Edition bei retrospektiven Verlaufsdaten: dargestellt an der Lebensverlaufsbefragung des Sfb 3. Arbeitspapier Nr. 90. Universität Frankfurt/Main.