Survey Research Methods (2020)
Vol. 14, No. 2, pp. 109-114
doi:10.18148/srm/2020.v14i2.7741

ISSN 1864-3361
http://www.surveymethods.org

Problems and pitfalls of retrospective survey questions in COVID-19 studies

Lena Hipp
Berlin Social Science Center (WZB), Germany, and
Faculty for Econmoics and Social Sciences
University of Potsdam, Germany

Mareike Bünning
Berlin Social Science Center (WZB)
Germany

Stefan Munnes
Berlin Social Science Center (WZB)
Germany

Armin Sauermann
Berlin Social Science Center (WZB)
Germany

This paper examines and discusses the biases and pitfalls of retrospective survey questions that are currently being used in many medical, epidemiological, and sociological studies on the COVID-19 pandemic. By analyzing the consistency of answers to retrospective questions provided by respondents who participated in the first two waves of a survey on the social consequences of the COVID-19 pandemic, we illustrate the insights generated by a large body of survey research on the use of retrospective questions and recall accuracy.

Keywords: COVID-19; retrospective questions; recall accuracy

1 Introduction

The urgent need to learn about the transmission of the SARS-CoV-2 virus and the pandemic's social, economic, and public health consequences has given rise to an enormous number of surveys across disciplines and countries. Many of these surveys are being conducted outside the context of existing panel studies and therefore lack important information about respondents' pre-crisis situations. Yet, to evaluate respondents' current situations, researchers often need to compare the current situation to a baseline before the outbreak. One way of doing this is to ask retrospective questions (e.g., Giorgio, Riso, Mioni, & Cellini, 2020; Li et al., 2020; Liang et al., 2020; Ran et al., 2020; SORA, 2020). Respondents, however, tend to give less accurate answers when asked about their past than when asked about the present (e.g., Coughlin, 1990; Schnell, 2019; Solga, 2001).

Measurement error arises for several reasons: Retrospective questions place high cognitive demands on respondents (e.g. Durand, Deslauriers, & Valois, 2015; Himmelweit, Biberian, & Stockdale, 1978; Yan & Tourangeau, 2007). Respondents, moreover, have difficulties to remember particular details, especially if the topic in question is not important to them (e.g., Bound, Brown, & Mathiowetz, 2001; Coughlin, 1990; Pina Sánchez, Koskinen, & Plewis, 2014). They also tend to report past attitudes and feelings that are more consistent with their current situation (Barsky, 2002; Jaspers, Lubbers, & Graaf, 2009; Schmier & Halpern, 2004; Yarrow, Campbell, & Burton, 1970) and current societal norms and values (Coughlin, 1990; Himmelweit et al., 1978).

To illustrate some major pitfalls and biases associated with using retrospective survey questions, we analyze data from a nonprobability online panel survey on the social consequences of the lockdown following the coronavirus outbreak in Germany. In particular, we investigate the consistency of respondents' answers to several types of retrospective questions asked at two different points in time. Based on the analyses of these Covid-19-specific data and the general insights generated in survey methodology, we conclude our research note with recommendations for research projects that necessarily rely on retrospective questions.

Data & Measures

The data used in our analyses stem from a nonprobability online survey on individuals' everyday experiences during the Covid-19 lockdown in Germany. The survey contained a number of retrospective questions, as has been the case for many other Covid-19 studies in medicine and social science (Betsch et al., 2020; Giorgio et al., 2020; Li et al., 2020; Liang et al., 2020; Ran et al., 2020; SORA, 2020).[1]

Contact information: Lena Hipp, Reichpietschufer 50, D-10785 Berlin (E-mail: lena.hipp@wzb.eu)

[1] Data collection started on March 23, 2020. Participants learned about the study via email lists, newspaper announcements, and instant message services. Participants who agreed to be interviewed again were sent follow-up questionnaires 3.5 to 4 weeks after they filled out the first questionnaire. The study's codebook and all repli-

Due to a change in the questionnaire, a considerable number of participants ($n = 1,486$) were asked five retrospective questions in both Wave 1 and Wave 2.[2] Two of these questions referred to respondents' working schedules, i.e., whether and how often they worked in the evening (7–10 pm) or at night (10 pm–6 am). One question asked about respondents' self-rated physical health and two questions asked about their mental health (Giesinger, Rumpold, & Schüßler, 2008; Kessler et al., 2002). All items were measured on a five-point scale. We used these items to examine the consistency of respondents' answers to retrospective questions.

2 Variation in measurement error due to subjectivity and complexity of retrospective survey questions

These questions allowed us to compare recall consistency between questions that sought to elicit objective information (on respondents' schedules) and subjective information (on their physical and mental health) between time 1 (T1) and time 2 (T2). Objective information should be easier to recall than subjective information (Schmier & Halpern, 2004). Moreover, we compared the consistency of the answers to three retrospective questions that differed in their degree of complexity. The question on physical health included only one answer dimension, whereas the questions on mental health included three answer dimensions each (Giesinger et al., 2008; Kessler et al., 2002)[3] Recalling past events and experiences is cognitively demanding and these demands further increase if the wording of questions is overly complicated or the questions include several aspects (e.g., Yan & Tourangeau, 2007).

To assess how these two question characteristics affected the consistency of answering behaviors in our sample, we first compared aggregate sample means at T1 and T2. Figure 1 shows that the means of the retrospective assessments aligned closely across the two waves for all five items. Although these data do not tell us how accurately these retrospective assessments reflected respondents' actual situations before the pandemic, they show fairly consistent response patterns on the aggregate level. The only exception was when respondents were asked how often they felt "distressed, hopeless, and strained" (mental health 2 measure)(mean difference of around 0.08 out of 5 scaling points; 95%CI = $[0.03, 0.12]$). The fact that this difference is quite small in size aligns with findings from previous studies (e.g., Beckett, Vanzo, Sastry, Panis, & Peterson, 2001; Jaspers et al., 2009)

Next, we will turn to the kappa-statistic measure of agreement between respondents' answers to questions on the pre-pandemic time at T1 and T2. Using Cohen's kappa with linear weights, we found substantial agreement in answering questions on working evenings and nights ($\kappa = 0.68$ and 0.63), but among our three health measures, agreement was only fair to moderate ($\kappa = 0.35$ and 0.36 for the two mental health items and $kappa = 0.41$ for physical health) (Landis & Koch, 1977). In line with previous studies (Schmier & Halpern, 2004), we hence observed greater consistency in answers to questions on more objective information (i.e., working time schedules) than on more subjective information (i.e., physical and mental health). In contrast to previous research (Yan & Tourangeau, 2007, p. 62), however, we only found small differences between the simpler physical health item and the more complex mental health items.

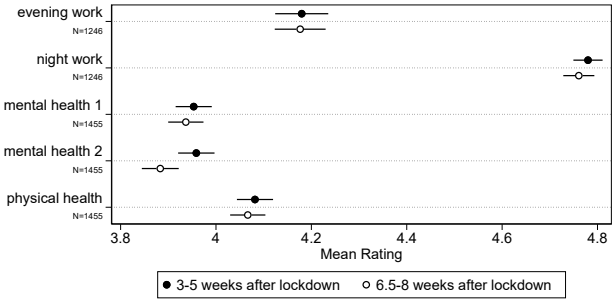

Figure 1. Average rating of pre-pandemic situation 3–5 weeks and 6.5–8 weeks after the start of the lockdowns. Answer categories for questions on evening/night work were *1 – every day, 2 – few times a week, 3 – every couple of weeks, 4 – less often, 5 – never*, for mental health *1 – all the time, 2 – most of the time, 3 – sometimes, 4 – rarely, 5 – never*, and for physical health *1 – bad, 2 – poor, 3 – satisfactory, 4 – good, 5 – very good*; only respondents aged 18-65 were included for the analyses of evening/night work. To assess whether the mean ratings of the pre-pandemic situation differed between T1 and T2 we ran paired t-tests (two-tailed).

sectionVariation in measurement error due to respondent characteristics and time between the interview and the event of interest

Our data also allowed us to examine the degree to which measurement error in retrospective questions varies with individual-level characteristics, in particular respondents'

cation materials are available at https://osf.io/qf3js/

[2]Respondents who were included in the analysis first participated in the survey between April 6–19, 2020; the total number of valid observations in Wave 1 was 14,888; 9,963 respondents provided valid contact information to be included in Wave 2, of whom 3,516 were asked a couple of retrospective questions in both waves; at the time we conducted our analyses for this paper, 1,486 of these respondents had participated in Wave 2.

[3]Physical health question: "How would you describe your physical health before the measures related to the coronavirus pandemic were first introduced?" Mental health question: "How often have you experienced the following feelings since the measures related to the Coronavirus pandemic were first introduced?"; a) "anxious, nervous, restless", b) "depressed, hopeless, strained."

current situations (e.g. Schnell, 2019), and the time span between the interview and the experience/event of interest (e.g., Beckett et al., 2001; Bound et al., 2001; Jaspers et al., 2009; Schmier and Halpern, 2004; Warrington and Silberstein, 1970 found that recall accuracy decreased over time). Health-related research has found that respondents "tend to over-report symptoms that correspond to current illness, and under report symptoms that predate the illness" (Schmier & Halpern, 2004). In our analyses, we therefore captured changes in the status quo by including both the "improvement" and "deterioration" of the respondents' status quo between T1 and T2 (i.e., less/more evening and night work and better/worse mental and physical health). We also included the number of days between the first and the second interview as a predictor in our analyses. Last but not least, although most socio-demographic characteristics have been found to be unrelated to recall accuracy (Coughlin, 1990), some studies found recall accuracy to be higher among more highly educated respondents (Joslyn, 2003) and lower among older respondents compared to younger ones (Schmier & Halpern, 2004; Warrington & Silberstein, 1970; Yarrow et al., 1970). We therefore included age (as a categorical variable) and education (dummy variable for tertiary degree) in our multivariate analyses as well as gender (dummy variable), parental status (dummy variable), geographic region (categorical variable), and size of the community (dummy variable). These adjustments were also important because our data were not drawn from a probability sample.

The outcome variable in our multinomial regressions for each item was a three-level categorical variable that distinguishes between consistent answering behaviors, a more negative assessment of the past at T2 than at T1, and a more positive assessment of the past at T2 than at T1. This trichotomization was necessary due to the unequal spacing of the answer categories for the different variables and the existence of cells with small numbers.

For each variable of interest, Figure 2 displays the adjusted predictions for the consistency of respondents' answers to retrospective questions depending on changes in the status quo. Column 1 displays the predicted probabilities for assessing the pre-pandemic situation as more negative at T2 than at T1, Column 2 shows the probabilities for giving the same answers to the retrospective questions at T1 and T2, and Column 3 reveals the probability of reporting better pre-pandemic outcomes at T2 than T1. A table with the predicted probabilities for the control variables can be found in the Online Appendix.

For each variable, the first row displays the proportion of respondents who experienced a deterioration in their current situation between T1 and T2 (i.e., more frequent evening and night work and poorer mental and physical health); the second row displays the predictions for those respondents who did not report a change, and the third row shows the predictions for those who reported an improvement in the variables of interest between T1 and T2.

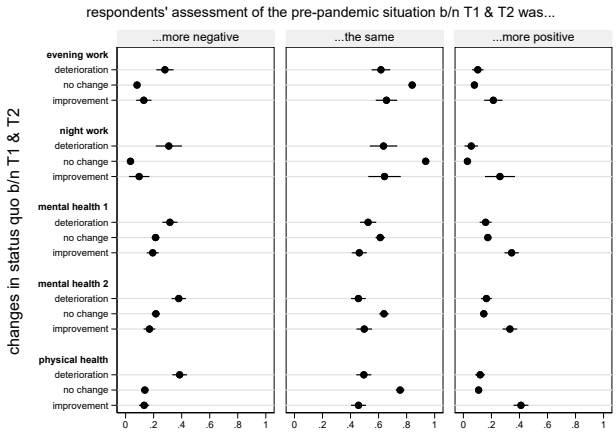

Figure 2. Consistency in answers to retrospective questions on working times, mental health, and physical health between T1 and T2. Predicted margins from multinomial logistic regression models adjusted for gender, age, tertiary education, parenthood, size of residence, region, and time span between waves; N(evening work) = 1133, N(night work) = 1129 (only respondents aged 18-65 were included), N(mental health 1) = 1427, N(mental health 2) = 1416, N(physical health) = 1429.

For all five items, we found a consistent pattern. Respondents who did not experience a change in their current situation between T1 and T2 were more likely to give consistent reports of their pre-pandemic situation than those who experienced an improvement or deterioration between waves. For instance, more than 80% of respondents who did not experience a change in evening work between T1 and T2 reported consistent pre-pandemic frequencies of evening work compared to only about 60% of respondents who experienced an increase or decrease in evening work between waves. Furthermore, across all items, we found that respondents who reported that their situation improved (deteriorated) in between waves were also more likely to report improved (deteriorated) conditions for the pre-pandemic situation in T2 compared to T1. This indicates that using retrospective questions leads to an underestimation of change between the current and pre-pandemic situation.

The number of days between T1 and T2 was not related to recall consistency for any of our five items (please see replication files for more information). With one exception, this also applied to the socio-demographic characteristics: Only respondents with a tertiary degree were more likely to provide consistent estimates of feeling depressed/hopeless/strained (mental health 2) prior to the pandemic.

Recommendations

What recommendations can researchers draw from these and related analyses when they have to rely on data collected using retrospective questions, as in the case of the many ongoing Covid-19 studies? First, even though data elicited using retrospective questions tend to be unreliable at the individual level, they are pretty consistent at the aggregate level (see also Jaspers et al., 2009). However, as respondents' assessments of their pre-pandemic conditions were positively correlated with changes in their present conditions, the aggregate amount of change between the past and present is likely underestimated and aggregate estimates may still be biased if changing societal norms (e.g., washing hands etc.) yield collective shifts in answering behaviors (see Jaspers et al., 2009; Joslyn, 2003).

Second, researchers should minimize the cognitive effort associated with (retrospective) questions (Krosnick, 1991; Stull, Leidy, Parasuraman, & Chassany, 2009). Questions should be short, easy to understand, and should not include multiple items. Introductory texts for questions can also help respondents by providing them with extra time to search their memory (Schnell, 2019). Questions that require respondents to provide objective facts prompt a higher recall accuracy than subjective evaluations (see also e.g. Schmier & Halpern, 2004). Asking broader questions and offering broader answering categories may also help to increase recall accuracy (Beckett et al., 2001; Schnell, 2019). If researchers need to collect detailed or sensitive information, they are well advised to do so by asking summary questions first and more specific questions later (Beckett et al., 2001; Schnell, 2019).

Third, the time span between the interview and the event or the experience that has to be recalled is another critical issue for recall accuracy. Using specific anchor points can help respondents to remember their health status and other items of interest (Barsky, 2002; Pearson, Ross, & Dawes, 1992). For example, if researchers ask questions about health status before the outbreak of the COVID-19 pandemic, as we did in our study, these will likely prompt more accurate answers than questions about their health status two months ago.

References

Barsky, A. J. (2002). Forgetting, fabricating, and telescoping. *Archives of Internal Medicine*, *162*(9), 981–984. doi:10.1001/archinte.162.9.981

Beckett, M., Vanzo, J. D., Sastry, N., Panis, C., & Peterson, C. (2001). The quality of retrospective data: An examination of long-term recall in a developing country. *The Journal of Human Resources*, *36*(3), 593–625. doi:10.2307/3069631

Betsch, C., Korn, L., Felgendreff, L., Eitze, S., Schmid, P., Sprengholz, P., … Schlosser, F. (2020). *German covid-19 snapshot monitoring (cosmo) - welle 9 (28.04.2020)*. PsychArchives. doi:http://dx.doi.org/10.23668/psycharchives.2890

Bound, J., Brown, C., & Mathiowetz, N. (2001). Measurement error in survey data. In *Handbook of econometrics* (pp. 3705–3843). doi:10.1016/s1573-4412(01)05012-7

Coughlin, S. S. (1990). Recall bias in epidemiologic studies. *Journal of Clinical Epidemiology*, *43*(1), 87–91. doi:10.1016/0895-4356(90)90060-3

Durand, C., Deslauriers, M., & Valois, I. (2015). Should recall of previous votes be used to adjust estimates of votingintention? *Survey Insights: Methods from the Field, Weighting: Practical Issues and'How to' Approach.* Retrieved from surveyinsights.org/?p=3543

Giesinger, J., Rumpold, M. G., & Schüßler, G. (2008). Die k10-screening-skala für unspezifischen psychischen distress. *Psychosomatik und Konsiliarpsychiatrie*, *2*(2), 104–111. doi:10.1007/s11800-008-0100-x

Giorgio, E. D., Riso, D. D., Mioni, G., & Cellini, N. (2020). The interplay between mothers' and children behavioral and psychological factors during COVID-19: An italian study. doi:10.31234/osf.io/dqk7h

Himmelweit, H. T., Biberian, M. J., & Stockdale, J. (1978). Memory for past vote: Implications of a study of bias in recall. *British Journal of Political Science*, *8*(3), 365–375.

Jaspers, E., Lubbers, M., & Graaf, N. D. D. (2009). Measuring once twice: An evaluation of recalling attitudes in survey research. *European Sociological Review*, *25*(3), 287–301. doi:10.1093/esr/jcn048

Joslyn, M. R. (2003). The determinants and consequences of recall error about gulf war preferences. *Political Science*, *47*(3), 440–452. doi:10.1111/1540-5907.00032

Kessler, R. C., Andrews, G., Colpe, L. J., Hiripi, E., Mroczek, D. K., Normand, S.-L. T., … Zaslavsky, A. M. (2002). Short screening scales to monitor population prevalences and trends in non-specific psychological distress. *Psychological Medicine*, *32*(6), 959–976. doi:10.1017/s0033291702006074

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, *5*(3), 213–236. doi:10.1002/acp.2350050305

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174. Retrieved from http://www.jstor.org/stable/2529310

Li, Y. K., Peng, S., Li, L.-q., Wang, Q., Ping, W., Zhang, N., & Fu, X.-n. (2020). Clinical and transmission characteristics of COVID-19 - a retrospective study of 25 cases from a single thoracic surgery department. *Current Medical Science*, *40*(2), 295–300. doi:10.1007/s11596-020-2176-2

Liang, L., Ren, H., Cao, R., Hu, Y., Qin, Z., Li, C., & Mei, S. (2020). The effect of COVID-19 on youth mental health. *Psychiatric Quarterly*. doi:10.1007/s11126-020-09744-3

McCormick, T. H., Salganik, M. J., & Zheng, T. (2010). How many people do you know? Efficiently estimating personal network size. *Journal of the American Statistical Association*, *105*(489), 59–70. doi:10.1198/jasa.2009.ap08518

Pearson, R. W., Ross, M., & Dawes, R. M. (1992). Personal recall and the limits of retrospective questions in surveys. In J. M. Tanur (Ed.), *Questions about questions. inquiries into the cognitive bases of surveys.* Russel Sage Foundation. Retrieved from https://www.jstor.org/stable/10.7758/9781610445269.10

Pina Sánchez, J., Koskinen, J., & Plewis, I. (2014). Measurement error in retrospective work histories. *Survey Research Methods*, *Vol 8*(1), 43–55. doi:10.18148/SRM/2014.V8I1.5144

Ran, L., Chen, X., Wang, Y., Wu, W., Zhang, L., & Tan, X. (2020). Risk factors of healthcare workers with corona virus disease 2019: A retrospective cohort study in a designated hospital of Wuhan in China. *Clinical Infectious Diseases*. doi:10.1093/cid/ciaa287

Raphael, K. (1987). Recall bias: A proposal for assessment and control. *International Journal of Epidemiology*, *16*(2), 167–170. doi:10.1093/ije/16.2.167

Schmier, J. K., & Halpern, M. T. (2004). Patient recall and recall bias of health state and health status. *Expert Review of Pharmacoeconomics & Outcomes Research*, *4*(2), 159–163. doi:10.1586/14737167.4.2.159

Schnell, R. (2019). *Survey-Interviews. methoden standardisierter Befragungen* (2nd ed.). doi:10.1007/978-3-531-19901-6

Solga, H. (2001). Longitudinal surveys and the study of occupational mobility: Panel and retrospective design in comparison. *Quality and Quantity*, *35*(3), 291–309. doi:10.1023/a:1010387414959

SORA. (2020). *Sars-cov-2 in austria. pcr tests in a representative sample. study report.* SORA Institute for Social Research and Consulting. Vienna. Retrieved from https://www.sora.at/fileadmin/downloads/projekte/Austria_Spread_of_SARS-CoV-2_Study_Report.pdf

Stull, D. E., Leidy, N. K., Parasuraman, B., & Chassany, O. (2009). Optimal recall periods for patient-reported outcomes: Challenges and potential solutions. *Current Medical Research and Opinion*, *25*(4), 929–942. doi:10.1185/03007990902774765

Warrington, E. K., & Silberstein, M. (1970). A questionnaire technique for investigating very long term memory. *Quarterly Journal of Experimental Psychology*, *22*(3), 508–512. doi:10.1080/14640747008401927

Yan, T., & Tourangeau, R. (2007). Fast times and easy questions: The effects of age, experience and question complexity on web survey response times. *Applied Cognitive Psychology*, *22*(1), 51–68. doi:10.1002/acp.1331

Yarrow, M. R., Campbell, J. D., & Burton, R. V. (1970). Recollections of childhood a study of the retrospective method. *Monographs of the Society for Research in Child Development*, *35*(5), iii. doi:10.2307/1165649

Commentary

The article shows the agreement of retrospective questions (before COVID-19) between two different times, relating them to the same questions with current reference in each of those two times. Therefore, there is greater agreement between the retrospective questions when there have been no changes between the current answers at all times. Does this mean that the retrospective response on the second measurement is more reliable when there have been no changes in the current responses between the two measurements?

In addition, a similar result is also observed when there is no agreement, that is, when the retrospective questions differ from one measurement to another. In this way, the deterioration (or improvement) between the current responses of each time also obtains the highest coefficients for the more negative (or more positive) retrospective responses from one moment to the next. Based on this, do the bias in the retrospective response given in the second measurement and the change in the current response between the times (deterioration, no change or improvement, respectively) have the same sign?

This may be of special interest to better understand the social, economic and health impact that COVID-19 is having on the population through studies with retrospective questions.

On the other hand, just as the article gives suggestions in the formulation of retrospective questions to reduce memory bias, what statistical methods would help to obtain better results in this type of retrospective question?

Maria del Mar Rueda
University of Granada, Spain

Andrés Cabrera
Andalusian School of Public Health
Granada, Spain

Reply to del Mar Rueda and Cabrera

In accordance with the existing literature (Barsky, 2002; Jaspers et al., 2009; Schmier & Halpern, 2004; Yarrow et al., 1970)[4], we found more consistent answering behaviors to retrospective questions when the item of interest had not changed between the two time points at which our respondents were interviewed compared to when there was a change. For all the items in our study, we observe that respondents were more likely to report more negative pre-pandemic conditions (T0) in the later interview (T2) if they experienced a deterioration between the first interview (T1) and the second interview (T2) than if there was no change in conditions between T1 and T2. Likewise, respondents were more likely to report more positive pre-pandemic conditions at T2 if they experienced an improvement between T1 and T2. Hence, respondents seem to remember their past condition as having been more similar to their present condition than they actually said it had been at T1. This finding suggests that data relying on retrospective survey questions may underestimate the amount of change that occurred between past and present (also see paragraph one in our recommendations section), i.e., retrospective questions result in conservative estimates of the amount of change between T0 (unobserved) and T1 (observed). Unfortunately, there are few options for adjusting and correcting such recall biases in those instances. Previous research has proposed techniques such as calibration curves, which have been used to address recall error in social network research (McCormick, Salganik, & Zheng, 2010), or validity scales, which have been applied in epidemiological research (Raphael, 1987). Yet, these methods require some baseline information for at least a proportion of the sample, general information about the population of interest, or information about a comparison group.

Lena Hipp, Mareike Bünning, Stefan Munnes, and Armin Sauermann

[4]The references for the citations are given in the references of the main paper.