Data Quality and Recall Bias in Time-Diary Research

Te Braak, Petrus; van Tienoven, Theun Pieter; Minnen, Joeri; Glorieux, Ignace

Published in:
Sociological Methodology

DOI:
10.1177/00811750221126499

Publication date:
2023

License:
Unspecified

Document Version:
Accepted author manuscript

Data quality and recall bias in time-diary research

The effects of prolonged recall periods in self-administered online time-use surveys

Accepted manuscript prior to typesetting

te Braak, P., van Tienoven, T.P., Minnen, J. & Glorieux, I. (2023). Data quality and recall bias in time-diary research: The effects of prolonged recall periods in self-administered online time-use surveys. *Sociological Methodology, 53*(1), 115-138. DOI: 10.1177/00811750221126499

ABSTRACT

Previous research has shown that a prolonged recall period is associated with lower data quality in time-diary research. In these studies, the recall period is roughly estimated on the basis of the period between the assigned diary day and the agreed collection day. Because this is so rudimentary, little is known about the duration of the mean recall period and its consequences for data quality. Recent advances in online methodology now allow a better investigation of the recall period using time stamps. Using a refined indicator, the authors examine the duration of the recall period, to what extent this duration is related to socio-economic characteristics, and how a prolonged recall period affects data quality. The authors demonstrate that using online time-diary data collected from 8,535 teachers in Belgium, the mean recall period is less than 24 hr for most respondents, although respondents with many time constraints have extended recall periods. Additionally, a prolonged recall period indeed has negative consequences for data quality. Quality deterioration already arises several hours after an activity has been completed, much sooner than previous research has indicated.

KEYWORDS

Recall period; data quality; data validity; memory issues; time-diary methodology

INTRODUCTION

Time-diary methodology is generally the preferred way to study estimates of daily time use (Bonke, 2005; Juster, 1985a; Otterbach & Sousa-Poza, 2010; Robinson, 1985; Sonnenberg et al., 2012), because it is specifically designed to account for recall problems (Allen et al., 2020; Niemi, 1993). This methodology attempts to record how individuals spend their time by having them keep time diaries for specific periods of time (e.g., one day, multiple days, one week). Nevertheless, time-diary methodology is not completely free of measurement errors, particularly related to recall (Juster, 1985b, 1986; Niemi, 1993). To avoid recall issues as much as possible, activities should ideally be logged close to real time (Eurostat, 2020; Harvey, 1993; Juster, 1985b, 1986; United Nations, 2004; Westat, 2013) to keep the recall period as limited as possible. The recall period thus refers to the period between the occurrence of activities and their logging in time diaries.

In time-diary research, the operational definition of this recall period has changed in recent years, simultaneously with innovations in methodology. Previously, time-diary data collection took place using either a leave-behind diary or a recall time diary. A leave-behind diary is a time diary the researcher leaves with the respondent; it is autonomously filled in by the respondent on an assigned day, after which it is retrieved by an interviewer on an agreed date. A recall time diary is completed through a (telephone) interview asking respondents retrospectively about the chronological sequence of activities they did over a defined period of time. In recall time-diary methodology, the recall period is defined similar to survey research in the phrasing of the question (e.g., by asking which activities a respondent has been doing during the past 24 hr). In a leave-behind diary methodology, the recall period is less clear. Only the assigned diary day and agreed collection date are known. Consequently, the recall period can be approximated by the period between both dates. This is a very rough estimate of the recall period, as the researcher does not know when exactly activities are logged during that period. Precisely because the recall period could only be measured so rudimentarily, little is known about its consequences for time-use research.

Contemporary time-diary studies increasingly take place in an online context in which interviewers are no longer present. Online time-diary surveys are less costly, speed up fieldwork, and eliminate interviewer bias (Minnen et al., 2014). Online time-diary methodology is similar to leave-behind time-diary methodology in that a respondent is assigned a day to keep a time diary, and the diary must be completed before the online time-diary closes (i.e., comparable with the agreed "collection" date). The major difference between the two is that online time-diary methodology allows one to track when activities are logged in a time diary. Hence, the recall period for each activity can be accurately estimated as the period between the ending time of an activity (i.e., a time actively entered by the respondent) and the logging time of the activity (i.e., a time stamp of the log recorded by the platform on which the time diary is logged).

Despite consensus on the association between recall bias and measurement error in time-use estimates, not much is known about the extent to which variations in the recall period of time-diary methodology itself affects time-use estimates. By using a large-scale, seven-day, online time-use survey from 2018, we aim to address this gap in knowledge by defining the recall period of activities as the time between the end time and log time of activities and associate it with quality indicators of time-use data.

BACKGROUND

Tracking How People Spend Their Time and Time-Diary Research

The purpose of time-use research is to study how people spend their time (Pentland et al., 2002). Direct observation of individual household activities is considered the most accurate way of obtaining data on time use(Juster, 1985a). The main advantage of direct observation is that it provides insights into activities that respondents do not consider worth recording, that are forgotten, or that are not recorded for reasons of social desirability. Yet direct observation has several significant disadvantages (Juster, 1985a). First, the continuous presence of an observer in the household is extremely expensive. Second, this observer's presence can influence the social dynamics of the environment and thus influence time allocation. Finally, data are captured from the observer's perspective, making the data subjective. At the other end of the spectrum, some

research uses stylized questions to collect time-use estimates (Juster et al., 2003; Kan & Pudney, 2008; Otterbach & Sousa-Poza, 2010); this method is significantly cheaper and has a lower respondent burden compared with direct observation. Yet this method has disadvantages related to cognitive problems regarding recall and summing issues that lead to measurement error and thus less valid estimates. Bateson (1984) thus considered time-diary methodology to be the golden mean between direct observation and stylized questions.

One of the most significant advantages of time diaries is that they can keep the recall period as limited as possible. In survey research, the recall period is commonly defined in the question, by asking respondents to recall what they have been doing for the past day, week, or month. In time-use research, however, the recall period can be defined as the period of time between the occurrence of activities and their recording for research purposes. Limiting this period reduces recall bias (Allen et al., 2020), making time-diary research the preferred method "for measuring phenomena that are highly susceptible to recall problems" (Niemi, 1993, p. 232), such as time spent working in the home environment (Niemi, 1993), unpaid work (Hill, 1985), and short-term tasks (Bonke, 2005).

Two different mechanisms are cited for the susceptibility of activities to recall problems. First, routine activities are systematically forgotten because of recall problems. Niemi (1993) demonstrated that passive kinds of physical exercise are forgotten, whereas visits to the cinema are easy to recall. In line with this logic, Michelson (2005) showed that there are differences in the extent to which weekdays and weekends can be recalled. Memory blending is a larger problem during weekdays, which are very similar to each other (see, e.g., van Tienoven et al., 2017). Second, with a longer recall period, activities that were incidental to daily time use (e.g., short, interrupting activities within larger activities, such as minor maintenance tasks) are often forgotten (Hill, 1985; Juster, 1986).

Testing the Data Quality in Time-Diary Research

Although direct observation is considered the "gold standard of reliability" (Vannest & Hagan-Burke, 2010, p. 131) for studies on time use, such studies are scarce because of the drawbacks mentioned earlier. Vannest and Hagan-Burke (2010) asked special education teachers to record the number of minutes spent per activity per 15 or 60 min of observation; they concluded, on the basis of a comparison of both methods, that teachers can reliably self-report their working time retrospectively. Gershuny et al. (2020) tested the suitability of self-reported time diaries for recording time-use patterns using cameras and accelerometers. They demonstrated that self-report time diaries are indeed a good data source for valid and reliable estimates of time-use patterns at both individ- ual and aggregate (sample) levels. These studies demonstrate the validity of time diaries by using direct, objective measures, whereas previous research only used indirect measures to support the validity of time diaries.

These indirect measures come from comparing diary data with findings from other methodologies. Robinson (1985) compared the results of time-diary data with findings of a study in which respondents were asked at a random moment by means of a beeper or a call to register what they were doing at that moment. As an alternative to these methods and absent comparative data, Juster(1985b, 1986) suggested five indicators of data quality. First, he assumed the number of logged activities is an indicator of data quality. The longer the recall period, the fewer activities will likely be effectively recalled. Second, he assumed that a longer recall period reduces the variation of activities (i.e., the number of different activities recorded). Third, this would also result in fewer secondary activities being recorded (i.e., activities that occurred simultaneously with the primary activity). Fourth, he considered "time not ascertained" an indicator: gaps in time diaries are interpreted as time periods that respondents cannot remember. Finally, he argued that the exact start and end times of an activity are hard to remember over time. This causes people to disproportionately record activities on whole and half hours, which is then a quality indicator.

Using these quality indicators, Juster (1985b, 1986) investigated the consequences of a longer recall period on time-diary data quality and compared recall with leave-behind time diaries. He defined the recall period rudimentarily as the time between the drop-off date of the time diary (i.e., the possibility to start the time diary) and the pickup date (i.e., the moment the diary needs to be completed). He showed that the length of the recall period in recall diaries has a substantial effect on the number of activities recorded on weekdays (Monday to Thursday): completed recall time diaries with a recall of longer than one day contained 10 percent to 20 percent fewer activities than diaries that were completed within the day; no differences were observed for weekends (Fridays to Sundays) (for similar findings, see Michelson, 2005). He found identical results for the number of different activities and no significant differences for three other quality indicators. Juster concluded that leave-behind time diaries lead to better data quality than recall time diaries and that data quality significantly declines after the recall period is extended beyond 24 hr.

According to Juster (1985b, 1986), these indicators for data quality are also influenced by personal characteristics (e.g., gender, age, family status) and characteristics related to time use (e.g., the number of hours worked, experienced time pressure). Paid work is often registered as a single activity and might lead to a smaller number of activities reported. In contrast, respondents with multiple responsibilities (i.e., family, work) might register more activities and a greater variety of activities. Such characteristics may be a consequence of real differences in behavior, but they may also be related to differences in the accuracy of registering activities. In Juster's studies, these characteristics are explicitly included as control variables only in terms of their effect on the quality indicators. However, his reasoning can also be used more comprehensively: personal characteristics and characteristics related to time use, such as number of hours worked and feelings of time pressure, might be directly related to the length of the recall period. People who experience time pressure may take longer before logging activities. Respondents with family responsibilities may give priority to

these responsibilities and thus let the recall period increase. To date, no study on recall bias has provided insight into this relationship.

The absence of substantial empirical evidence on the association between (the length of) the recall period and time-diary data quality indicators means that guidelines on conducting time-use surveys have been hesitant to recommend the maximum permissible recall period. The most recent edition of *Harmonised European Time Use Surveys (HETUS) 2018 Guidelines* (Eurostat, 2020) states that respondents should record their time diaries as close to the completion of activities as possible. If this cannot be attained—a maximum permissible recall period is not defined—assigning another date is preferable to allowing an extended recall period. Similarly, the *Guide to Producing Statistics on Time Use* (United Nations, 2004) is indecisive about whether to stick to the des- ignated day and allow a longer recall period or reassign the designated day.

In "Guidelines for Time Use Data Collection," Harvey (1993) concluded only that a recall period should not exceed two days and that empirical evidence is lacking to make more definitive recommendations on recall period. His evidence was limited to a study by Klevmarken (Keller et al., 1982) demonstrating that a 48-hour recall period leads to a smaller number of activities compared with a 24-hour recall period and research by the Survey Research Center (1984) demonstrating that leave-behind diaries generate 5 percent to 10 percent more activities than do 24-hour recall time diaries. Kalton (1985) went a little further and suggested to disallow a recall period of more than two to three days after a designated weekday, because these data are subject to memory decay. For weekend days this can be somewhat longer, as weekend days differ more from each other.

The indecisiveness of guidelines for time-use research on permissive recall period is due to limited empirical evidence on the association between the length of the recall period and data quality indicators and the use of rudimentary operationalizations of the recall period and ignorance of other influences on data quality indicators in the available evidence. Although leave-behind diaries are considered superior to recall diaries, researchers still do not know when entries are made in a leave-behind diary or the exact

recall period. Consequently, recall periods shorter than 24 hr have never been investigated. This lack of knowledge seems to downplay the argued strength of time-use research, that is, the close-to-real-time registration of activities.

However, today's online time-diary methodology allows a more profound investigation of the recall period. Each registered activity comes with a logged time stamp, which allows one to compare the end time and logged time of the activity. Online time-diary methodology can define recall periods not only to the minute but also for each activity separately. Using online time-diary data, we investigate the duration of the mean recall period and its association with data quality indicators. Specifically, we address the following research questions: (1) What are the duration and distribution of the mean recall period in online time-diary research? (2) How does the mean recall period differ according to socioeconomic characteristics? and (3) How does the duration of the recall period relate to quality indicators of time-diary data, net of variations by socioeconomic characteristics?

METHODS

Data

Data come from a 2018 time-diary study of teachers in Flanders, Belgium. This study was commissioned by the Flemish Ministry of Education and Training and conducted by the Research Group TOR of Vrije Universiteit Brussel between February 22 and May 15, 2018. All teachers in Flanders were eligible to participate in the study. A communications campaign recruited teachers through various channels, including teachers' unions, school management, magazines for teachers, and social media. The study took place online using the data collection platform MOTUS[1] (Minnen et al., 2020), which was developed to conduct time-use surveys. The study was available on the MOTUS Web and mobile applications, allowing respondents to register activities at any time using any devices they had available.

[1] More information on the MOTUS platform can be found at https://www.motusresearch.io.

The study consisted of four stages. First, teachers registered with their personal teacher numbers, which served as an exclusion mechanism for nonteachers and prevented repeated participation. Second, teachers completed a short prequestionnaire with limited socioeconomic questions. Third, they kept a seven-day (168-hour) time diary. Fourth, teachers completed a questionnaire about job characteristics immediately after finishing their time diaries.

Time Diaries

In their time diaries, teachers recorded all their activities using a predefined activity list. The activity list was presented to participants as a tree structure that contained 15 main categories, of which 9 were work related and 6 were not work related. The nine categories contained 49 detailed work-related activities. This list of activities was developed in consultation with an advisory committee consisting of representatives from teachers' unions, various educational networks (i.e., the governing body of Flemish education), and the Flemish Ministry of Education and Training. The other 6 main categories consisted of 24 detailed non-work-related activities (for the full activity list, see the Appendix in te Braak et al., 2022). Although respondents were asked to log their nonwork activities in addition to work activities, we limit ourselves in this article to work activities, because (1) logging nonwork activities, although encouraged, was not obligatory (only 50.4 percent of respondents reached the minimum criteria for sufficient data quality of nonwork activities), and (2) work activities were surveyed in much more detail than nonwork activities, which directly affects the data quality indicators. Working time includes all primary recorded working activities; commuting was excluded.

All activities were recorded in open intervals. Teachers provided the exact start and end time of each activity and were able to record activities as short as one minute in duration. The end time of the previous entry was suggested as the start time of a new entry. For each activity, respondents also recorded a possible secondary activity, where the activity took place, who was present, and if they enjoyed it. Only teachers who kept full seven-day time-diaries are included. The data set consists of 59,997 diary days of 8,571 teachers. To avoid effects of item nonresponse, we use only data from

respondents who do not have any missing variables that are crucial for this study (8,535 teachers, or 99.6 percent of the data set).

Concepts

The recall period is calculated as follows:

$$Mean_t = \frac{\sum_{i=1}^{n} t_{i\,registered} - t_{i\,ended}}{n_i},$$

where the mean recall period $mean_t$ is derived from dividing the sum of the recall periods for all registered activities during the seven-day period $i = 1$ to $i = n$ by the sum of recorded activities n_i. The recall period is calculated as the difference between the time stamp of the activity log $t_{i\,registered}$ minus the end time of the activity $t_{i\,ended}$. No maximum permissible recall period was applied. The actual maximum recall period is the period between the activity and the conclusion of the fieldwork (on May 15, 2018). For teachers who started in the first week of fieldwork (January 22, 2018), this theoretical maximum is thus almost four months; for respondents who participated toward the end of fieldwork, a theoretical maximum of two weeks was possible. In practice (see "Results" section), the recall period exceeded seven days for only 0.4 percent of all respondents.

We measure time-diary data quality using four of the five indicators proposed by Juster (1985b, 1986). Time not ascertained is excluded as a quality indicator, as we focus on work activities only, and time not ascertained does not necessarily mean one cannot remember what one did during this period but rather that one did something other than work. The number of activities is measured as the sum of individual work activities per respondent. The variety of activities is measured by the number of different work activities registered on the second level (49 work activities; see also the Appendix in te Braak et al., 2022). Unlike Juster (1985b, 1986), we do not use the number of secondary activities because of a strong correlation with the number of primary activities ($r = 0.69$). Instead, we use the percentage of primary work activities for which a secondary activity is registered. Finally, we construct the percentage of work activities that are registered exactly on the full or half hour.

Analytic Strategy

In the first step, we examine the duration and distribution of the mean recall period. Because respondents are asked to answer as close to real time as possible, we expect the mean recall period to be nonnormally distributed, with a peak close to zero and an extended upper tail. For the next steps, the mean recall period will therefore be constructed as dummies. The construction of these dummies is decided according to the results from this step.

In the second step, we examine how the mean recall period differs according to socioeconomic characteristics using multiple logistic regression analyses in which we compare the probability of answering within a certain recall period (the first dummy) compared with all other recall periods (all other dummies). Next, we take the first two dummies of the mean recall period (i.e., the two groups with the shortest recall period) together versus all other dummies, then the first three dummies versus all other dummies, and so on. We investigate for the following characteristics: sex (male, female), age (20–29, 30–39, 40–49, and ≥ 50 years), family status (living with parents, living alone, single parent, with partner without children, with partner and with children, other), the number of hours worked (centered on the mean of 43 hr 40 min), and experienced time pressure (a sum scale with minimum = –5 and maximum = +5; based on (van Tienoven, Minnen, et al., 2017).

In the final step, we examine the relationship between the mean duration of the recall period and the data quality by means of the four indicators mentioned above, net of socioeconomic characteristics, using multivariate linear regression analysis. Each quality indicator is presented in a separate regression model.

RESULTS

Description of the Mean Recall Period per Respondent

The population mean recall period for all respondents is 15 hr 30 min; the median is 8 hr 27 min. The maximum recall period during the fieldwork was 31 days. Figure 1 shows the distribution of the mean recall period where we cap the maximum recall period at

more than 168 hr. Among all respondents, 25.2 percent registered activities within 4 hr of completing them; 47.8 percent registered their work activities within 8 hr and 83.4 percent within 24 hr. Among all activities, 94.3 percent were registered within 48 hr and 97.2 percent within 72 hr. Thirty-five respondents, or 0.4 percent, took more than a week to register activities after completion.

Figure 1. Histogram of the number of respondents by the mean recall period, in hours (n = 8,535)

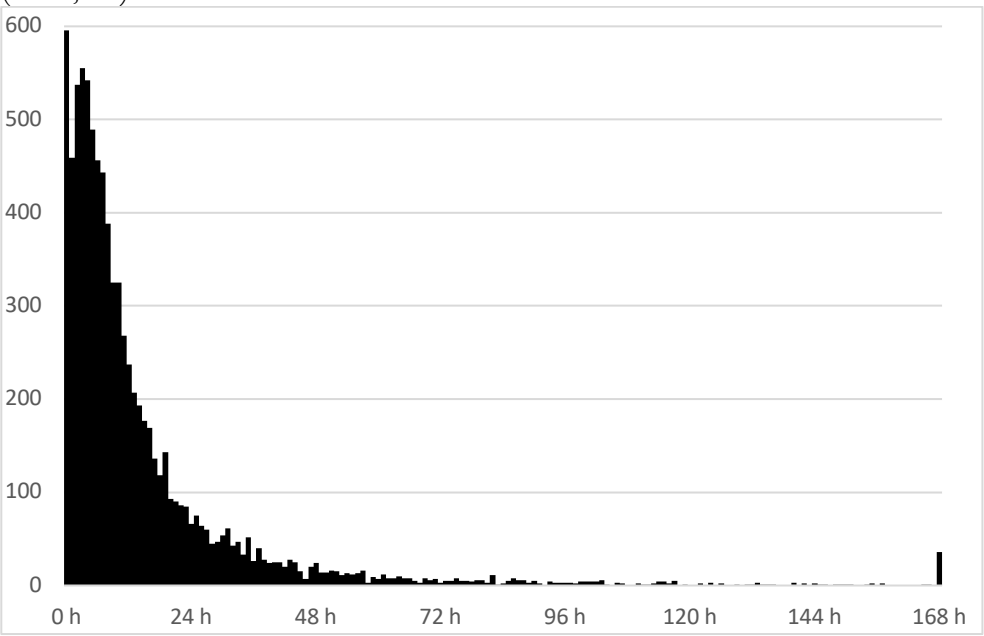


Figure 1 shows that most respondents had mean recall periods within 24 hr, which is the cutoff point for decreasing data quality suggested by Juster (1985b, 1986); only a few had mean recall periods longer than 48 hr (i.e., Harvey's [1990, 1993] cutoff point). Substantial variation occurs within a recall period less than 24 hr. For subsequent analyses, we therefore use four different cutoff points to create five groups: respondents with recall periods (1) less than 4 hr (the 25th percentile), (2) between 4 and 8 hr (~50th percentile), (3) between 8 and 24 hr (Juster's cutoff, 83th percentile), (4) between 24 and 48 hr (Harvey's cutoff, 94th percentile), and (5) beyond 48 hr.

Social Differences in the Length of the Recall Period

Table 1 shows the results of four logistic regression analyses that examine the stratification of the mean recall period by socioeconomic characteristics. Models show the odds ratios (ORs) for different recall periods. The first model compares the first dummy (0–4 hr) versus all other dummies (4 hr to a mean recall period of ≥ 48 hr). In the second model, we combine the first two dummies (0–4 and 4–8 hr) versus the other dummies (8 to ≥ 48 hr), and the third model shows the first three dummies (0–4, 4–8, and 8–24 hr) versus the other dummies (24 to ≥ 48 hr). This allows us to compare how the fastest responders (operationalized with a sliding scale of 0- to 4-hr, 0- to 8-hr, 0- to 24-hr, and 0- to 48-hr recall period, depending on the model) differ according to background compared with slower responders. The ORs in model 1 can be interpreted as a higher chance (i.e., OR > 1) or a lower chance (i.e., OR < 1) of registering an activity within 4 hr.

Model 1 shows significant differences according to socioeconomic characteristics. Age does not appear to be linearly related to whether people register within four hours: 30- to 39-year-olds have the highest odds of registering an activity within four hours. All other age groups are less likely to register an activity within this time frame (OR for 20–29 years = 0.80, OR for 40–49 years = 0.81, OR for ≥ 50 years = 0.65). Respondents living with their parents (OR = 1.51), living alone (OR = 1.35), or living with a partner without children (OR = 1.32) all have a substantially higher chance to register activities within four hours, on average, than do people living with a partner and children. Single parents are no different from people living with a partner and children regarding their recall period. In summary, the absence of children from the household is clearly related to registering within four hours. In addition, the model shows that, as the number of hours worked increases, the probability of registering within four hours decreases (OR = 0.99). The same applies to time pressure. The higher the time pressure, the smaller the OR (OR = 0.95) of registering activities within four hours. No effects were found by gender.

Table 1. Logistic Regression Models of the Mean Recall Period (n = 8,535)

				[CI 95%]	
	B	OR	Sig.	Lower	Upper
Model 1: 0-4 hours average recall period					
Constant	-1.014	0.363	***		
Gender *(ref.: female)*					
Male	0.042	1.043	n.s.	0.910	1.194
Age *(ref.: 30-39 years)*					
20-29 years	-0.223	0.800	**	0.683	0.937
40-49 years	-0.213	0.808	***	0.710	0.921
50+ years	-0.427	0.652	***	0.559	0.762
Family status *(ref.: with partner and child(ren)*					
Living with parents	0.414	1.513	***	1.191	1.922
Living alone	0.302	1.353	***	1.125	1.627
Single parent	0.115	1.122	n.s.	0.867	1.453
Partner, no children	0.277	1.319	***	1.139	1.527
Other	0.028	1.028	n.s.	0.683	1.548
Working time in hours *(centered to the mean)*	-0.015	0.985	***	0.981	0.989
Time pressure *(sum scale, min=-5, max=+5)*	-0.053	0.948	***	0.921	0.976
Nagelkerke R^2	.024				
Model 2: 0-8 hours average recall period					
Constant	-0.029	0.972	n.s.		
Gender *(ref.: female)*					
Male	-0.051	0.951	n.s.	0.845	1.070
Age *(ref.: 30-39 years)*					
20-29 years	-0.157	0.854	*	0.743	0.982
40-49 years	-0.084	0.920	n.s.	0.822	1.029
50+ years	-0.120	0.887	n.s.	0.780	1.010
Family status *(ref.: with partner and child(ren)*					
Living with parents	0.265	1.304	*	1.052	1.616
Living alone	0.280	1.323	***	1.122	1.561
Single parent	0.051	1.052	n.s.	0.845	1.311
Partner, no children	0.164	1.178	*	1.037	1.339
Other	0.096	1.101	n.s.	0.784	1.546
Working time in hours *(centered to the mean)*	-0.010	0.990	***	0.987	0.994
Time pressure *(sum scale, min=-5, max=+5)*	-0.075	0.928	***	0.904	0.952
Nagelkerke R^2	.016				

Notes. B=regression coefficient. OR=odds ratio. Sig.=significance. CI=confidence interval. Levels of significance: *** $p \leq 0.001$. ** $p \leq 0.01$. * $p \leq 0.05$. n.s. not significant

Table 1. Continued

	B	OR	Sig.	[CI 95%] Lower	Upper
Model 3: 0-24 hours average recall period					
Constant	1.596	4.935	***		
Gender *(ref.: female)*					
Male	-0.173	0.841	*	0.721	0.980
Age *(ref.: 30-39 years)*					
20-29 years	-0.078	0.925	n.s.	0.768	1.114
40-49 years	0.095	1.100	n.s.	0.947	1.277
50+ years	0.115	1.122	n.s.	0.942	1.335
Family status *(ref.: with partner and child(ren)*					
Living with parents	0.288	1.334	n.s.	0.992	1.793
Living alone	0.319	1.376	**	1.087	1.742
Single parent	-0.098	0.907	n.s.	0.683	1.204
Partner, no children	0.168	1.183	n.s.	0.994	1.407
Other	0.073	1.076	n.s.	0.681	1.699
Working time in hours *(centered to the mean)*	-0.002	0.998	n.s.	0.994	1.003
Time pressure *(sum scale, min=-5, max=+5)*	-0.062	0.940	***	0.908	0.973
Nagelkerke R^2	.006				
Model 4: 0-48 hours average recall period					
Constant	2.847	17.241	***		
Gender *(ref.: female)*					
Male	-0.355	0.702	**	0.556	0.886
Age *(ref.: 30-39 years)*					
20-29 years	-0.117	0.889	n.s.	0.665	1.189
40-49 years	0.248	1.281	*	1.003	1.638
50+ years	0.206	1.229	n.s.	0.927	1.629
Family status *(ref.: with partner and child(ren)*					
Living with parents	0.294	1.342	n.s.	0.839	2.148
Living alone	0.298	1.347	n.s.	0.920	1.973
Single parent	-0.308	0.735	n.s.	0.479	1.129
Partner, no children	0.121	1.129	n.s.	0.856	1.489
Other	-0.320	0.726	n.s.	0.384	1.373
Working time in hours *(centered to the mean)*	-0.007	0.993	n.s.	0.986	1.000
Time pressure *(sum scale, min=-5, max=+5)*	-0.098	0.907	***	0.858	0.959
Nagelkerke R^2	.012				

Notes. B=regression coefficient. OR=odds ratio. Sig.=significance. CI=confidence interval. Levels of significance: *** $p \leq 0.001$. ** $p \leq 0.01$. * $p \leq 0.05$. n.s. not significant

The results in model 2, which shows ORs for a mean recall period between zero and eight hours, are similar to model 1. The largest differences are found in terms of age: 20- to 29-year-olds are less likely than 30- to 39-year-olds to record their activities within eight hours (OR = 0.85); the other effects of age become insignificant. The effects for family status (OR for living with parents = 1.30, OR for living alone = 1.32, OR for

living with partner without children = 1.18) and working time in hours (OR = 0.99) remain very similar, but the effect size decreases slightly, and the effect size for time pressure increases slightly (OR = 0.93). Again, no effects were found by gender.

Model 3 compares the mean recall period between 0 and 24 hr with one of at least 24 hr and shows limited effects. Men have a lower probability than women to answer within 24 hr (OR = 0.84), and single persons have a higher probability (OR = 1.38). Also in model 3, the effect of time pressure is similar to the effects found in the previous models (OR = 0.94). No effects were found by age and working time in hours.

In model 4, we compare a mean recall period between 0 and 48 hr. Again, men have a substantially lower probability (OR = 0.70) of registering activities within 48 hr after the end of that activity. In addition, 40- to 49-year-olds (OR = 1.28) are more likely to register their activities within 48 hr after completing an activity than are 30- to 39- year-olds. Finally, the higher the time pressure, the lower the OR (OR = 0.91) of registering activities within 48 hr.

In short, the models consistently show that time pressure contributes to a longer recall period. The effects of other characteristics are somewhat more complex. Men have a longer mean recall period, as they are significantly more likely than women to record their activities only after 24 hr or later. Regarding age, 30- to 39-year-olds stand out because their propensity to register activities within 4 hr is higher than all other age groups, yet these effects subside as the recall period increases. Thus, they stand out mainly for their very short mean recall period. The same applies to respondents without children: they are more likely to register their activities very quickly (more often within 8 hr) than are respondents with children. Finally, an increasing number of working hours mainly results in respondents being less inclined to register their activities within 8 hr.

Data Quality and the Effect of the Duration of the Recall Period

Next, we examine the relationship between the mean duration of the recall period and four indicators of data quality: the number of logged work activities (Table 2), the number of different logged activities (Table 3), the percentage of primary work activities

with a secondary activity (Table 4), and the percentage of activities that started on the (half) hour exactly (Table 5). In each table, model 1 presents the association between the recall period and the quality indicator, and model 2 assesses this association net of socioeconomic characteristics.

In Table 2, model 1 reveals that the number of activities is strongly related to the mean recall period. Respondents with mean recall periods of 8 to 24 hr log an average of 38.5 activities. A mean registration within 4 hr after completion results in 3.3 more activities. A recall period of more than 24 hr, on the other hand, results in 4.3 fewer activities, and a recall period of more than 48 hr generates 6 fewer activities. Model 2 shows that, controlling for gender, age, family status, working time in hours, and time pressure, the effects of shorter recall periods increase: a mean recall period lower than 4 hr induces 5.5 more activities, and a period between 4 and 8 hr leads to 4 additional logged activities. A mean recall period of 24 to 48 hr results in 3.4 fewer activities being logged, and a recall period of more than 48 hr produces 6.5 fewer activities. The control variables also demonstrate that, on average, men register 2.9 fewer activities than women, and the number of activities decreases for the older age groups (2.2 fewer activities if 40–49 years old and 4.7 fewer activities if ≥ 50 years old). Also, a longer working time in hours is related to more activities. For each additional hour worked, the number of registered activities increases by 0.7.

Table 2. Linear Regression of Number of Registered Activities (n = 8,535)

	Model 1					Model 2				
				[CI 95%]					[CI 95%]	
	B	Sig.	β	Lower	Upper	B	Sig.	β	Lower	Upper
Constant	38.490	***		37.84	39.14	39.386	***		38.60	40.17
Average recall period *(ref.: 8-24h)*										
<4h	3.337	***	0.078	2.326	4.349	5.490	***	0.128	4.624	6.355
4-8h	3.181	***	0.072	2.138	4.225	3.992	***	0.090	3.105	4.878
24-48h	-4.273	***	-0.072	-5.617	-2.930	-3.420	***	-0.058	-4.561	-2.280
>48h	-6.044	***	-0.076	-7.792	-4.295	-6.455	***	-0.081	-7.941	-4.970
Gender *(ref.: female)*										
Male						-2.872	***	-0.057	-3.781	-1.962
Age *(ref.:30-39 years)*										
20-29 yrs						-0.038	n.s.	-0.001	-1.113	1.036
40-49 yrs						-2.150	***	-0.051	-3.018	-1.282
50+ yrs						-4.714	***	-0.096	-5.713	-3.715
Family status *(ref.: with partner and child(ren)*										
Living with parents						0.780	n.s.	0.010	-0.876	2.436
Living alone						-0.953	n.s.	-0.014	-2.228	0.322
Single parent						-0.451	n.s.	-0.005	-2.145	1.244
Partner, no children						-0.139	n.s.	-0.003	-1.125	0.847
Other						2.817	*	0.019	0.151	5.484
Working time in hours *(centered to the mean)*						0.737	***	0.517	0.710	0.763
Time pressure *(sum scale, min=-5, max=+5)*						0.158	n.s.	0.015	-0.039	0.355
Adjusted R^2	.025					.301				

Notes. B=unstandardized regression coefficient, Sig.=significance, b=standardized regression coefficient, CI=confidence interval. Levels of significance: *** $p \leq 0.001$, * $p \leq 0.05$, n.s. not significant

In Table 3, model 1 shows that the variety of activities for a mean recall period of 8 to 24 hr is 5.1 different activities, on average. This increases by 0.3 activities if they are recorded within 4 to 8 hr; it decreases if registered after a longer mean recall period (–0.2 activities at a recall period of 24–48 hr, –0.7 activities at ≥ 48-hr recall period). In model 2, which controls for effects of socioeconomic characteristics, these effects are largely preserved. A mean recall period of less than 4 hr is now significant and generates a 0.2 higher amount of variety. In addition, the model shows that men register 0.2 fewer different activities than women, the variety of activities is lower in the youngest age group (20–29 years), and people living with parents (–0.5), people living alone (–0.3), and single parents (–0.3) register a lower variety of activities compared with people with a partner and children. The number of hours worked is positively related to the variety of activities logged. For each additional hour worked, the variety increases by 0.05 activities.

Table 3. Linear Regression of Variety of Registered Activities (n = 8,535)

	Model 1					Model 2				
				[CI 95%]					[CI 95%]	
	B	Sig.	β	Lower	Upper	B	Sig.	β	Lower	Upper
Constant	5.076	***		4.988	5.164	5.199	***		5.078	5.320
Average recall period *(ref.: 8-24h)*										
<4h	0.063	n.s.	0.011	-0.075	0.201	0.238	***	0.041	0.104	0.372
4-8h	0.283	***	0.047	0.141	0.425	0.333	***	0.056	0.196	0.471
24-48h	-0.239	**	-0.030	-0.422	-0.057	-0.186	*	-0.023	-0.362	-0.009
>48h	-0.660	***	-0.061	-0.898	-0.422	-0.675	***	-0.063	-0.905	-0.445
Gender *(ref.: female)*										
Male						-0.183	*	-0.027	-0.324	-0.043
Age *(ref.:30-39 years)*										
20-29 yrs						-0.217	*	-0.036	-0.383	-0.051
40-49 yrs						-0.071	n.s.	-0.012	-0.206	0.063
50+ yrs						-0.051	n.s.	-0.008	-0.205	0.104
Family status *(ref.: with partner and child(ren)*										
Living with parents						-0.511	***	-0.048	-0.768	-0.255
Living alone						-0.289	**	-0.032	-0.486	-0.092
Single parent						-0.287	*	-0.023	-0.549	-0.025
Partner, no children						-0.133	n.s.	-0.021	-0.286	0.019
Other						-0.080	n.s.	-0.004	-0.493	0.332
Working time in hours *(centered to the mean)*						0.050	***	0.262	0.046	0.055
Time pressure *(sum scale, min=-5, max=+5)*						0.013	n.s.	0.009	-0.018	0.043
Adjusted R^2	.008					.078				

Notes. B=unstandardized regression coefficient, Sig.=significance, b=standardized regression coefficient, CI=confidence interval. Levels of significance: *** $p \leq 0.001$, ** $p \leq 0.01$, * $p \leq 0.05$, n.s. not significant

Model 1 in Table 4 shows that the mean percentage of registered activities accompanied by a secondary activity is 50.5 percent if the mean recall period is between 8 and 24 hr. This percentage decreases by 3.6 percentage points if the recall period is within 4 hr and by 2 percentage points if the recall period is within 4 to 8 hr. A recall period within 24 to 48 hr leads to an increase of 2.1 percentage points, and a recall period of more than 48 hr provides 5.2 percentage points more activities accompanied by a secondary activity.

These effects remain very similar when controlling for socioeconomic characteristics. Model 2 also shows that 40- to 49-year-olds log a slightly lower number of activities with a secondary activity (2.2 percentage point decrease), and higher time pressure leads to a slightly higher percentage of activities with a secondary activity (0.3 percentage point increase).

Table 4. Linear Regression of the Proportion of Activities with Secondary Activity (n = 8,535)

	Model 1					Model 2				
				[CI 95%]					[CI 95%]	
	B	Sig.	β	Lower	Upper	B	Sig.	β	Lower	Upper
Constant	0.505	***		0.497	0.513	0.516	***		0.504	0.528
Average recall period *(ref.: 8-24h)*										
<4h	-0.036	***	-0.067	-0.049	-0.023	-0.034	***	-0.063	-0.048	-0.021
4-8h	-0.020	**	-0.035	-0.033	-0.006	-0.019	**	-0.033	-0.032	-0.005
24-48h	0.021	**	0.028	0.004	0.038	0.022	*	0.029	0.004	0.039
>48h	0.052	***	0.051	0.029	0.074	0.051	***	0.050	0.028	0.073
Gender *(ref.: female)*										
Male						-0.013	n.s.	-0.020	-0.026	0.001
Age *(ref.:30-39 years)*										
20-29 yrs						-0.013	n.s.	-0.022	-0.029	0.004
40-49 yrs						-0.022	***	-0.040	-0.035	-0.009
50+ yrs						-0.007	n.s.	-0.012	-0.022	0.008
Family status *(ref.: with partner and child(ren)*										
Living with parents						-0.007	n.s.	-0.007	-0.032	0.018
Living alone						-0.002	n.s.	-0.002	-0.021	0.018
Single parent						-0.017	n.s.	-0.014	-0.042	0.009
Partner, no children						-0.006	n.s.	-0.010	-0.020	0.009
Other						0.018	n.s.	0.010	-0.022	0.059
Working time in hours *(centered to the mean)*						0.000	*	0.026	0.000	0.001
Time pressure *(sum scale, min=-5, max=+5)*						0.003	*	0.025	0.000	0.006
Adjusted R^2	.009					.012				

Notes. B=unstandardized regression coefficient, Sig.=significance, b=standardized regression coefficient, CI=confidence interval. Levels of significance: *** $p \leq 0.001$, ** $p \leq 0.01$, * $p \leq 0.05$, n.s. not significant

Model 1 in Table 5 presents the regression coefficients of the mean recall period on the percentage of activities whose starting time was logged at the exact (half) hour. The proportion of activities starting at the exact (half) hour increases steeply as the mean recall period increases. At a mean recall period of 8 to 24 hr, 30.1 percent of activities are logged with a starting time at the exact (half) hour. This decreases by 2.9 percentage points if activities are logged within 4 to 8 hr; it decreases by 10.1 percentage points if activities are logged within 4 hr of the end of an activity. When the recall period increases to 24 to 48 hr, the number of activities starting at the exact (half) hour increases by 3.9 percentage points, and a mean recall period of more than 48 hr produces 6.8 percentage points more activities logged with a starting time at the (half) hour. After controlling for socioeconomic characteristics in model 2, these effects remain. In addition, men log activities on the exact (half) hour slightly more often than women (+1.6 percentage points), as do people living with parents (+3.5 percentage points). People older than 50 years register 1.2 percentage points fewer activities on the (half) hour. For each additional hour worked, the number of activities logged on the exact (half) hour decreases slightly by 0.1 percentage points.

Table 5. Linear Regression of the Proportion of Activities that Start on the (Half) Hour Exactly (n = 8,535)

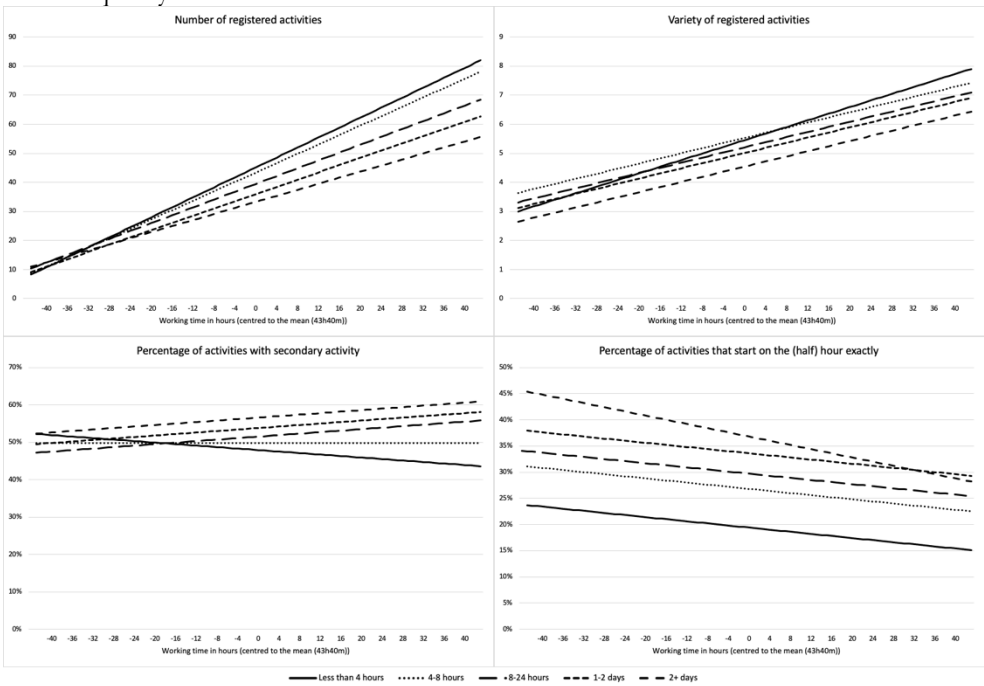
	Model 1					Model 2				
				[CI 95%]					[CI 95%]	
	B	Sig.	β	Lower	Upper	B	Sig.	β	Lower	Upper
Constant	0.301	***		0.296	0.307	0.299	***		0.291	0.307
Average recall period *(ref.: 8-24h)*										
<4h	-0.101	***	-0.265	-0.110	-0.093	-0.105	***	-0.274	-0.113	-0.096
4-8h	-0.029	***	-0.074	-0.038	-0.020	-0.030	***	-0.075	-0.039	-0.021
24-48h	0.039	***	0.073	0.027	0.050	0.038	***	0.072	0.026	0.050
>48h	0.068	***	0.096	0.053	0.084	0.068	***	0.095	0.053	0.083
Gender *(ref.: female)*										
Male						0.016	***	0.035	0.006	0.025
Age *(ref.:30-39 years)*										
20-29 yrs						-0.001	n.s.	-0.002	-0.012	0.010
40-49 yrs						-0.001	n.s.	-0.001	-0.009	0.008
50+ yrs						-0.012	*	-0.027	-0.022	-0.002
Family status *(ref.: with partner and child(ren)*										
Living with parents						0.035	***	0.049	0.018	0.052
Living alone						0.010	n.s.	0.017	-0.003	0.023
Single parent						-0.013	n.s.	-0.016	-0.030	0.004
Partner, no children						0.008	n.s.	0.018	-0.002	0.018
Other						-0.005	n.s.	-0.004	-0.032	0.022
Working time in hours *(centered to the mean)*						-0.001	***	-0.051	-0.001	0.000
Time pressure *(sum scale, min=-5, max=+5)*						0.000	n.s.	0.000	-0.002	0.002
Adjusted R^2	.095					.101				

Notes. B=unstandardized regression coefficient, Sig.=significance, b=standardized regression coefficient, CI=confidence interval. Levels of significance: *** $p \leq 0.001$, * $p \leq 0.05$, n.s. not significant

Interactions between the Mean Recall Period and Socioeconomic Characteristics

All models control for interaction effects between the mean recall period and all other characteristics. Consistent interactions were found only between the mean recall period and working time in hours. These effects are not shown in the tables but are summarized in Figure 2 (detailed results are available upon request from the corresponding author). The effects presented in the text with significance levels always refer to the main and interaction effect cumulated.

Figure 2. Interaction effects between mean recall period and working time, in hours, on data quality indicators



The number of registered activities increases as the recall period is shorter and the working time is longer (see Table 2). Recall periods of 8 to 24 hr generates effects of 0.68 ($p \leq .001$) additional activities registered per hour worked, respectively. The effects for respondents with a mean recall period of less than 4 hr (0.86; $p \leq .001$) and between 4 and 8 hr (0.81; $p \leq .001$) indicate that the shorter the recall period, the faster the number of logged activities increases when respondents work more hours. A recall

period of more than 48 hr induces an effect of 0.52 ($p \leq .01$) additional activities registered. In conclusion, the shorter the recall period, the faster the increase in activities per hour worked.

Regarding the variety of activities, we found a significant interaction effect if the recall period is shorter than 4 hr (0.06; $p \leq .01$). The effects for recall periods of 4 to 8 hr (0.05; not significant), 24 to 48 hr (0.05; not significant), and more than 48 hr (0.04; not significant) do not differ significantly from the reference recall period of 8 to 24 hr (0.04; $p \leq .001$). This interaction effect indicates that a short recall period produces a slightly faster increase in the variety of activities as the hours worked rises, compared with those with a prolonged recall period.

Per hour worked, the proportion of work activities with secondary activities decreases with a recall period of less than 4 hr –0,001; $p \leq .001$); a recall of 4 to 8 hr produces a stable effect (0.00; $p \leq .05$), which does not change as one works more hours. Recall periods of 24 to 48 hr (0.001; not significant) and more than 48 hr (0.001; not significant) do not differ significantly from the reference of 8 to 24 hr (0.001; $p \leq .001$). In other words, a short recall period produces a lower proportion of work activities with secondary activities, whereas a prolonged recall period generates a higher proportion of activities with secondary activities.

Finally, the percentage of activities registered on the exact (half) hour generates one significant interaction effect: a recall period of more than 48 hr induces a lower proportion of activities with a registered starting time on the (half) hour (–.002; $p \leq .05$) per hour worked. The effects of recall periods of less than 4 hr (–0.001; not significant), between 4 and 8 hr (–0.001; not significant), and between 24 and 48 hr (–0.001; not significant) do not differ significantly from a recall period of 8 to 24 hr (–0.001; not significant).

DISCUSSION

To study estimates of daily time use, time-diary methodology is generally recommended because it is specifically designed to account for recall bias. Yet time-diary methodology is not immune to recall problems. Limited empirical evidence suggests that the length of the recall period in time-diary research affects its data quality. However, these findings are based on rough estimates of the recall period, in which recall is calculated as the period between the drop-off or designated diary day and the return of the diary. As a result, not much is known about the actual recall period and its consequences for data quality. However, time-diary research is increasingly con- ducted using an online methodology that makes it feasible to ascertain more precisely when activities are registered using time stamps, allowing the recall period to be calculated more specifically per activity than was previously possible.

On the basis of these new opportunities, we investigated the mean duration of the recall period in online time-diary research, how the recall period differs according to socioeconomic characteristics, and how it affects time-diary data quality. Using online time-diary data from 2018 on teachers' work activities, in which 8,535 teachers recorded how they spent their time for one week, we showed that more than 4 out of 5 respondents in self-administered online time diaries logged their activities within 24 hr, on average, but there is a high degree of variation in the recall period within that 24 hr. Previously, this variation was undetected because of less accurate measurement methods of the recall period. Our results show that 1 out of 4 respondents logged their work activities within 4 hr, and 1 out of 2 respondents logged within 8 hr of finishing an activity.

The duration of the mean recall period relates to socioeconomic characteristics. The busier people are, the higher the mean recall period: a lower degree of experienced time pressure and fewer hours worked are both related to a shorter mean recall period. The effects of family status can also be understood within this logic. Respondents with children, whether single or with a partner, generally have a prolonged recall period; this period is shorter for participants living alone, with parents, or with a partner without

children. In other words, the larger the time constraints due to multiple social roles and responsibilities, the less likely one is to log activities close to real time.

A prolonged recall period is not without negative consequences. The effects of the mean recall period on various data quality indicators generally demonstrate that a prolonged recall period leads to lower time-diary data quality. In line with previous find ings, the results presented here also show that data quality decreases substantially as the recall period increases to 24 to 48 hr or more than 48 hrs. In addition, a shorter recall period results in a larger number of recorded activities, a slightly higher variety of activities, and a lower percentage of activities with a logged start time on the exact (half) hour. It is remarkable that data quality is substantially higher if the recall period is shorter than 8 hr. As a rule, a mean recall period of 4 hr results in even better data quality. We believe that these findings are a substantial addition to the existing literature on recall bias. Because of less precise parameters measuring the mean recall period, prior work assumed data quality declined substantially if the recall period was longer than 24 hr. The results presented here demonstrate that the loss of quality occurs much earlier, as soon as the mean recall period exceeds 4 hr.

To improve data quality in time-diary research, we recommended encouraging participants to log their activities as close to real time as possible by providing respondents with different platforms to use. Smart phones and tablets can enable the registration of work activities in the absence of a computer and during commutes and breaks, which can substantially shorten the recall period. In addition, automatic notifications on their devices can remind respondents to fill in their time diaries regularly. We recommended sending respondents a notification on their devices per part of the day (e.g., in the morning around 8 a.m., during lunch around 12:30 p.m., after work around 6 p.m., and shortly before going to sleep around 9 p.m.) to encourage them to limit the recall period to four hours. Passive registration is a further possible improvement to support recall. For example, movement can be detected by means of the global positioning system. This registered movement can then be provided to respondents to confirm in their time diaries, which prevents respondents from forgetting events because of longer recall periods.

The lack of substantial interaction effects also shows that a longer recall period has similar effects on data quality for different groups. The exception is the number of hours worked: the data quality measured in number and variety of activities increases faster per hour worked if the recall period is limited in duration. The importance of a short recall period is thus greater for respondents who work many hours and therefore must register more activities.

The quality indicator that examines the proportion of activities with a secondary activity initially shows contradictory results. Previous studies have argued that more secondary activities indicate better data quality; we argue, however, that the proportion of secondary activities is not a suitable indicator for data quality in time-diary research. Our results demonstrate that a shorter mean recall period is associated with a slightly lower proportion of activities accompanied by a secondary activity. In addition, the interaction effects show that the proportion of secondary activities increases more rapidly for respondents with a prolonged recall period than if the recall period is shorter. These findings, in conjunction with the preceding findings that the number of logged activities decreases with a prolonged recall period, indicate that some respondents log two different activities as a primary and secondary activity within one activity. Respondents who experience time pressure more often opt for an activity with a secondary activity instead of two activities. These findings suggest that the registration of secondary activities offers a preferred alternative for respondents with extended recall periods and those who must register a lot of activities because of longer working hours. It may be more difficult for these respondents to recall the exact start and end time for each individual activity, which may be why, aside from saving time, they prefer to log one activity with a secondary activity instead of two separate activities. However, removing the option of secondary activities in time-diary methodology is not recommended, as their logging is still useful for recording simultaneous activities (e.g., childcare supervision during other activities, reading while commuting). Yet we advise caution in including three or more simultaneous activities as an option in time diaries, because this may create less accurate entries.

The quality indicators regarding the number and variety of registered activities are strongly related to the taxonomy of activities surveyed in the time diary in combination with the number of working hours. In this study, we focused on working hours, as the time diaries were specifically designed to measure working time. In general time-diary research, working time is only surveyed in one or a few categories. Respondents who work many hours may log fewer and less different activities. The results presented here, as well as future results in similar studies on other time-diary data, can thus not be read as suggestions for benchmarks, because the interpretation of the numbers is clearly linked to the design of the time-diary methodology. The percentage of activities registered on the (half) hour is the only quality indicator that can possibly be used as a benchmark, as it is relatively stable across different designs of time diaries. Further research, including non-work-related time-diary research, will have to indicate what is considered a reasonable percentage.

This is one of the limitations of this study. It currently remains unknown to what extent the registration of nonwork activities, at moments other than usual working hours, is related to an extended recall period and data quality. A second limitation is that the study concerned a single professional group that is not a representative reflection of the population. In future time-diary research, more specific operationalizations of the recall period by means of time stamps should clarify the extent to which recall period is related to data quality in the general population. This would also allow one to examine the extent to which a prolonged recall to log nonwork activities for specific subgroups affects the data quality. The forthcoming online time-diary study among Belgians in 2022 and 2023 will provide suitable data for this purpose.

CONCLUSION

The duration of the mean recall period is strongly related to data quality in time-diary research. In contrast to previous research, our findings show that data quality already decreases if respondents log activities after four hours. In particular, respondents with many time constraints, such as parents and people with longer working hours, have a prolonged recall period. To substantially shorten the recall period, and thus increase

data quality in online time-diary research, we recommend providing options to log activities as soon as possible, such as via smart phones and tablets, as this would allow activities to be logged at convenient moments (e.g., during breaks and commutes, and between activities), as well as notifying respondents to log their activities close to real time.

REFERENCES

Allen, R., Benhenda, A., Jerrim, J., & Sims, S. (2020). New evidence on teachers' working hours in England. An empirical analysis of four datasets. *Research Papers in Education*, *36*(6), 657–681. https://doi.org/10.1080/02671522.2020.1736616

Bateson, N. (1984). *Data Construction in Social Surveys*. Allen & Unwin.

Bonke, J. (2005). Paid work and unpaid work: Diary information versus questionnaire information. *Social Indicators Research*, *70*(3), 349–368. https://doi.org/10.1007/s11205-004-1547-6

Eurostat. (2020). *Harmonised European Time Use Surveys (HETUS) 2018 Guidelines. Re-edition. 2020 edition*. Publications Office of the European Union. https://doi.org/10.2785/160444

Gershuny, J., Harms, T., Doherty, A., Thomas, E., Milton, K., Kelly, P., & Foster, C. (2020). Testing Self-Report Time-Use Diaries against Objective Instruments in Real Time. *Sociological Methodology*, *50*(1), 318–349. https://doi.org/10.1177/0081175019884591

Harvey, A. S. (1990). Time use studies for leisure analysis. *Social Indicators Research*, *23*, 309–336. https://doi.org/10.1007/BF00572527

Harvey, A. S. (1993). Guidelines for time use data collection. *Social Indicators Research*, *30*, 197–228. https://doi.org/10.1007/BF01078728

Hill, M. S. (1985). Investments of time in houses and durables. In F. T. Juster & F. P. Stafford (Eds.), *Time, Goods, and Well-Being* (pp. 205–244). Institute for Social Research, University of Michigan.

Juster, F. T. (1985a). Conceptual and Methodological Issues Involved in the Measurement of Time Use. In F. T. Juster & F. P. Stafford (Eds.), *Time, Goods, and Well-Being* (pp. 19–31). Institute for Social Research, University of Michigan.

Juster, F. T. (1985b). The Validity and Quality of Time Use Estimates Obtained from Recall Diaries. In F. T. Juster & F. P. Stafford (Eds.), *Time, Goods, and Well-Being* (pp. 63–91). Institute for Social Research, University of Michigan.

Juster, F. T. (1986). Response errors in the measurement of time use. *Journal of the American Statistical Association*, *81*(394), 390–402. https://doi.org/10.1080/01621459.1986.10478283

Juster, F. T., Ono, H., & Stafford, F. P. (2003). An assessment of alternative measures of time use. *Sociological Methodology*, *33*, 19–54. https://doi.org/10.1111/j.0081-1750.2003.t01-1-00126.x

Kalton, G. (1985). Sample Design Issues in Time Diary Studies. In F. T. Juster & F. P. Stafford (Eds.), *Time, Goods, and Well-Being* (pp. 93–112). Institute for Social Research, University of Michigan.

Kan, M. Y., & Pudney, S. (2008). Measurement Error in Stylized and Diary Data on Time Use. *Sociological Methodology*, *38*, 101–132. https://doi.org/10.1111/j.1467-9531.2008.00197.x

Keller, J., Kempter, D., Timmer, S. G., & Young-Demarco, L. (1982). *Proceedings of the International Time Use Workshop*.

Michelson, W. M. (2005). *Time use: Expanding the explanatory power of the social sciences.* Paradigm.

Minnen, J., Glorieux, I., Tienoven, T. P. van, Weenas, D., Deyaert, J., Bogaert, S. van den, & Rymenants, S. (2014). Modular Online Time Use Survey (MOTUS) – Translating an existing method in the 21 st century. *Electronic International Journal of Time Use Research*, *11*(1), 73–93. https://doi.org/dx.doi.org/10.13085/eIJTUR.11.1.73-93

Minnen, J., Nagel, E., & Sabbe, K. (2020). *SOURCE TM: Software Outreach and Redefinition to Collect E-data Through MOTUS. Towards a Modular Online Time Use Survey*. https://ec.europa.eu/eurostat/documents/9986036/10970806/847218_BE_final+report_Main+document.pdf (Accessed August 2021)

Niemi, I. (1993). Systematic error in behavioural measurement: Comparing results from interview and time budget studies. *Social Indicators Research*, *30*(2–3), 229–244. https://doi.org/10.1007/BF01078729

Otterbach, S., & Sousa-Poza, A. (2010). How accurate are German work-time data? a comparison of time-diary reports and stylized estimates. *Social Indicators Research*, *97*(3), 325–339. https://doi.org/10.1007/s11205-009-9504-z

Pentland, W. E., Harvey, A. S., Lawton, M. P., & Mccoll, M. A. (2002). *Time Use Research in the Social Sciences*. Kluwer Academic.

Robinson, J. P. (1985). The Validity and Reliability of Diaries versus Alternative Time Use Measures. In F. T. Juster & F. P. Stafford (Eds.), *Time, Goods, and Well-Being* (pp. 33–62). Institute for Social Research, University of Michigan.

Sonnenberg, B., Riediger, M., Wrzus, C., & Wagner, G. G. (2012). Measuring time use in surveys - Concordance of survey and experience sampling measures. *Social Science Research*, *41*(5), 1037–1052. https://doi.org/10.1016/j.ssresearch.2012.03.013

Survey Research Center. (1984). *Proposal to the National Science Foundation for support of a 1985-86 Study of Time Allocation Among American Households*. Institute for Social Research, The University of Michigan.

te Braak, P., van Droogenbroeck, F., Minnen, J., van Tienoven, T. P., & Glorieux, I. (2022). Teachers' working time from time-use data: Consequences of the invalidity of survey questions for teachers, researchers, and policy. *Teaching and Teacher Education*, *109*, 1–15. https://doi.org/10.1016/j.tate.2021.103536

United Nations. (2004). *Guide to Producing Statistics on Time Use: Measuring Paid and Unpaid Work*. United Nations Publication.

van Tienoven, T. P., Glorieux, I., & Minnen, J. (2017). Exploring the stable practices of everyday life: A multi-day time-diary approach. *Sociological Review*, *65*(4), 745–762. https://doi.org/10.1177/0038026116674886

van Tienoven, T. P., Minnen, J., & Glorieux, I. (2017). *The statistics of the time pressure scale*. Vrije Universiteit Brussel, Onderzoeksgroep TOR.

Vannest, K. J., & Hagan-Burke, S. (2010). Teacher time use in special education. *Remedial and Special Education*, *31*(2), 126–142. https://doi.org/10.1177/0741932508327459

Westat. (2013). *American Time Use Survey Substitution of Days. Final Report*. Westat. http://bls.gov/tus/ (accessed July 2022)