Longitudinal and Life Course Studies • vol xx • no xx • 1–23
 • Paper received August 2020/Revised February 2021/Published online June 2021
Online ISSN 1757-9597 • https://doi.org/10.1332/175795921X16209324334818

RESEARCH ARTICLE

Consistency of data collected through online life history calendars

André Berchtold, andre.berchtold@unil.ch
Bertil Wicht, Bertil.Wicht@unil.ch
Joan-Carles Surís, Joan-Carles.Suris@chuv.ch
Davide Morselli, davide.morselli@unil.ch
University of Lausanne, Switzerland

The collection of longitudinal data is crucial in some domains such as life course studies. However, prospective studies are considerably costly, and thus retrospective data are an appealing alternative. A life history calendar is a tool specifically conceived to collect retrospective data. However, although it is designed to enhance the recall process of the respondents, the accuracy of the data collected through this approach remains unknown, particularly when data is collected online. In this study, we conducted a secondary analysis of data collected from n = 5,181 respondents through an online survey regarding their sexual health. Because we inquired about the occurrence of certain events twice during the survey, once using a life calendar and once through a traditional questionnaire, we were able to perform three types of consistency checks: (1) reporting of single events, (2) age when the events occurred and (3) correct timing between two events. The main results indicated that it is generally more difficult to remember the exact age of occurrence of an event than the event itself, that the report of related events is generally coherent, and that women are generally more accurate in their answers than men. Based on our results, it is therefore possible to identify a subset of persons whose answers are more consistent throughout the survey. This study also shows that data obtained through an online life history calendar can be of a quality similar to that obtained through a traditional online questionnaire.

Key words life history calendar • online data collection • retrospective data • consistency checks • data accuracy

Key messages

- Online life history calendars can be used to collect accurate retrospective data.
- The quality of data can be very heterogeneous among participants.
- Data quality controls should be implemented within the life history calendar tool.

To cite this article: Berchtold, A., Wicht, B., Surís, J. and Morselli, D. (2021) Consistency of data collected through online life history calendars, *Longitudinal and Life Course Studies*, vol xx, no xx, 1–23, DOI: 10.1332/175795921X16209324334818

Introduction

Longitudinal data are being increasingly used in social sciences, and they are becoming indispensable in specific domains such as life course studies. However, such data are notoriously more costly to collect than cross-sectional data. Researchers have two main possibilities in obtaining longitudinal data: either they must prospectively follow a cohort of subjects, or they must ask the subjects to retrospectively remember their past. The second option is often preferable because only one questionnaire is required to obtain complete longitudinal information; this is advantageous in terms of time and cost. Currently, life history calendars (LHCs), also known as event history calendars, are seen as the best solution for collecting retrospective data, especially on long-past periods (Freedman et al, 1988; Belli, 1998). The advantage of LHCs over traditional questionnaires is the combination of a graphical presentation of the life course (or other domain of interest) and the presence of cues designed to enhance the recall ability of the respondents (Shum, 1998; Glasner et al, 2015).

On the basis of research on autobiographical memories (Bradburn et al, 1987; Conway, 1996), different types of processes that can be mobilised when collecting memories have been identified: top-down, sequencing and parallel. We speak of top-down retrieval when we try to date a specific event in function of a larger one (for instance identifying the exact period of a two-week student job during the three-month summer term), which implies a kind of hierarchical storing of the information. Sequencing retrieval uses the succession of events belonging to a common domain to date them (for instance the succession of student jobs of one person during academic studies), and parallel retrieval implies the use of events belonging to one domain to better recall or date events from other domains (for instance, remembering where we lived, and with whom, can help remembering what kind of job we were doing). Contrary to traditional lists of questions, the design of an LHC as a calendar with all time periods appearing simultaneously and different columns for different domains helps mobilising all types of memory retrieval processes (Belli, 1998). In addition to the different memory retrieval processes, it has also been observed that many factors can impact the correct remembering of the past. Among them, we can cite the simple passage of time that can decrease the accuracy of information recall (Mathiowetz and Duncan, 1988), the so-called *telescoping effect*, the tendency to report recent events as having occurred a longer time ago, and to report older events as having occurred more recently than in reality (Neter and Waksberg, 1964; Thompson et al, 1988; Johnson and Schultz, 2005), and a greater facility to remember more salient events (Herz and Schooler, 2002). It has been shown that the use of a tool such as the LHC helps to limit some of these effects by allowing the combination of all retrieval mechanisms instead of relying on only one of them (Caspi et al, 1996; Luke et al, 2011; Schatz et al, 2020).

The quality of retrospective data encompasses at least three different dimensions: (1) Are all pertinent events correctly indicated on the calendar? (2) Is the time of occurrence of an event, as indicated on the calendar, correct? (3) Are the relationships between different events correct? For instance, some pairs of events must appear in a certain order; in addition, two events might be mutually exclusive, or the time separating two events might not be allowed to be lower than a given number of months or years. When the third dimension of data quality can be examined on the

basis of the sole data collected through an LHC, the first two dimensions require the knowledge of a gold standard, that is, reliable information about the true number of events that occurred to one person, as well as the exact timing of each of them. The problem of the quality of retrospective data has become even more important because the trend in data collection has been to apply web-based solutions. More precisely, in the case of an LHC, the standard settings are a paper-and-pencil survey completed by respondents with the help of an interviewer. A complete shift to an online system means that a digital version of the calendar is used, presumably without the assistance of an interviewer. An open version of an online LHC is not yet available for research purposes. Only tools displaying the collected information have been made available in R (Wieczorek et al, 2020), but without the possibility of interactive data entry. However, several trials have been conducted during the past few years (Callegaro et al, 2005; Brüderl et al, 2017), including a survey regarding the sexual health of young adults in Switzerland (Barrense-Dias et al, 2018). Previous studies showed that the use of a self-administered LHC is feasible (Morselli et al, 2019), and that the amount of information obtained through an online calendar is equal to or higher than that obtained through a traditional online survey (Morselli et al, 2016). However, the quality of the information collected using a self-administered online LHC remains unclear.

Even with well-designed retrospective calendars, it is still sometimes difficult to remember all past events, particularly when they occurred significantly long ago; in addition, even when they are remembered, the exact memory may be altered in some manner. For instance, the correct timing of an event may be erroneous. Therefore, some researchers can be reluctant to implement an LHC because they believe that data will be of lower quality than data collected prospectively (Bergman et al, 1991). Moreover, in most situations, the quality of the retrospective data is difficult to validate owing to the lack of gold standard. For instance, there is no systematic recording of many events of interest such as robberies or sexual encounters, or they can be very difficult to obtain. To evaluate the quality of retrospective data, the best solution is then to compare a prospective data set with data from a retrospective study that is conducted at the end of the prospective study. Some studies have followed this design, but highly heterogeneous results have been obtained, some showing good (Schlinkmann et al, 2017) or average (Reuben et al, 2016) quality of the retrospective data, whereas others achieve poor results (Kazemian and Farrington, 2005). However, we noted that in all such studies data on very specific topics were used. Moreover, retrospective data were not obtained using an LHC, and finally nothing ensures that data collected throughout a prospective study are free of error. Only the use of administrative data could perhaps lead to a perfect reference point for the evaluation of retrospective data quality, but it would drastically limit the extent of such comparisons, since administrative data are rather limited in terms of information diversity and of accessibility.

This study is based on a secondary analysis of data collected during an online survey that included both an LHC and traditional questions. While we did not have access to an error-free set of reference information about respondents, we did have the opportunity to compare two reports of the same information by the same respondents, thus allowing us to assess how answers might possibly differ between an online traditional questionnaire and an online LHC, using answers to the traditional questions as a benchmark. We hypothesised that: (1) the report of events will differ

between the LHC and the traditional questionnaire; (2) the report of the age of occurrence of the different events will also differ between both questionnaires; (3) the time ordering of pairs of events reported on the LHC will generally be correct; and (4) large differences will be observed from one respondent to another regarding the coherence of the data collected using the LHC. Indeed, a number of previous studies did already compare an LHC and a traditional questionnaire (Belli et al, 2001; Engel et al, 2001; Belli et al, 2007; Bilgen and Belli, 2010), but they were using data from paper-and-pencil and from computer-assisted telephone interviews. By contrast, we consider here data obtained from a self-administered online survey, and we want to determine whether previous results will be confirmed or not given this specific setting.

Methods

Data

Data were collected from the Sexual Health and Behaviour of Young People in Switzerland project, a self-administered Swiss national survey on sexual behaviours conducted in 2017 among a representative sample of young adults aged 25–27 (Barrense-Dias et al, 2018) whose mailing addresses were provided by Swiss Federal Statistical Office. The final sample included 7,142 participants (response rate 15.1%), with 5,181 having answered the entire questionnaire (72.5%). The quite low response rate can be explained mainly by the very subject of the survey (sexual behaviour).

The survey was conducted online and consisted of three parts: a first part with 19 traditional questions (mainly socio-demographic), followed by an LHC, a graphical tool designed to report when 38 different events belonging to six domains (family, education/work, transition, sexuality, health and substances) occurred from birth to the present day. The time resolution of the calendar was a quarter of the year, meaning that we know in which quarter, from birth to present, an event occurred. Figures 4 and 5 provided in the appendix show the general appearance of the LHC as well as one of the pop-up windows used to provide information on each event. Finally, the third part of the questionnaire included 329 traditional questions, mainly on all aspects of sexuality. Most of these questions were conditional to the answer to a previous question, so no one had to answer to all questions. From the 7,142 respondents who took the survey, 12 dropped out during part 1 of the questionnaire (three females, five males, four unknown), 1,508 during part 2 (51.6% females), and 441 during part 3 (48.5% females). Therefore, most of the dropouts are likely to be caused by the unusual aspect of the second part of the questionnaire, the LHC. In contrast, even if the questionnaire was quite long, the number of individuals who dropped out during part 3 is low, indicating a good motivation to complete the survey. It is also worth mentioning that apart from a small unconditional gift (a pen) in the invitation letter, no further incentives were given during or after completion of the questionnaire. More details about the survey and data collection are available in the main report of the project (Barrense-Dias et al, 2018). The study was reviewed by the research ethics committee of the canton of Vaud (CER-VD), indicating that it was in compliance with Swiss law.

When the respondents accessed the LHC part of the survey, a certain number of events were pre-filled on the calendar based on answers provided in the first part of

the survey: first arrival in Switzerland, end of mandatory school, first regular paid job, departure from parents' home, first move after departure from the home of one's parents, first driver's licence, first smartphone and first menstrual cycle (women only). Depending on the respondent, the number of pre-positioned events varied between zero and eight. The respondents then had the opportunity to add further events related to the six different domains, listed previously. In addition, the respondents had the option to modify the events already entered into the calendar (including pre-positioned events), for instance by changing their quarter of occurrence, and to delete events in the case of a mistake. The numbers of added, modified and deleted events were also stored in the database.

In this study, we analysed the quality of data collected through the LHC using data from the two other parts of the questionnaire as a benchmark. To have a consistent sample throughout all the analyses, we decided to use only those 5,181 participants who answered the entire survey. Moreover, because the complete questionnaire was considerably lengthy, with some very sensitive questions, we believed that people who answered the entire questionnaire should also have provided data of a better overall quality.

Consistency checks

Given the available data, we performed three types of consistency check:

1 Event: For questions asked in both the calendar and third parts of the questionnaire, when an event was reported on the calendar, it must also have been reported in the third part, and vice versa.
2 Age: For questions asked in both the calendar and third parts of the questionnaire, when the occurrence of an event was reported twice, the age at occurrence must be the same in the two reports.
3 Temporality: When two events asked for on the calendar are only possible in a given order, or must be separated by a minimal amount of time, their report on the calendar must respect this timing.

These three types of consistency checks are related with our first three hypotheses, while the fourth hypothesis is using all the consistency checks. We detail the three types of consistency checks in the following.

Events reported twice

The possible occurrence of several events was asked twice on the questionnaire, once on the LHC and once in the third part of the survey. This occurred mainly because in the last part of the questionnaire, when using traditional questions, it was sometimes difficult to introduce a subject without asking a basic question beforehand. For instance, in the substance use section of the survey, it was illogical to ask the respondents whether they had smoked tobacco during the last 12 months, without asking first whether they had ever smoked tobacco. However, this does not apply to all events in the calendar, because some events were related to subjects not explored further in the traditional part of the questionnaire (for instance, the death of a relative), or subjects that were sufficiently described by the events on the calendar. This explains

why only ten events were available twice, eight for all respondents (beginning of the first stable relationship, end of the first stable relationship, beginning of the current relationship, first sexual intercourse, first unwanted sexual experience, first cigarette, first cannabis consumption and first drunkenness episode), and two for females only (first appointment with a gynaecologist and first pregnancy).

Reported ages

For the different events asked about twice on the questionnaire, we can also compare the age reported for the event. Based on the calendar, we know during which quarter an event occurred, and in the third part of the questionnaire we asked the age in years when the event occurred. For the analysis, we considered that an event indicated in a specific quarter of the calendar occurred in the middle of that quarter. Similarly, in the first part of the questionnaire, the year and month of birth were elicited, and we considered that the birthday is in the middle of the given month. Therefore, the maximum timing error is 1.5 months for an event on the calendar, and 0.5 months for age. We took this inherent imprecision into account and considered a reported age difference between the LHC and the third part of the survey of up to 2 months for greater accuracy. For instance, if the age given in the third part of the questionnaire for the occurrence of an event is 24, then it is coherent with ages reported in the calendar between 23 years and 10 months, and 25 years and 2 months.

Ten age comparisons were possible from our data, eight for all respondents (beginning of the first stable relationship, end of the first stable relationship, beginning of the current relationship, first sexual intercourse, first unwanted sexual experience, first cigarette, first cannabis consumption and first drunkenness episode), and two for females only (first appointment with a gynaecologist and first pregnancy). Of course, the comparison makes sense only when the event was actually reported on both the LHC and in the third part of the questionnaire.

Timing between events

It is also possible to check the timing accuracy of the report of pairs of events on the calendar. For instance, we asked about the moments of both the first and last cannabis consumption, without checking during the data collection whether the last consumption event was placed after the first one on the calendar. Therefore, we can use the relative position of these two events to perform an additional check. Note that the first consumption event can be correctly reported without the last consumption event, but the reverse is incoherent. Of course, if none of these events was reported, we cannot then determine whether an error occurred.

In addition, some pairs of events must be separated by a minimal amount of time. For instance, the time separating the beginning of an apprenticeship from the corresponding diploma is at least two years in Switzerland. Conversely, given the quarter time resolution of the LHC, some pairs of reported events must be considered as accurate even when they were indicated in the same quarter. For instance, it is possible for the first and last cannabis consumptions to have occurred within the same three-month period.

Only when the second event of the pair is reported can the pair be used to check the accuracy. If the first event of the pair was also reported, and if the order and

time gap between both events is correct, then the temporality of the pair of events is accurate. Otherwise, the temporality check is a failure. Sixteen pairs of events were testable from our data, 11 concerning all respondents (beginning/end of the first stable relationship; end of mandatory school / beginning apprenticeship; beginning apprenticeship / apprenticeship diploma; end of mandatory school / beginning of high school; beginning of high school / high school diploma; high school diploma / beginning of tertiary education; beginning of tertiary education / tertiary education diploma; beginning/end of first unemployment period; legal age / driving licence, first/last cigarette, first/last cannabis consumption), and five for females only (first/last gynaecologist appointment; first sexual intercourse / first pregnancy; menarche / first pregnancy; first pregnancy / birth of first child; menarche / first contraceptive pill).

Statistical analyses

In the first step, we performed the three separate types of consistency checks described previously. The results were reported as a number of possible checks based on the event and percentage of coherent reports, both overall and separated by gender. Since most variables used in this paper are related to sexual behaviour and substance use, and since significant differences are known to exist between women and men on these topics (Petersen and Hyde, 2010; McHugh et al, 2018), it was therefore sensible to suppose that differences could also exist regarding the remembering process of these topics, hence our choice of providing results not only overall, but also separately by gender.

We then considered the consistency of the answers provided by each respondent. For each of the three types of consistency checks, we computed the number of possible checks by the respondents and the corresponding consistency of the data (the percentage of coherent answers among the possible answers checked). We also computed the overall number of checks by the respondents and the corresponding consistency score. All computations were also performed separately based on gender, and the results are shown graphically. A t-test between females and males is also provided. Note that for a better comparison of all respondents, only the events queried to both females and males were included in the consistency scores.

In the case of age and temporality consistency checks, we also explored the relationship between the number of possible checks and the corresponding consistency score. We did not do the same for event consistency because their number (eight) was the same for all respondents.

Finally, we separated the entire sample in function of the consistency score of the respondents, and compared the socio-demographic characteristics of the respondents above and below the threshold. Six thresholds were considered: overall consistencies of 90% or more, 75% or more and 50% or more, and consistencies of 90% or above, 75% or above and 50% or above for each of the three types of consistency checks. The socio-demographics were gender (women/men), language (French/German/Italian), and education level. Using the same thresholds, we also compared the number of events that were pre-positioned, added, modified or deleted from the calendar.

The type I error was set to 5% for all statistical computations. The open-source R statistical language (R Core Team, 2018) and the RStudio environment (RStudio Team, 2016) were used for all computations.

Results

A total of n = 5,181 respondents (57.54% female, 42.35% male, and 0.11% other) completed the full survey, including the LHC and traditional questionnaire, and were included in this study. Given the small number of 'other' answers to the gender question (n = 6), these persons were included in the overall computations, but not in the comparisons by gender.

Events reported twice (hypothesis 1)

Table 1 shows the percentage of consistency for all events queried for in both the LHC and the traditional part of the questionnaire. In addition to the global consistency percentage, we also provided for each event the details of the four possible combinations of reports or non-reports of the event in the LHC and during the third part of the questionnaire. Separate results for females and males are also given.

Overall, the consistency ranged between 73.42% and 86.28%. The results by gender indicated that females were generally more coherent in their reporting of events than males with two exceptions: first unwanted sexual experience ($p < .001$) and the beginning of the current relationship (non-significant). The lowest consistency score (69.69%) was found for the beginning of the relationship among males, and the highest score appeared for the first pregnancy (95.71%) among females.

In both the overall results and the results separated by gender, the event that was most often reported as having occurred in both the LHC and the last part of the questionnaire was the individual's first sexual intercourse. By contrast, the event less often reported both times as having occurred was the first unwanted sexual experience. Regarding the discrepancies between the LHC and the last part of the questionnaire, the beginning of the current relationship was the event most often reported in the calendar but not in the last part of the questionnaire, both overall and when separated by gender. Conversely, the beginning of the first relationship was the event most reported overall in the last part of the questionnaire, but not in the LHC, followed by the first cannabis consumption and the first drunkenness episode. The results were similar for males only, whereas for females the highest discrepancy concerned the first appointment with a gynaecologist.

Reported ages (hypothesis 2)

Table 2 shows the percentage of consistency between the age at which an event occurred as deduced from the LHC, and the age given for the same event in the traditional part of the questionnaire. Notice that this comparison was possible only for events reported as having occurred twice by a participant, both in the LHC and in the last part of the questionnaire. Therefore, the number of possible comparisons varied significantly among events. The results provided are both the overall outcomes and those separated by gender.

Age consistency was globally lower than consistency concerning the reporting of events. Overall, they ranged from 40.16% (first cigarette) to 83.28% (beginning of the current relationship). Age consistency was higher among females than males, but at times the difference was non-significant (first unwanted sexual experience, first cannabis consumption and first drunkenness episode).

Table 1: Event consistency

	Overall (n = 5,181)					Female (n = 2,981)					Male (n = 2,194)					
Event	**Global %**	**% NN**	**% YY**	**% NY**	**% YN**	**Global %**	**% NN**	**% YY**	**% NY**	**% YN**	**Global %**	**% NN**	**% YY**	**% NY**	**% YN**	***p***
Beginning of the first relationship	74.06	8.14	65.91	25.63	0.31	77.26	5.97	71.28	22.54	0.20	69.69	11.12	58.57	29.85	0.46	<.001
End of the first relationship	78.07	28.01	50.07	20.32	1.60	81.31	27.31	54.01	16.71	1.98	73.70	28.99	44.71	25.21	1.09	<.001
Beginning of the current relationship	73.42	36.33	37.10	9.57	17.00	72.56	31.77	40.79	7.55	19.89	74.57	42.53	32.04	12.35	13.08	.105
First sexual intercourse	82.74	4.73	78.02	15.44	0.50	84.90	3.99	80.91	13.52	0.60	79.81	5.74	74.07	18.05	0.36	<.001
First unwanted sexual experience	86.28	82.18	4.09	13.55	0.17	81.58	75.24	6.34	18.32	0.10	92.62	91.61	1.00	7.11	0.27	<.001
First cigarette	78.94	47.67	31.27	9.88	11.18	80.31	49.11	31.20	7.88	11.81	77.07	45.76	31.31	12.63	10.30	.005
First cannabis consumption	74.81	37.83	36.98	24.32	0.87	75.75	42.37	33.38	23.15	1.11	73.56	31.72	41.84	25.93	0.50	.075
First drunkenness episode	75.56	12.22	63.35	23.76	0.68	76.89	15.46	61.42	22.34	0.77	73.75	7.79	65.95	25.75	0.50	.010
First appointment with a gynaecologist						75.58	4.39	71.18	24.19	0.23						
First pregnancy						95.71	88.33	7.38	3.36	0.13						

Note: For each of the ten events asked about twice, this table shows the percentage of consistency between the answer given on the life calendar and that provided in the traditional part of the questionnaire. Results are provided for both males and females together and separately. We provide the global percentage of consistency, the percentage of respondents who did not declare the event in both parts of the questionnaire (% NN), the percentage of respondents who declared the event in both parts of the questionnaire (% YY), the percentage of respondents who did not declare the event on the LHC, but who declared it in the traditional part of the questionnaire (% NY), and the percentage of respondents who declared the event on the LHC, but who did not declare the event in the traditional part of the questionnaire (% NY). For the eight events possible for both females and males, the last column reports the *p*-value of a Student t-test.

Table 2: Age consistency

	Overall (n = 5,181)		Female (n = 2,981)		Male (n = 2,194)		
Age comparison	**n**	**% coher**	**n**	**% coher**	**n**	**% coher**	***p***
Beginning of the first stable relationship	3,408	69.01	2,120	70.99	1,283	65.71	.001
End of the first stable relationship	2,590	67.88	1,608	70.90	979	62.82	<.001
Beginning of the current relationship	1,920	83.28	1,215	84.61	702	80.91	.041
First sexual intercourse	2,699	60.95	1,596	63.53	1,099	57.05	<.001
First unwanted sexual experience	146	54.11	133	54.89	12	41.67	.409
First cigarette	1,616	40.16	927	42.72	686	36.88	.018
First cannabis consumption	1,912	56.85	991	58.12	918	55.56	.258
First drunkenness episode	3,276	59.46	1,827	59.93	1,445	58.96	.574
First appointment with a gynaecologist			2,117	67.97			
First pregnancy			218	75.23			

Note: Consistency in age reported for the ten events asked for twice. For each event, we provide the number of checks that were performed (that is, the number of times the event was reported twice on the questionnaire), and the corresponding consistency score. Results are provided for both males and females together and separately. For the eight events possible for both females and males, the last column reports the *p*-value of a Student t-test.

Timing between events (hypothesis 3)

Table 3 shows results regarding the timing between pairs of reported life events. In contrast to the data shown in Tables 1 and 2, all data here were derived from the LHC only, without a comparison with the traditional part of the questionnaire. We considered both the logical order of the two events (for instance, high school graduation must occur after the beginning of high school) and the time gap between the two events (for instance, in Switzerland, an apprenticeship cannot be shorter than two years).

Overall, the temporal consistency was higher than the two other types of consistency, often above 90%. However, there were a few exceptions, the most important being the relation between the end of mandatory school and the beginning of high school with only 57.59% consistency, followed by the relation between the high school diploma and the beginning of tertiary education (63.39%). In contrast to the other two types of consistency, the percentages were generally not significantly different between females and males, with three exceptions, being the beginning of high school / high school diploma, legal age / driving licence, and first/last cigarette consumption, showing slightly higher scores among females.

Inter-respondent variability (hypothesis 4)

The next step was to identify those respondents who could be considered as having provided reliable answers. To allow a better comparison between all respondents, we chose to consider only events pertinent for all respondents (eight events in Tables 1 and 2, and 11 pairs of events in Table 3). Moreover, scores regarding age and temporality took into account only the number of events (age) or pairs of events (temporality) that were reported by the respondent. For instance, if a respondent did not indicate first cannabis consumption in either the LHC or third part of the questionnaire, then no age comparison could be made regarding this event. Similarly, the comparison of the timing between the high school diploma and the beginning of tertiary education was performed only among respondents having indicated these two events on their life calendar. Figure 1 shows the distribution of the number of possible checks of each type by respondent, both overall and by gender.

Overall, the average number of possible consistency checks was slightly higher for females (16.06) than for males (15.40; $t = 5.72$, $df = 4{,}580$, $p < .001$), indicating that females reported more events in average than males. In detail, when the number of checks was identical for all respondents regarding the events (eight), the number of possible consistency checks was larger for females in terms of both age (3.49 versus 3.25; $t = 3.82$, $df = 4{,}698$, $p < .001$) and temporality (4.57 versus 4.16; $t = 6.70$, $df = 4{,}534$, $p < .001$), leading to more skewed distributions.

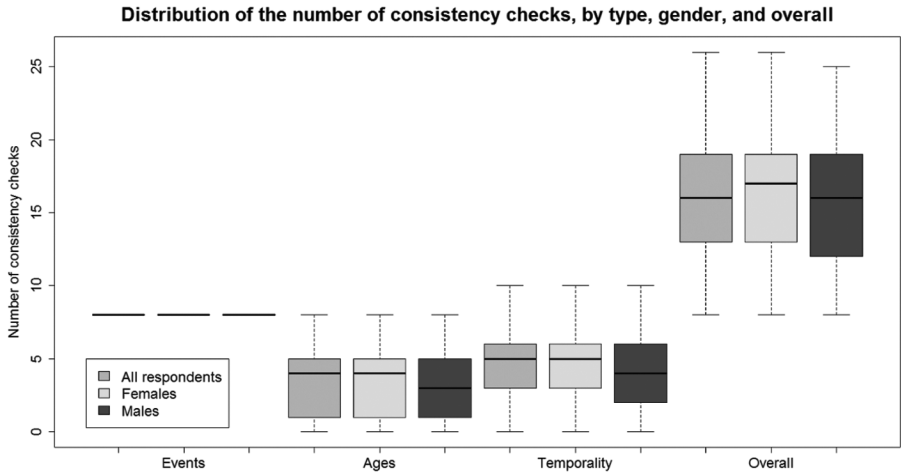
Figure 2 shows the distribution of the consistency score among the respondents. The results are provided separately for each type of consistency check by gender and overall, and indicate that females were slightly more coherent in their answers than males, with an overall score of 74.93% as compared with 71.91% among males ($t = 5.53$, $df = 4{,}452$, $p < .001$). In detail, the consistency was higher among females (78.82%) than among males (76.85%; $t = 2.91$, $df = 4{,}330$, $p < .004$) for the reporting of events. It was also higher among females (65.70%) than among males (60.94%; $t = 4.70$, $df = 3{,}655$, $p < .001$) for age consistency. However, there was no difference

Table 3: Temporality consistency

	Overall (n = 5,181)		Female (n = 2,981)		Male (n = 2,194)		
Timing of events	**n**	**% coher**	**n**	**% coher**	**n**	**% coher**	***p***
Beginning of first stable relationship / end of first stable relationship	2,677	97.91	1,669	98.08	1,005	97.61	.423
End of mandatory school / beginning of apprenticeship	1,778	94.49	960	93.85	815	95.21	.207
Beginning of apprenticeship / apprenticeship of diploma	1,871	81.40	1,010	80.89	857	82.03	.528
End of mandatory school / beginning of high school	2,273	57.59	1,434	57.39	837	57.95	.797
Beginning of high school / high school diploma	2,252	78.73	1,411	80.79	839	75.33	.003
High school diploma / beginning of tertiary education	2,966	63.39	1,831	64.66	1,133	61.25	.062
Beginning of tertiary education / tertiary education diploma	2,380	87.94	1,506	87.52	872	88.65	.410
Beginning of first unemployment / end of first unemployment	482	97.51	289	96.54	192	98.96	.064
Legal age / driving licence	4,603	95.83	2,636	96.70	1,963	94.65	<.001
First cigarette / last cigarette	739	97.16	455	98.68	283	94.70	.006
First cannabis consumption / last cannabis consumption	735	97.14	411	98.05	322	95.96	.107
First gynaecologist appointment / last gynaecologist appointment			2,010	91.44			
First sexual intercourse / first pregnancy			224	92.41			
Menarche / first pregnancy			224	99.11			
First pregnancy / birth of first child			162	76.54			
Menarche / first contraceptive pill			2,078	96.20			

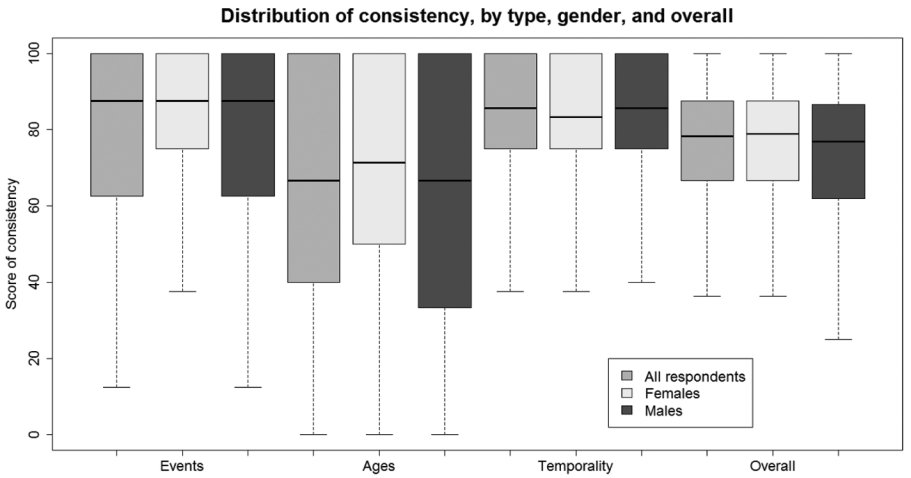
Note: Consistency in terms of timing between two events reported on the LHC, considering both the order of the two events and the time gap between them. Results are provided for both males and females together and separately. For the 11 events possible for both females and males, the last column reports the *p*-value of a Student t-test.

Figure 1: Boxplots for the distribution of the number of possible consistency checks by respondents



Note: The results are provided separated on the basis of gender and type of check. Regarding the events, all respondents have exactly eight possible checks.

Figure 2: Boxplots for the distribution of the consistency score

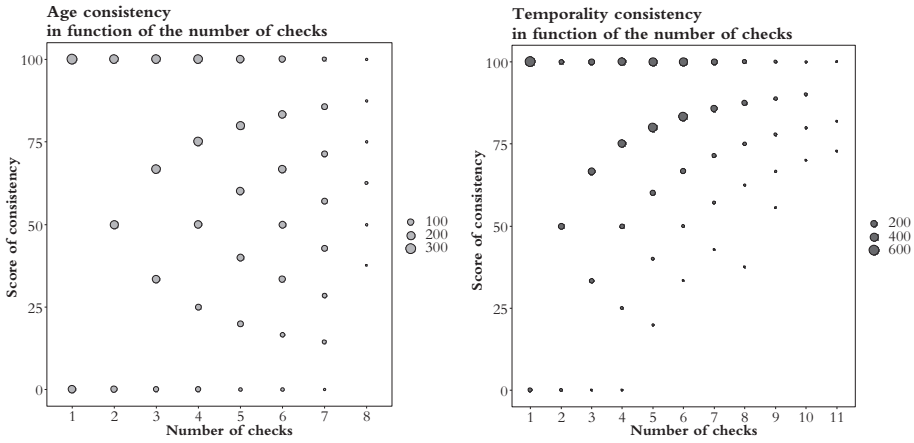


Note: We provide the scores both overall and separately for each type of consistency check and based on gender.

regarding temporality (83.28% versus 83.36%; t = −0.14, df = 4,387, $p < .886$). All distributions were asymmetric with a flattening towards the lowest scores.

We also investigated the relationship between the number of possible checks and the consistency score. The number of checks did not have any influence on the report of events because it was equal to eight for all respondents. However, it could influence the age and temporality consistencies. Figure 3 shows the relationship between the number of checks and the consistency scores for age and temporality. Regarding age, there was no apparent relationship between the number of checks

Figure 3: Distribution of age and temporality consistency as a function of the number of possible consistency checks



and the consistency score. We can see that when eight checks were possible, then the consistency score was never lower than 37.5%. However, regarding the temporality, a relationship appears: the consistency score tended to increase as the number of possible checks increased.

Finally, we compared the characteristics of the respondents with a high consistency score with respondents having a lower score. First, we considered the overall consistency score followed by the respondents having a score higher than a given percentage simultaneously in each of the three types of consistency checks. We considered three consistency thresholds: 90%, 75% and 50%. Three characteristics of the respondents were investigated: gender (female/male), country of birth (Switzerland/other), and level of education (mandatory school / apprenticeship / high school / tertiary education / other). Table 4 summarises the results as a function of the overall consistency score, and Table 5 provides the results when the consistency in each of the three domains was above a given level.

The results in Tables 4 and 5 show that respondents with a higher degree of consistency generally differed significantly from those under the threshold in terms of socio-demographic characteristics. The only exceptions can be explained by the strong unbalance between the two compared groups, resulting in a significantly low power of the chi-2 test. The more consistent respondents were more often females, born in Switzerland, and with a higher level of education. The results also showed that the consistency score could be substantially different among the three types of checks. For instance, only 146 of the respondents had a score equal to or higher than 90% in each of the three domains, but 936 had an overall score equal to or higher than 90%.

Tables 4 and 5 also show information on the number of events reported on the LHC, and on the changes to these events. To allow for a fair comparison, only events possible for both women and men were used. More consistent respondents tended to enter a significantly larger number of events onto their calendar than less consistent respondents. The same was also true regarding the number of modified and deleted events, even if such possibilities were seldom used in practice. This tends

Table 4: Comparison of respondents as a function of their overall score of consistency

	Threshold at 90%			Threshold at 75%			Threshold at 50%		
	<90% (n = 4,245)	≥90% (n = 936)	*p*	<75% (n = 2,069)	≥75% (n = 3,112)	*p*	<50% (n = 580)	≥50% (n = 4,601)	*p*
Gender									
Female	57.3	59.2	.273	53.4	60.4	<.001	47.0	58.9	<.001
Male	42.7	40.8		46.6	39.6		53.0	41.1	
Country of birth									
Switzerland	87.4	92.3	<.001	85.3	90.2	<.001	84.1	88.8	.001
Other	12.6	7.7		14.7	9.8		15.9	11.2	
Education									
Mandatory school	1.5	0.9	<.001	1.7	1.1	<.001	3.1	1.2	<.001
Apprenticeship	32.9	27.4		37.2	28.3		46.7	30.0	
High school	11.8	9.5		11.6	11.3		13.4	11.1	
Tertiary education	50.5	58.5		45.8	56.0		33.1	54.3	
Other	3.3	3.7		3.6	3.3		3.6	3.4	
Events on the calendar									
n. pre-positioned events	5.14	4.97	<.001	5.17	5.07	.002	5.04	5.12	.097
n. added events	9.66	12.72	<.001	6.93	12.39	<.001	1.11	11.36	<.001
n. modified events	0.46	0.66	<.001	0.36	0.60	<.001	0.05	0.56	<.001
n. deleted events	0.52	0.73	<.001	0.33	0.70	<.001	0.07	0.62	<.001

Note: The two groups of respondents are compared using a chi-2 test of independence for socio-demographic characteristics, and a t-test for the number of events.

Table 5: Comparison of respondents as a function of their consistency score based on type of check

	Threshold at 90%			Threshold at 75%			Threshold at 50%		
	<90% (n = 5,035)	≥90% (n = 146)	*p*	<75% (n = 3,794)	≥75% (n = 1,387)	*p*	<50% (n = 2,206)	≥50% (n = 2,975)	*p*
Gender									
Female	57.5	61.6	.316	55.5	63.5	<.001	52.6	61.3	<.001
Male	42.5	38.4		45.5	36.5		47.4	38.7	
Country of birth									
Switzerland	88.1	95.2	.008	87.0	91.6	<.001	86.2	89.8	<.001
Other	11.9	4.8		13.0	8.4		13.8	10.2	
Education									
Mandatory school	1.4	0.7	.613	1.7	0.6	<.001	2.1	0.8	<.001
Apprenticeship	32.0	26.7		34.0	26.0		37.8	27.5	
High school	11.4	11.6		11.8	10.2		11.3	11.5	
Tertiary education	51.8	57.5		48.9	60.3		45.6	56.7	
Other	3.4	3.4		3.6	2.9		3.2	3.6	
Events on the calendar									
n. pre-positioned events	5.12	4.88	.008	5.11	5.11	.892	5.05	5.16	<.001
n. added events	10.14	12.81	<.001	9.18	13.02	<.001	7.14	12.49	<.001
n. modified events	0.49	0.69	.122	0.46	0.61	<.001	0.37	0.59	<.001
n. deleted events	0.55	0.84	.030	0.50	0.71	<.001	0.41	0.66	<.001

Note: For instance, 90% indicates that the respondents have a consistency score equal to or higher than 90% in all three domains of checks (event, age and temporality). The two groups of respondents are compared using the chi-2 test of independence for socio-demographic characteristics, and the t-test for number of events.

to demonstrate that the more consistent respondents paid more attention to their calendar, and that they made corrections more often. In some cases, although, the number of pre-positioned events was significantly higher for respondents below the consistency threshold. This can be explained by the fact that more of these respondents were born outside of Switzerland, meaning that they had to report their 'first arrival in Switzerland' as an event.

Discussion

The purpose of this study was to evaluate the quality of retrospective data collected through an online LHC by comparing them with other data collected using traditional questions. Different consistency checks involving the indication of specific events, correct reporting of the age of occurrence of these events, and relative timing of pairs of events were performed. We hypothesised that (1) the report of events would differ between the LHC and the traditional questionnaire; (2) the report of the age of occurrence of the different events would also differ between both questionnaires; (3) the time ordering of pairs of events reported on the LHC would generally be correct; and (4) large differences would be observed from one respondent to another.

Our results confirm these four hypotheses. Table 1 shows that large differences are observed between the LHC and the traditional questionnaire regarding the report of specific events (hypothesis 1), the percentage of error being often larger than 20%. Table 2 shows that the same is true concerning the correct report of the age at which these events occurred (hypothesis 2), results being the worst among the three types of consistency checks (Figure 2). Of course, our measure suffers from the fact that we did not have access to the actual timing of these events. Instead, we relied on a comparison between the timing given on the LHC with a second timing deduced from the self-reported age when the event occurred. However, even with this limitation in mind, it indicates that there is less coherence in the timing than in a simple report of the same events. Inconsistencies between the two measures of the time an event occurred are not surprising, since Janssen and colleagues demonstrated that the way a date is asked for, either in an absolute or relative sense, has an influence on the answer (Janssen et al, 2006). However, it must be noted that, depending on the objectives of the research project, a maximal difference of approximately two months for defining the accuracy (as applied in this study) may be substantially restrictive. Larger time differences can sometimes be considered sufficient to define an accurate report.

Regarding the third type of consistency check (temporality of pairs of events, hypothesis 3), the results are generally equal to or better than those of the first two types of control, especially in men, although they must be considered extremely carefully, because the order of occurrence of these pairs is generally evident (for example, the last cigarette smoking event must be placed after the first cigarette smoking event on the calendar), and in several cases, their temporal spacing was based on legal conditions (for instance, in Switzerland an apprenticeship cannot be shorter than two years). Moreover, in many situations, the correct order of two events does not imply that the reported time of occurrence is also correct.

Our results also indicate that the variability in the data accuracy reported on an LHC can be very high from one respondent to another (hypothesis 4), with some overall consistency scores lower than 40% (Figure 2). The lowest consistency scores were obtained in the age indicators, which compare the LHC and answers to traditional

questions. Because the right answer is not known, we cannot assess whether the bias was in the LHC, in the traditional questionnaire or in both. However, keeping the inconsistency of these results in mind is particularly important because the overall questionnaire was considerably lengthy, and only those respondents who completed it were included in our analyses. Moreover, the LHC was placed near the beginning of the survey before the respondents could become tired by the length of the questionnaire. We deduced from these observations that even people extremely motivated by the study (based on the fact that they finished it) can produce data with low accuracy. It is worth noting that our consistency score was computed as a function of the number of events placed on the calendar, and thus the low accuracy of the data from some of the respondents is not explained by the respondents having placed only a few events on the calendar. By contrast, we observed that, regarding the temporality checks, the accuracy of the data seems to increase with the number of possible checks. Because the number of possible checks is a function of the number of events placed on the life calendar, it indicates that the more events on the calendar, the more accurate their relative timing is (Figure 3, right). We hypothesise that, following one of the basic assumptions of LHCs, events already on the calendar can act as a trigger to both increase the number of reported events and to enhance the accuracy of their timing (Loftus and Marburger, 1983; Brüderl et al, 2017).

The three types of consistency check we performed are not similarly distributed among the respondents. This is demonstrated by the fact that the number of respondents having an overall consistency score higher than 90% for all three types of checks (n = 146, Table 5), as an example, is much lower than the number of respondents having an overall consistency score of 90% (n = 936, Table 4). The same is true for the other thresholds of consistency. Therefore, some people are perhaps better at remembering the occurrence of an event than its exact timing, whereas others might pay more attention to the correct sequence of events. Finally, the socio-demographic characteristics of the respondents are also related to the overall accuracy of the LHC data. In particular, women generally provided better data than men. The achieved level of education is also positively correlated with the consistency score, indicating that the degree of computer literacy required to answer an online LHC might be insufficient among less-educated people, particularly among manual workers who spend less time working with computers. On the other hand, the lower consistency in some strata of our sample is coherent with other studies on survey response, showing that young and poorly educated men are the most reluctant to engage in social surveys (Groves, 2006; Watson and Wooden, 2009; Voorpostel, 2010). Either way, further investigation is needed on these aspects. Finally, it is not surprising to see that respondents with higher consistency scores were also significantly more prone to edit and correct the events placed on their calendar.

This study is one of the first to compare results from an online LHC with results from a traditional questionnaire. Moreover, it is based on a large sample (n = 5,181), and it considers three different types of accuracy checks. However, in addition to those already mentioned, there are many additional limitations to this report. First, no randomisation occurred between parts 2 (the LHC) and 3 (traditional questions) of the survey. Therefore, all respondents answered the traditional questions after they completed the LHC. It would be interesting to know whether the results would have been different, especially regarding the number of reported events and their timing, had they filled in the LHC after answering the traditional questions. It is

also important to recall that all reported events are directly related to the respondents themselves, and thus we have no opportunity to check whether they really occurred or not, and whether their timing (even if reported similarly twice) is correct. In a further study, it could be useful to ask for public events, whose date of occurrence is certain. Other limitations include both an inaccuracy in terms of time and the different temporality between the LHC and traditional questions. For the LHC, the answers were given with a quarter recall precision when the ages were reported in whole years for the traditional questions. Therefore, we considered the inherent imprecision regarding the correct timing of events. It must also be noted that all respondents were approximately of the same age (25–27). If, as we hypothesise, the level of digital literacy is important to obtain accurate data by means of an LHC, we can imagine that older people, who are not digital natives, could experience more difficulties with this type of tool. More studies are required for this. Finally, we must also restate that the data analysed in this study were obtained from a survey on the extremely sensitive topic of sexuality. Therefore, some respondents could have been reluctant to provide information on certain events, perhaps lowering the quality of the data.

Conclusion

The need for longitudinal studies, particularly in social and human sciences, is well established. Therefore, all methods that can provide longer, richer and more accurate data series within a short period of time are welcome. Moreover, online self-administered approaches are considerably encouraged, often for costs reasons, but also to handle special circumstances such as during the worldwide COVID-19 crisis, or in conflict areas in which security cannot be guaranteed. We thus believe that our study is a milestone in the field of online LHC research.

First, we found that obtaining accurate data is possible with an online LHC, although a high variability in the data quality occurs from one respondent to another. Therefore, adequate checking mechanisms should be designed to evaluate the quality of the collected data. In this study, we performed a secondary analysis of data that were not collected for this purpose. However, these mechanisms should be explicitly defined and tested before the data collection. For instance, questions that are asked twice, on the LHC and elsewhere in the questionnaire, should be selected to reflect the main themes of the study. It would also be better to have the same timing precision in both cases, for instance, both in years or months. Moreover, in contrast to a traditional LHC, which is completed with the help of a human interviewer, in the online setting considered here, the respondents had only a limited opportunity to ask questions when something was unclear, and they had to rely on the sole information provided by the online help system. Similarly, when the respondents thought they had completed the calendar, an extensive reminder to warn against possible missing events, with the exception of the displayed list of non-used events, was not provided. This is important to note, because different studies have demonstrated the importance of the interviewer's role and of his/her verbal interaction with the respondent when filling in an LHC (Belli et al, 2004; Dijkstra et al, 2009; Morselli et al, 2019). However, the types of verification usually done by the interviewer can easily be implemented in the conception of an online LHC, and the same is true for the guidelines related to how the tool is used and what information to enter into

each column. In other words, the interviewer's role can be replaced by integrated verification processes, using a tutorial and clear documentation. This would ensure the performance of the consistency checks in real time, hence possibly increasing the quality of the collected data. This could also bridge another gap faced by this type of research, related to non-reported events: we never know for sure whether a non-reported event is an omission, conscious or unconscious, or if this event never actually occurred. Such implementation of checks and feedback in real time may provide insight into the occurrence of errors or omissions and help diminish them.

Finally, our analysis of online behaviours showed that a better consistency was obtained by the respondents who reported a higher number of events, and who also re-edited their answers on the LHC, correcting certain events and deleting others. This mechanism is consistent with the literature indicating that the interviewer–respondent interaction in an LHC interview helps correct inconsistencies and increase the quality of the data (see, for example, Freedman et al, 1988; Yoshihama et al, 2005). In an online self-administered LHC, this must be achieved through a user interface. Editing and re-editing procedures should be intuitive and encouraged throughout the entry process. Hence, researchers should pay specific attention to the graphical user interface and usability of their LHC. Consequently, high-quality longitudinal data can be easily collected using LHC tools.

Conflict of interest

The authors declare that there is no conflict of interest.

References

Barrense-Dias, Y., Akre, C., Berchtold, A., Leeners, B., Morselli, D. and Suris, J.-C. (2018) *Sexual Health and Behavior of Young People in Switzerland* (No. 291; Raisons de Santé). Institut universitaire de médecine sociale et préventive. https://www.iumsp.ch/Publications/pdf/rds291_fr.pdf

Belli, R.F. (1998) The structure of autobiographical memory and the event history calendar: potential improvements in the quality of retrospective reports in surveys, *Memory*, 6(4): 383–406. doi: 10.1080/741942610

Belli, R.F., Lee, E.H., Stafford, F.P. and Chou, C.H. (2004) Calendar and Question-list survey methods: association between interviewer behaviors and data quality, *Journal of Official Statistics*, 20(2): 185–218.

Belli, R.F., Shay, W.L. and Stafford, F.P. (2001) Event history calendars and question list surveys: a direct comparison of interviewing methods, *Public Opinion Quarterly*, 65(1): 45–74. doi: 10.1086/320037

Belli, R.F., Smith, L.M., Andreski, P.M. and Agrawal, S. (2007) Methodological comparisons between CATI event history calendar and standardized conventional questionnaire instruments, *Public Opinion Quarterly*, 71(4): 603–22. doi: 10.1093/poq/nfm045

Bergman, L.R., Eklund, G. and Magnusson, D. (1991) Studying individual development: problems and methods, in D. Magnusson, L.R. Bergman, G. Rudinger and B. Torestad (eds) *Problems and Methods in Longitudinal Research: Stability and Change*, Cambridge: Cambridge University Press, pp 1–27.

Bilgen, I. and Belli, R.F. (2010) Comparison of verbal behaviors between calendar and standardized conventional questionnaires, *Journal of Official Statistics*, 26(3): 481–505.

Bradburn, N.M., Rips, L.J. and Shevell, S.K. (1987) Answering autobiographical questions: the impact of memory and inference on surveys, *Science*, 236(4798): 157–61. doi: 10.1126/science.3563494

Brüderl, J., Castiglioni, L., Ludwig, V., Pforr, K. and Schmiedeberg, C. (2017) Collecting event history data with a panel survey: combining an electronic event history calendar and dependent interviewing, *Methods, Data, Analyses*, 11(1): 45–66.

Callegaro, M., Yu, M., Cheng, F.W., Hjermstad, E., Liao, D. and Belli, R.F. (2005) Comparison of computerized event history calendar and question-list interviewing methods: a two-year hospitalization history study, in *2004 Proceedings of the American Statistical Association, 59th Annual Conference of the American Association for Public Opinion Research*, Alexandria, VA: American Statistical Association, pp 4746–53.

Caspi, A., Moffitt, T.E., Thornton, A., Freedman, D., Amell, J.W., Harrington, H., Smeijers, J. and Silva, P.A. (1996) The life history calendar: a research and clinical assessment method for collecting retrospective Event-history data, *International Journal of Methods in Psychiatric Research*, 6(2): 101–14. doi: 10.1002/(SICI)1234-988X(199607)6:2<101::AID-MPR156>3.3.CO;2-E

Conway, M.A. (1996) Autobiographical knowledge and autobiographical memories, in D.C. Rubin (ed) *Remembering our Past: Studies in Autobiographical Memory*, Cambridge: Cambridge University Press, pp 67–93.

Dijkstra, W., Smit, J.H. and Ongena, Y.P. (2009) An evaluation study of the event history calendar, in R.F. Belli, F.P. Stafford and D.F. Alwin (eds) *Calendar and Time Diary Methods in Life Course Research*, Thousand Oaks, CA: Sage, pp 257–75.

Engel, L.S., Keifer, M.C. and Zahm, S.H. (2001) Comparison of a traditional questionnaire with an icon/calendar-based questionnaire to assess occupational history, *American Journal of Industrial Medicine*, 40(5): 502–11. doi: 10.1002/ajim.1118

Freedman, D., Thornton, A., Camburn, D., Alwin, D. and Young-DeMarco, L. (1988) The life history calendar: a technique for collecting retrospective data, *Sociological Methodology*, 18: 37–68. doi: 10.2307/271044

Glasner, T., van der Vaart, W. and Dijkstra, W. (2015) Calendar instruments in retrospective web surveys, *Field Methods*, 27(3): 265–83. doi: 10.1177/1525822X14554882

Groves, R.M. (2006) Nonresponse rates and nonresponse bias in household surveys, *Public Opinion Quarterly*, 70(5): 646–75. doi: 10.1093/poq/nfl033

Herz, R.S. and Schooler, J.W. (2002) A naturalistic study of autobiographical memories evoked by olfactory and visual cues: testing the Proustian hypothesis, *The American Journal of Psychology*, 115(1): 21–32. doi: 10.2307/1423672

Janssen, S.M.J., Chessa, A.G. and Murre, J.M.J. (2006) Memory for time: how people date events, *Memory & Cognition*, 34(1): 138–47.

Johnson, E.O. and Schultz, L. (2005) Forward telescoping bias in reported age of onset: an example from cigarette smoking, *International Journal of Methods in Psychiatric Research*, 14(3): 119–29. doi: 10.1002/mpr.2

Kazemian, L. and Farrington, D.P. (2005) Comparing the validity of prospective, retrospective, and official onset for different offending categories, *Journal of Quantitative Criminology*, 21(2): 127–47. doi: 10.1007/s10940-005-2489-0

Loftus, E.F. and Marburger, W. (1983) Since the eruption of Mt. St. Helens, has anyone beaten you up? Improving the accuracy of retrospective reports with landmarkevents, *Memory & Cognition*, 11(2): 114–20.

Luke, N., Clark, S. and Zulu, E.M. (2011) The relationship history calendar: improving the scope and quality of data on youth sexual behavior, *Demography*, 48(3): 1151–76. doi: 10.1007/s13524-011-0051-2

Mathiowetz, N. and Duncan, G. (1988) Out of work, out of mind: response errors in retrospective reports of unemployment, *Journal of Business & Economic Statistics*, 6(2): 221–9.

McHugh, R.K., Votaw, V.R., Sugarman, D.E. and Greenfield, S.F. (2018) Sex and gender differences in substance use disorders, *Clinical Psychology Review*, Dec;66: 12–23. doi: 10.1016/j.cpr.2017.10.012

Morselli, D., Le Goff, J.M. and Gauthier, J.-A. (2019) Self-administered event history calendars: a possibility for surveys?, *Contemporary Social Science*, 14(3-4): 423-446, doi: 10.1080/21582041.2017.1418528

Morselli, D., Berchtold, A., Suris, J.-C. and Berchtold, A. (2016) On-line life history calendar and sensitive topics: A pilot study, *Computers in Human Behavior*, 58: 141–149. https://doi.org/10.1016/j.chb.2015.12.068

Neter, J. and Waksberg, J. (1964) A study of response errors in expenditures data from household interviews, *Journal of the American Statistical Association*, 59(305): 18–55. doi: 10.1080/01621459.1964.10480699

Petersen, J.L. and Hyde, J.S. (2010) A meta-analytic review of research on gender differences in sexuality, 1993–2007, *Psychological Bulletin*, 136(1): 21–38,. doi: 10.1037/a0017504

R Core Team (2018) *R: A Language and Environment for Statistical Computing [online]*, Vienna: R Foundation for Statistical Computing, https://www.r-project.org/.

Reuben, A., Moffitt, T.E., Caspi, A., Belsky, D.W., Harrington, H., Schroeder, F., Hogan, S., Ramrakha, S., Poulton, R. and Danese, A. (2016) Lest we forget: comparing retrospective and prospective assessments of adverse childhood experiences in the prediction of adult health, *Journal of Child Psychology and Psychiatry*, 57(10): 1103–12. doi: 10.1111/jcpp.12621

RStudio Team (2016) *RStudio: Integrated Development Environment for R [online]*, Boston, MA: RStudio, https://www.rstudio.com/.

Schatz, E., Knight, L., Belli, R.F. and Mojola, S.A. (2020) Assessing the feasibility of a life history calendar to measure HIV risk and health in older South Africans, *PLOS ONE*, 15(1):1–15 art e0226024, https://doi.org/10.1371/journal.pone.0226024. doi: 10.1371/journal.pone.0226024

Schlinkmann, K.M., Bakuli, A. and Mikolajczyk, R. (2017) Incidence and comparison of retrospective and prospective data on respiratory and gastrointestinal infections in German households, *BMC Infectious Diseases*, 17(1):1–8 art 336, https://doi.org/10.1186/s12879-017-2434-5. doi: 10.1186/s12879-017-2434-5

Shum, M.S. (1998) The role of temporal landmarks in autobiographical memory processes, *Psychological Bulletin*, 124(3): 423–42. doi: 10.1037/0033-2909.124.3.423

Thompson, C.P., Skowronski, J.J. and Lee, D.J. (1988) Telescoping in dating naturally occurring events, *Memory & Cognition*, 16(5): 461–8.

Voorpostel, M. (2010) Attrition patterns in the Swiss Household Panel by demographic characteristics and social involvement, *Swiss Journal of Sociology*, 36(2): 359–77.

Watson, N. and Wooden, M. (2009) Identifying factors affecting longitudinal survey response, in P. Lynn (ed) *Methodology of Longitudinal Surveys [online]*, Chichester: John Wiley, pp 157–81, https://doi.org/10.1002/9780470743874.ch10.

Wieczorek, L.L., Tata, C.S., Penke, L. and Gerlach, T.M. (2020) Online implementation of an event history calendar with formr: a tutorial, *Personal Relationships*, 27(1): 176–208. doi: 10.1111/pere.12305

Yoshihama, M., Gillespie, B., Hammock, A.C., Belli, R.F. and Tolman, R.M. (2005) Does the life history calendar method facilitate the recall of intimate partner violence? Comparison of two methods of data collection, *Social Work Research*, 29(3): 151–63. doi: 10.1093/swr/29.3.151

Appendix

Figures 4 and 5 show the general design of the life history calendar used to collect data within the sexuality survey. Figure 4 shows the calendar itself with each row corresponding to a quarter and each column to a specific domain, and Figure 5 shows one of the pop-up windows used to collect precise information about an event.

Figure 4: Main window of the life history calendar

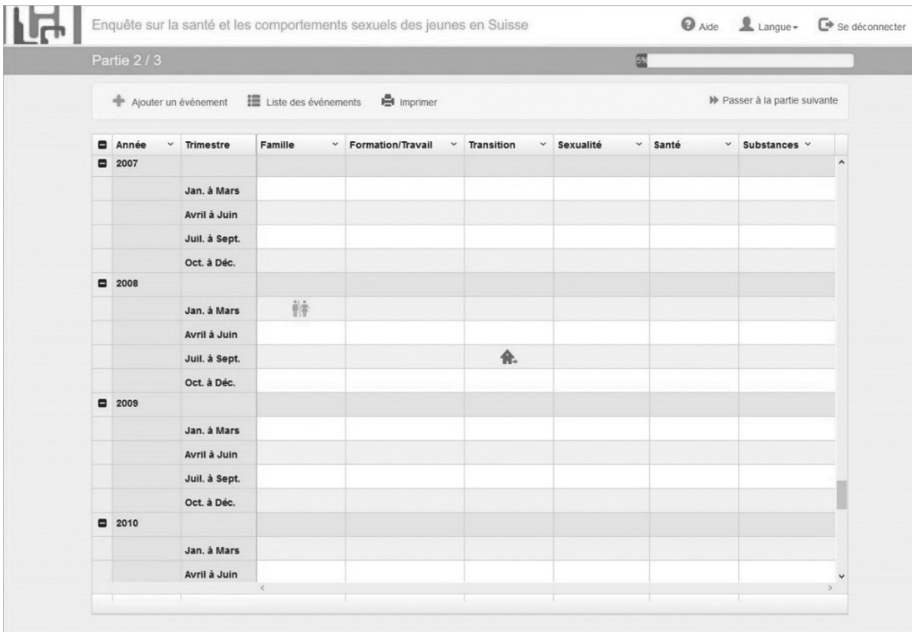


Figure 5: Pop-up window used to define an event precisely