International Journal of Public Opinion Research

doi:10.1093/ijpor/edv032

Research Note

Reducing Underreports of Behaviors in Retrospective Surveys: The Effects of Three Different Strategies

Peter Lugtig[1], Tina Glasner[2], and Anja J. Boevé[3]

[1]Department of Methods and Statistics, Utrecht University, P.O. Box 80.140, 3508 TC Utrecht, The Netherlands; [2]University of Humanistic studies, Utrecht, The Netherlands; [3]Faculty of Social and Behavorial Sciences, University of Groningen, Groningen, The Netherlands

Longitudinal social science surveys typically collect data at regular intervals. In most ongoing panel surveys, the time between two consecutive waves of measurement is 1 year. This interval is often chosen because year-to-year changes often suffice to answer the research questions of interest. This article focuses on the accuracy of data that are collected retrospectively on events that occur between two interviews: the use of the medical services of a Family Physician (FP). Surveys like the Survey of Health and Retirement in Europe (Börsch-Supan et al., 2013) and the Health and Retirement Survey (Wallace & Herzog, 1995) annually ask the survey question:

"During the last 12 months, about how many times in total have you seen or talked to a medical doctor about your health?"

Earlier studies on the quality of such survey reports have found substantial inaccuracies in retrospective reports of behavioral frequencies that were caused by underreporting, overreporting, or a combination of the two.

Respondents can follow different strategies to answer retrospective questions. One strategy is to try to recall every specific behavioral event along with details of such events. Another strategy, used more often, is to estimate the frequency of events without recalling every event specifically (Conrad, Brown, & Cashman, 1998; Schwarz, 1990). Both strategies involve memory retrieval, which is often followed by a process of adding up or averaging the behavioral frequencies and giving an answer. At each of these stages, respondents can make mistakes.

When underreporting occurs in retrospective interviews, this is often assumed to be because of failure to recall events, although other factors such as social desirability or inefficient estimation strategies also play a role. Earlier studies have shown that temporally distant events are less likely to be recalled, and, thus, remain

All correspondence concerning this article should be addressed to Peter Lugtig, Department of Methods and Statistics, Utrecht University, P.O. Box 80.140, 3508 TC Utrecht, The Netherlands. E-mail: p.lugtig@uu.nl

underreported (Belli, 2001; Pascale, Roemer, & Resnick, 2009; Wagenaar, 1986). Even when the recall period is short, behaviors can remain underreported when the behavior is not salient to the respondent (Becker, 2003; Wagenaar, 1986).

Overreporting can be affected by social desirability bias, but it is often because of forward telescoping—the fact that respondents report events as having occurred more recently than they occurred in reality. As with underreports, telescoping occurs more often when the recall period is longer (Belli, 2001).

This article reports the results of an experiment with different question formats to estimate the annual use of a FP. We compare the estimates based on three different questions that use different reference periods or recall aids.

Improving Retrospective Recall

There are generally three ways to improve the quality of retrospective questions. The recall and reference period can be shortened, or the respondent can be assisted in the recall task by the survey researcher. Shortening the recall period can be done by asking about events that occurred in the past month instead of the whole year, simplifying the recall task. If annual estimates are of interest, this requires asking respondents such a question each month or relying on extrapolation.

Another approach currently used by many panel studies is to shorten the reference period to 1-month periods, while keeping a recall period of 1 year. By forcing the respondent to think about what happened in 12 separate months, recall should be facilitated and more events recalled.

Another way to improve recall is to use recall aids. Two of the most frequently used aids are timelines (Belli, 1998) and landmarks (Loftus & Marburger, 1983). Timelines consist of one or a few lines of time units, such as weeks or months, in which life events for the reference period can be recorded (Westerberg, 1998). The wide use of timelines has included diary studies (Sobell, 2001) and retrospective reports spanning months (Westerberg, 1998) or even decades (Van Der Vaart, 2004).

Timelines have been found to reduce omissions because of forgetfulness, as well as the magnitude of telescoping error (Belli, 2001). This finding does not depend on the length of the total reference period, although studies have noticed that the longer the reference period, the less effective the timeline is in aiding recall (Glasner & Van der Vaart, 2009). Furthermore, timelines have been found to work better when one timeline incorporates monthly states on multiple but related variables of interest. As job status changes often coincide with address changes for example, timelines that combine monthly records of both job status and household address will yield higher data accuracy (Glasner & Van der Vaart, 2009).

Timelines are often used in combination with landmarks, which are typically autobiographical events for which the respondent remembers the date (Van der Vaart & Glasner, 2011). After the respondent has identified some landmark events within the reference period, a timeline is then used as a visual aid to connect the dates of landmark events to the variable of interest. Landmarks might serve as a cognitive cue, but only when they are closely connected to the subject of interest (Callegaro, 2008; Van der Vaart & Glasner, 2011).

In this study, we use a within-subjects design to compare estimates obtained by:

1. A 1-month retrospective question asked each month over a 12-month period. This format includes no additional recall aids.
2. An annual retrospective question. This format includes a 1-year reference period and no recall aids
3. A 12-month retrospective question with timeline. This format includes a 1-year recall period and twelve 1-month reference periods. To facilitate recall, a 12-month timeline was used. In this question, a between-factor experiment was embedded in which respondents were randomly assigned to receive landmarks or not.

The 1-month retrospective question should lead to the fewest underreports because both the recall and reference period are 1 month. Underreports should be higher in the 12-month retrospective question, where the reference period is 1 month but the recall period is 12 months. In the annual retrospective question, underreports should be highest.

These effects on underreports may be offset to some degree by an increase in overreports because of forward telescoping. In the 1-month retrospective question, the annual estimate of FP use is constructed with 12 monthly variables. If forward telescoping occurs in each month, there is a risk of overestimating annual FP use. We expect forgetting because of recall and estimation problems of events to outweigh forward telescoping in all question formats.

Because of this, we expect the estimate of annual FP use to be highest in the 1-month retrospective question, followed by the 12-month retrospective question, and then the annual retrospective question.

Finally, we expect that landmarks in combination with a timeline result in a higher estimate compared with the sole use of a timeline.

Methods

Data

The data for our study stem from the Longitudinal Internet Studies for the Social Sciences (LISS).[1] This panel was started in 2007, and respondents were interviewed monthly on a wide range of topics. Panel members were recruited using a simple random sample of Dutch households, who were contacted using a mixed-mode design. After initial contact, all household members were asked to participate in the panel survey, and those without Internet access were provided it for the study. The panel recruitment rate in Wave 1 amounted to 49% (AAPOR, 2009; Callegaro & DiSogra, 2008). Among the initial respondents, respondents >75 years of age were underrepresented, as well as respondents living in single households, respondents renting (as opposed to being a homeowner), and respondents from highly urban areas (Leenheer & Scherpenzeel, 2013; Lugtig, Das, & Scherpenzeel, 2014). For this reason, we weighted our analyses on age, household composition, home ownership, and urbanicity. Because of a small number of respondents >75 years, we chose to limit our analyses to people between the ages of 15 and 75.

[1]More information about the recruitment of the panel, response percentages for all waves, and the full questionnaires used in this study, can be found on www.lissdata.nl.

Instruments

In this study, we compare three versions of survey questions that can be used to estimate annual visits to a FP. We use a within-subjects design, so that each respondent receives every question format.

The annual retrospective question asked about the frequency of using the medical services of a FP in the past 12 months and was administered in November 2011 to all LISS panel respondents ($n = 6{,}533$). The wording of the question was: "How often did you use the following health services over the past 12 months?" This question was followed by a number of medical services among which the first was "Family Physician." Any answer between 0 and 999 was allowed. Although the question did not stress that only personal FP visits had to be included, the entire questionnaire only asked about personal health issues, thus excluding visits to accompany family members.

For the 1-month retrospective question, a random subset of 1,511 respondents from the LISS panel was invited on the first Monday of every month from January 2011 until December 2011 to complete a short questionnaire about their perceived health, weight, and use of medical services, including FP use. These respondents received the same question in every month: "How often in the past 4 weeks did you use the following health services for yourself?... Family physician," with answers between 0 and 100 allowed.

Third, for the 12-month retrospective question, the same 1,511 respondents who completed the 1-month retrospective question were asked in January 2012 to complete a questionnaire about their health in the past year. To help respondents recall and reconstruct their FP use in every month, respondents were shown a timeline as depicted in the lower panel of Figure 1. The instruction accompanying the timeline was: "On the timeline below, can you please indicate per month how often, in the past year, you used the following health service for yourself?"

In the landmarks condition, the timeline was preceded by the question: "Have there been any events during the past year that have affected your health? If so, could you please indicate what happened to you, and when that happened? If you do not remember the exact date, you can also just indicate in which month it happened." There was room to report six possible landmarks, with a date in DD/MM format. After completing this page, the landmarks were fed forward to a next page where the landmarks with their dates were included in the timeline for respondents to complete (see Figure 1).

In summary, we have three sets of information from 1,511 respondents about FP use:

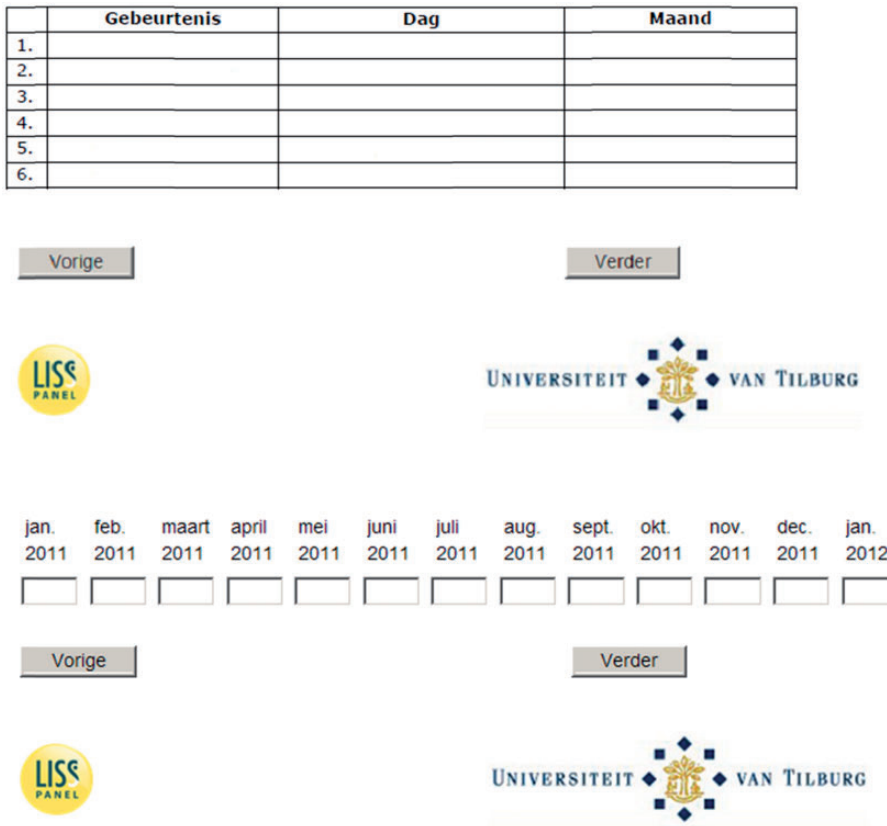
1. An annual retrospective question asked in November 2011.
2. A 1-month retrospective question asked every month from January to December 2011.
3. A 12-month retrospective question asked in January 2012 that includes a timeline and an experimental design using landmarks.

Missing Data

Because the 1-month retrospective FP data were collected in 12 separate waves of data collection, missingness was most problematic for this question. Table 1 shows the

Figure 1

Landmarks and timeline as implemented in the Longitudinal Internet Studies for the Social Sciences panel (original Dutch version). Landmarks and timelines were shown on separate pages, but all landmarks that were entered by respondents on the first page were fed forward to the second page



observed sample sizes for each question format. Missing data amount to 12% for the annual retrospective question, about 6% for the 12-month retrospective question, and ranges from 9% in January 2011 to 24% in December 2011 for the 1-month retrospective question. If some cells in the 12-month retrospective question were left empty, we assumed a value of 0.

We performed multiple imputation for the missing FP visits for all three question formats simultaneously using the Predictive Mean Matching algorithm as implemented in the MICE package in R (van Buuren & Groothuis-Oudshoorn, 2011), using all dependent variables and all predictors used in a model later in the article (see Table 2). We computed an annual estimate for the 1-month and 12-month retrospective questions by summing the monthly estimates to an annual estimate.

The estimate for annual FP use has a correlation of .01 with a count of the months with valid data for the 1-month retrospective data. This implies that cases with missing FP use did not underreport FP use more or less often than respondents

Table 1
Sample Sizes for Annual Retrospective, 12-Month, and 1-month Retrospective Questions about Use of a Family Physician

Question type	Month	Sample size	Percent missing (%)
1-month retrospective	January	1,334	9
	February	1,305	14
	March	1,294	14
	April	1,264	16
	May	1,248	17
	June	1,230	19
	July	1,229	19
	August	1,271	16
	September	1,217	19
	October	1,277	15
	November	1,178	22
	December	1,141	24
Annual retrospective		1,334	12
12-month retrospective		1,427	6

Note. $N = 1{,}511$. The landmark and no landmark conditions were only implemented in the 12-month retrospective question. $n_{\text{no landmarks}} = 758$; $n_{\text{landmarks}} = 753$.

with complete data. We conducted sensitivity analysis varying the covariates for the imputation model and found similar results using different specifications.

Results

Retrospective Annual, 12-Month Retrospective, and 1-Month Retrospective Estimates

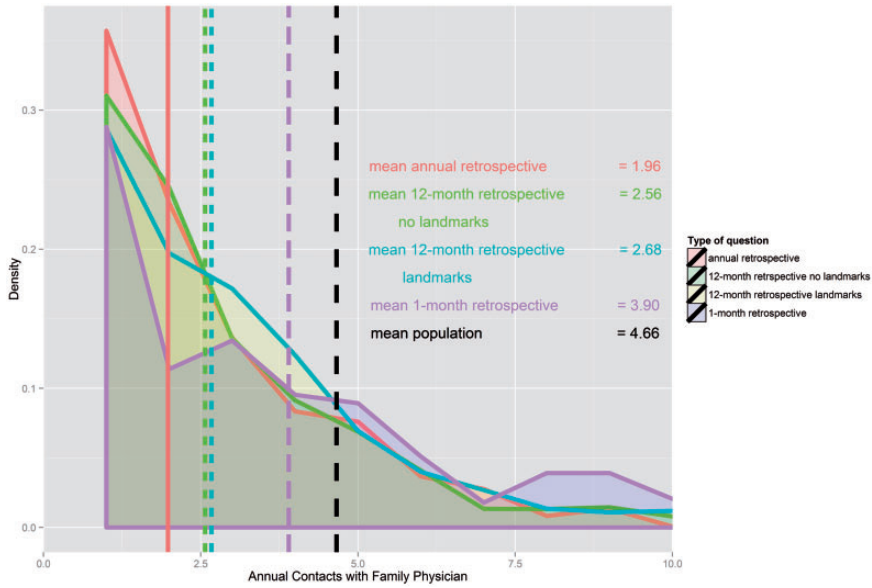
Annual estimates of FP use vary greatly depending on whether the annual retrospective, 12-month retrospective, or 1-month retrospective survey questions are used. Figure 2 shows that the annual retrospective and 1-month retrospective questions generate a different distribution and mean estimate of annual FP visit, with the 12-month retrospective estimate in between these two.

The lowest estimate is found for the annual retrospective estimate ($M = 1.96$), followed by the 12-month retrospective estimate without landmarks ($M = 2.56$) and the 12-month retrospective estimate with landmarks ($M = 2.68$). The highest estimate is found for the 1-month retrospective questions ($M = 3.90$).

Effect of Landmarks on Inconsistency between 12-Month and 1-Month Retrospective Questions

Because the estimates of annual FP use from the three survey questions vary greatly, we conducted further analyses on the difference in reports between the 12-month retrospective and 1-month retrospective questions. Both questions ask about monthly FP use; yet, our estimates differ greatly because of the difference in the recall period.

Figure 2
Smoothed density plots for four survey estimates and one population record estimate of annual use of a Family Physician



In comparing these question formats, we look specifically at the effect of our experiment. Did the landmarks change retrospective reporting of monthly FP use in the 12-month retrospective question?

The prevalence (yes or no) of FP use is modeled instead of the frequency of use. If the monthly data between the 1-month and 12-month retrospective questions are consistent, we code this as (0), and, if the data differ, we code an inconsistency (1). We conservatively assumed that if monthly data were missing for the 1-month or 12-month retrospective question, the data were consistent. We checked whether inconsistencies because of overreports in month t could be related to underreports in either month $t+1$ or $t-1$, implying that respondents do recall events without error, but that the inconsistency is introduced in the dating of the FP use. Our results using these two definitions were however similar.

To explain the total annual difference between the reports, we used a quasi-Poisson regression model,[2] with the sum of inconsistencies in 1 year (ranging from 0–12) as the dependent variable. In a first model, we only used the experimental condition (landmarks or no landmarks) as a predictor, and, in a second model, we added respondent-focused covariates, which could be related to both the frequency (and inconsistency) in FP visits: gender, age, age squared, subjective health status in

[2]We started fitting the model as a Poisson regression model, but because of overdispersion in the absolute difference between the 12-month and 1-month retrospective reports, we chose to model a quasi-poisson model where standard errors are corrected for overdispersion.

Table 2
Quasi-Poisson Regression Explaining Sum of Absolute Difference Between 12-Month Retrospective and 1-Month Retrospective Reports

Predictor Variable	Model 1		Model 2		Model 3	
	B	*SE*	*B*	*SE*	*B*	*SE*
Intercept	.643*	.078	.012	.267	−.073	.269
Condition (1 = landmarks)	.118	.048	.099	.043	.058	.045
Gender (1 = female)			.112	.044	.088	.044
Age/10			−.137	.072	−.106	.072
Age/10 (squared)			.023*	.007	.018	.007
Health (ref = bad)						
Moderate			.473*	.156	.488*	.156
Good			.372*	.158	.412*	.158
Very good			.136	.172	.201	.173
Excellent			.005	.236	.083	.236
Education (ref = primary school)						
Basic secondary			−.094	.074	−.093	.074
Middle secondary			−.160	.093	−.182	.093
Vocational tertiary			−.052	.078	−.054	.078
Bachelor degree			−.057	.080	−.065	.079
Masters degree			−.076	.107	−.095	.107
Frequency of doctor visits (average of 12-month and 1-month formats)			.061*	.004	.058*	.004
Duration (min)					.017	.007
Difficulty					.066*	.016
Residual deviance (df)	2,459.7 (1,509)		1,986.4(1,496)		1,957.1 (1,494)	
Pseudo R^2	.004		.195		.207	

Note. Cell entries are coefficients (*B*) and standard errors (*SE*) from a quasi-poisson regression model. $N = 1{,}511$, Deviance (*df*) of null model (H_o) = 2,468.4 (1,510).
*Significant with $\alpha = .01$. Pseudo R^2 is computed as 1—[Deviance (H_1)/Deviance (H_o)]. In addition to the models shown, we also fitted a model where interactions between the experimental conditions and all other predictors were added to the model. This resulted in a model with a Deviance of 1,950.10 (*df* = 1,487). We found no significant interaction effects.

January 2012, highest level of education, and the average frequency of FP visits across the two formats. In a third model, we added two variables that are likely to interact with our experimental manipulation: the reported difficulty of the questionnaire as evaluated by the respondent and the total duration for completing the survey. All analyses were conducted using R 3.1.1 (R Core Development Team, 2014).

The results shown in Table 2 indicate that the effect of the landmarks is nonsignificant in all three models. The estimate of landmarks is, however, always

positive, implying that, if anything, landmarks lead to more inconsistencies between retrospective 12-month and 1-month reports. The estimate of the effect of the landmarks becomes smaller when perceived difficulty is added as a predictor to the model. Respondents who received the landmarks found the questionnaire more difficult (M (SE) in no landmarks = 1.79 (0.05), landmarks = 1.99 (0.05), $t(1,503.5) = -3.12$, $p < .01$).

Adding covariates did not lead to larger explanatory power of the models. The pseudo R^2 of the most elaborate model is .21. Age has a curvilinear effect on total inconsistencies: It first decreases, but it increases again after about age 60. People who report to be in moderate health make more errors. The frequency of FP visits itself is also a predictor of inconsistencies. The more often a respondent visits the FP, the more likely he or she is to underreport. Finally, we see that the time respondents take to complete the survey does not predict inconsistencies in FP visits.

Conclusion and Discussion

Asking respondents to estimate the number of annual FP visits resulted in different estimates of annual FP use, depending on the design of the question. To evaluate which question format was most accurate, we compared our estimates with register data from Statistics Netherlands on FP use. The register contains data from about 125 FPs in the Netherlands, covering details on FP use and diagnoses of about 300,000 patients (De Bruin et al., 2003).[3] The register data included (1) appointments at the physician's practice, (2) at-home appointments, and (3) telephone appointments. We used the combined total.[4] The patient data were subsequently matched to the population registry containing demographic data on all Dutch citizens. After matching, the data from the 300,000 patients were weighted to the Dutch population using the following variables: gender, age, ethnicity, urbanicity, income, and whether a person moved in the previous year (De Bruin et al., 2003). The official statistics were then split by age group, and we only used the population statistics for the age range of our sample (age 15–75 years).

Two of three question designs that we tested among the same respondents led to much lower estimates of annual FP use than the mean derived from the population records, which was 4.66 instances per year.

In the annual retrospective estimate, FP use remained 58% underreported compared with our population estimate. The underestimate was reduced to 44% using 12-month retrospective reports and further to 16% when the annual estimate for FP use was based on 1-month retrospective reports. Therefore, we conclude that

[3]The data we used are for 2010 instead of 2011. However, the total estimate of FP use in the population does not vary greatly over years, as shown in the records.

[4]If telephone appointments were to be excluded from the population total, the population estimate would be 3.62. The survey questionnaire did not explicitly instruct respondents whether to include telephone appointments in their estimate. Without including telephone appointments, the annual retrospective is 54% of the population estimate, the 12-month retrospective 72%, and the 1-month retrospective 108% of the population estimate.

shortening the reference period from 1 year to 1 month greatly reduces underreports, but not to the extent that underreports disappear.

Most existing panel surveys use the annual retrospective format for asking about medical use. This is likely to lead to large underreports of FP use in those surveys. Our findings likely extend to other uses of medical services that all use the same question format. Whether our results extend to questions that ask respondents to recall facts and behaviors more generally remains to be seen. Visits to a FP are probably salient events only to respondents in bad health, which could be one reason why we find these respondents to report more consistently than those in moderate health. If behaviors are more salient to respondents generally, underreports may be not as problematic as in this study. Still, panel surveys routinely ask respondents to recall monthly details on not-salient events that are nonetheless important for scientists and policymakers.

The fact that the 1-month retrospective estimate comes closest to our population estimate does not necessarily mean that the data accuracy of this question is best. The higher annual estimate in FP use in the 1-month retrospective question format could be caused by telescoping, rather than by a reduction of forgetting. However, given the size of the difference between the various question formats, we find it unlikely that the difference is only caused by increased forward telescoping. Rather, we postulate that both a large reduction in forgetting, and some increased telescoping could contribute to the fact that this format delivers the best estimate.

The introduction of landmarks in the 12-month retrospective monthly question did not lead to better data accuracy as compared with the timeline-only condition. When predicting the size of absolute inconsistency between 12-month retrospective and 1-month retrospective reports, we found no significant effect of the landmarks on inconsistencies, although the direction of the effect was always in the direction that landmarks lead to *more* inconsistencies. This may be explained by the fact that respondents who received landmarks, found the questionnaire more difficult because of the extra cognitive effort that reporting the landmarks required. Another related reason could be that the landmarks themselves could not be properly dated by respondents (Glasner, 2011). The landmarks that were listed by respondents were likely to have been salient events to respondents, but telescoping in the dates of landmark events could lead respondents to subsequently also misdate FP visits on the timeline. Another possibility is that some respondents who found these questions boring engaged in satisficing behavior, leading to more errors.

One of the limitations of this study is that the 1-month retrospective survey question used a reference period of 4 weeks instead of 1 month. Because our estimate of annual FP use with the 1-month retrospective question was made up of 12 separate estimates, this question only covered 48 weeks of the year instead of the full year. This slight difference in the reference period is something that occurred unintentionally in data collection. Given this, we would have expected higher FP use to be reported in the 12-month retrospective question. We found the opposite. Another limitation concerns the lack of validation data at the respondent level. Despite these limitations, we do find that different question formats on average

lead to widely different estimates, of which the estimate of the most regularly used format—the annual retrospective question—yielded the estimate that was furthest from our population estimate obtained from a register.

This study finds that landmarks did not help to reduce underreports of past behavior. As monthly collection of data is generally not feasible because of increased costs, we have to find different ways to improve the collection of data on behavior. In the future, administrative data may enable us not to rely on survey questions for several frequency measures anymore.

Acknowledgments

Design and data collection were supported by the LISS panel, at CentErdata, Tilburg University, funded by the Dutch Organization for Scientific Research (grant no. 176.010.2005.017). The authors are grateful to Annette Scherpenzeel for her feedback in the design of this study.

References

The American Association for Public Opinion Research (2009). *Standard definitions: Final dispositions of case codes and outcome rates for surveys* (6th ed.). AAPOR. Lenexa: Kansas.

Becker, S. (2003). Does use of the calendar in surveys reduce heaping? *Studies in Family Planning, 34*, 127–132. doi: 10.1111/j.1728-4465.2003.00127.x

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*, 383–406. doi: 10.1080/741942610

Belli, R. F. (2001). Event history calendars and question list surveys: A direct comparison of interviewing methods. *Public Opinion Quarterly, 65*, 45–74.

Börsch-Supan, A., Brandt, M., Hunkler, C., Kneip, C., Korbmacher, J., Malter, F., ... & Zuer, S. (2013). Data resource profile: The survey of health, ageing and retirement in Europe (SHARE). *The International Journal of Epidemiology, 42*, 992–1001. doi: 10.1093/ije/dyt088

Callegaro, M. (2008). Seam effects in longitudinal surveys. *Journal of Official Statistics, 24*, 387–403.

Callegaro, M., & DiSogra, C. (2008). Computing response metrics for online panels. *Public Opinion Quarterly, 72*, 1008–1032. doi:10.1093/poq/nfn065

Conrad, F. G., Brown, N. R., & Cashman, E. R. (1998). Strategies for estimating behavioural frequency in survey interviews. *Memory*, *6*, 339–366.

De Bruin, A., De Bruin, E. I., Gast, A., Kardaun, J. W. P. F., Van Sijl, M., & Verweij, C. G. C. (2003). Koppeling van LMB- en GBA-gegevens: Methode, resultaten en kwaliteitsonderzoek. [Merging LMB and GBA-data: Methods, results and data quality]. Statistics Netherlands. Retrieved from http://www.cbs.nl/NR/rdonlyres/BAF2D6C5-7A77-4771-859B-338CC2F6F589/0/koppelingLMGBAgegevens1203.pdf.

Glasner, T., & Van der Vaart, W. (2009). Applications of calender instruments in social surveys: A review. *Quality and Quantity, 43*, 333–349. doi: 10.1007/s11135-007-9129-8

Glasner, T. J. (2011). *Reconstructing event histories in standardized survey research: Cognitive mechanisms and aided recall techniques* (Ph.D. thesis, Vrije Universiteit: Amsterdam, the Netherlands). Retrieved from http://dare.ubvu.vu.nl/handle/1871/19216

Leenheer, J., & Scherpenzeel, A. C. (2013). Does it pay off to include non-internet households in an internet panel? *International Journal of Internet Science*, *8*, 17–29

Loftus, E. F., & Marburger, W. (1983). Since the eruption of Mt. St. Helens, has anyone beaten you up? Improving the accuracy of retrospective reports with landmark events. *Memory and Cognition*, *11*, 114–120. doi: 10.3758/BF03213465

Lugtig, P., Das, M., & Scherpenzeel, A. C. (2014). Nonresponse and attrition in a probability-based online panel for the general population. In M. Callegaro, P. J. Lavrakas, J. Krosnick, R. P. Baker, J. Bethlehem, & A. S. Göritz (Eds), *Online panel research: A data quality perspective* (pp. 135–153). Wiley series in survey methodology. New York: Wiley.

Pascale, J., Roemer, M. I., & Resnick, D. M. (2009). Medicaid underreporting in the CPS: Results from a record check study. *Public Opinion Quarterly*, *73*, 497–520. doi:10.1093/poq/nfp028

R Core Development Team (2014). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from www.r-project.org

Schwarz, N. (1990). Assessing frequency reports of mundane behaviors: Contributions of cognitive psychology to questionnaire construction. In C. Hendrick & M. S. Clark (Eds.), *Research methods in personality and social psychology: Review of personality and social psychology* (Vol. 11, pp. 98–119). Thousand Oaks, CA: Sage Publications.

Sobell, L. C. (2001). Cross-cultural evaluation of two drinking assessment instruments: Alcohol timeline followback and inventory of drinking situations. *Substance Use and Misuse*, *36*, 313–331.

Van Buuren, S., & Groothuis-Oushoorn, K. (2011). Mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software, 45*, 1–67.

Van Der Vaart, W. (2004). The time-line as a device to enhance recall in standardized research interviews: A split ballot study. *Journal of Official Statistics*, *20*, 301–317.

Van der Vaart, W., & Glasner, T. (2011). Personal landmarks as recall aids in survey interviews. *Field Methods*, *23*, 37–56.

Wagenaar, W. A. (1986). My memory: A study of autobiographical memory over six years. *Cognitive Psychology*, *18*, 225–252. doi: 10.1016/0010-0285(86)900137

Wallace, R. B., & Herzog, A. R. (1995). Overview of the health measures in the health and retirement study. *Journal of Human Resources*, *30*, S84–S107.

Westerberg, V. S. (1998). Reliability of form 90D: An instrument for quantifying drug use. *Substance Abuse, 19*, 179–189. doi: 10.1080/08897079809511386

Biographical Notes

Peter Lugtig is an assistant professor at the Department of Methods and Statistics, Utrecht University and a senior research officer at the Institute for Social and Economic Research at the University of Essex.

Tina Glasner is a research associate at the University of Humanistic Studies, Utrecht, the Netherlands.

Anja J. Boevé is a Ph.D. student at the department of Psychometrics and Statistics at the University of Groningen.