Survey Research Methods (2020)
Vol. 14, No. 2, pp. 205-209
doi:10.18148/srm/2020.v14i2.7759

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

Can we directly survey adherence to non-pharmaceutical interventions? Evidence from a list experiment conducted in Germany during the early Corona pandemic.

Simon Munzert
Hertie School
Berlin, Germany

Peter Selb
Department of Politics and Public Administration
University of Konstanz, Germany

Self-reports of adherence to non-pharmaceutical interventions in surveys may be subject to social desirability bias. Existing questioning techniques to reduce bias are rarely used to monitor adherence. We conducted a list experiment to elicit truthful answers to the question whether respondents met friends or acquaintances and thus disregarded the social distancing norm. Our empirical findings are mixed. Using the list experiment, we estimate the prevalence of non-compliant behavior at 28%, whereas the estimate from a direct question is 22%. However, a more permissively phrased direct question included later in the survey yields an estimate of 47%. All three estimates vary consistently across social groups. Interestingly, only the list experiment reveals somewhat higher non-compliance rates among the highly educated compared to those with lower education, yet the variance of the list estimates is considerably higher. We conclude that the list experiment compared unfavorably to simpler direct measurements in our case.

Keywords: COVID-19; compliance; item count technique; list experiment; non-pharmaceutical interventions; social desirability bias; social distancing

1 Motivation

Many countries, including Germany, have adopted non-pharmaceutical public health measures to contain the SARS-CoV-2 pandemic. Some of these measures both demand costly behavioral self-restrictions from individuals and are difficult for the authorities to enforce. Surveys such as the "WHO tool for behavioural insights on COVID-19" are used to monitor adherence to these measures (Betsch, Wieler, & Habersaat, 2020), but they rarely account that socially desirable – let alone officially required – behaviors are subject to overreporting in surveys (e.g., Roger Tourangeau & Yan, 2007).

We embedded a list experiment, a prominent technique to reduce social desirability bias (e.g., Glynn, 2013), in an online survey of a non-probability sample of Germany's residential population in late April–early May 2020. We were particularly interested in a directive enacted in all German states on March 22 to minimize physically meeting others in both the public and private spheres. Individual commitment to social distancing measures is considered key in containing the pandemic (e.g., Anderson, Heesterbeek, Klinkenberg, & Hollingsworth, 2020).

In a typical list experiment, respondents are asked to report how many (N.B. not: which) statements on a list apply to them. The treatment group receives an additional sensitive item on the list (in our case: whether they had met with friends or acquaintances within the last week). The prevalence of the sensitive item can then be estimated by comparing the means of the item counts in the experimental versus the control group. By way of comparison, we augment the typical setup with another control group in which all of the items (including the sensitive one) are polled directly.

We then use multivariate regression techniques from Imai (2011) to estimate how the probability that the sensitive item varies as a function of the respondents' social characteristics (age, gender, education). Group differences in adherence to NPIs is a recurring topic in epidemiological research (see Aiello et al., 2010). Most recently, lower adherence to social distancing practices has been suspected to have caused temporal rises in COVID-19 cases among younger cohorts in Germany (Goldstein & Lipsitch, 2020).

Our initial plan to analyze group-specific patterns of misreporting using a method developed by Eady (2017) with a direct question asked at the end of the survey failed because this item's prevalence far exceeded the estimate from the list experiment (see the preregistration linked in the Acknowledgement at the end of this article). We discuss potential sources of error and implications for survey-based studies of

Contact information: Peter Selb, Department of Politics and Public Administration, University of Konstanz, Box 85, D-78457 Konstanz (E-mail: peter.selb@uni-konstanz.de)

NPI compliance in the conclusion.

2 Experimental design

The experiment was embedded in an ad hoc survey initiated by the Cluster of Excellence "The Politics of Inequality" at the University of Konstanz, in which 5,015 participants were recruited from a commercial online access panel administered and remunerated by respondi. Participants were required to be 18 years of age or older, German-speaking, and residents of Germany. The quota reflected the resident population in terms of (the marginal distributions of) age group, gender, education, and region (see Table A1) in the Supplementary Information, SI). The survey was implemented and run by surveyLab at the University of Konstanz from April 29–May 8. Speeder respondents were excluded prior to the analysis to improve data quality in accordance with our pre-analysis plan (see the Acknowledgement). After excluding speeders and applying list-wise deletion of observations with missing variable values, the size of the sample used for the analysis was $n = 4,448$.

Respondents were assigned with equal probability to one of three conditions: a list of 4 items (control group I; $n = 1,396$); a list of 4+1 items (treatment group; $n = 1,493$); a direct question with 4+1 items (control group II; $n = 1,559$). Table A2 in the SI displays the wording of the items on the list. The introductory text clearly placed the items in the context of COVID-19 NPIs, which possibly acted to decrease prevalence estimates. The control items pertained to actions and behaviors which vary in their social desirability and presumed prevalence in order to avoid ceiling/floor effects and to reduce the measurement variance of the item count via negative correlations (see Glynn, 2013). The treatment group and control group II had the additional item of interest (*"I have met with friends or acquaintances"*) on the list. The order of the items was randomly varied to level order effects. All respondents were asked another direct question, separated by 15 buffer pages that contained questions which were unrelated to the experiment and NPIs, about the sensitive behavior with a 4-point response scale: *"How often have you met with friends or acquaintances in the last 7 days?"* The response categories were (1) daily, (2) several times, (3) once, (4) never. The item was adapted from the Mannheim Corona Study, a rolling panel survey based on the German Internet Panel (Blom et al., 2020). Both the lack of an NPI frame as well as the provision and order of nuanced response categories distinguish this item from the direct question embedded in the experiment and possibly made the sensitive behavior appear more acceptable to the respondents (e.g., Schwarz, Hippler, Deutsch, & Strack, 1985). In the figures and tables below, we denote the items 'Direct question (multivalue)' and 'Direct question (binary)', although we dichotomize the former (daily, several times, and once versus never) for better comparability.

Table 1
Estimated prevalence of sensitive behavior "I have met with friends or acquaintances"

	Est.	Std. Err.
List	0.28	0.04
Direct (binary)	0.22	0.01
Direct (multivalue)	0.47	0.01
Difference list − direct (binary)	0.06	0.04
Difference list − direct (multivalue)	−0.19	0.04

3 Empirical results

Table 1 reports the estimated prevalence of the sensitive behavior according to the different measurements (see Figure A1 in the SI for an overview of all rates including the estimated prevalence for the control items). As initially stated, the difference in means of the item counts in the experimental treatment group (2.29) versus the control group (2.01) identifies the prevalence in the list experiment. An inspection of the item counts by treatment group indicates that less than 10% of each group chose either the minimum or the maximum number of items and thereby offset the privacy-preserving property of the experiment (see Table A3 in the SI).

The prevalence estimated from the direct (binary) question embedded in the list experiment is somewhat lower at 22%. However, this difference is barely significant at conventional levels due to the vast measurement variance of the list estimate. On the other hand and contrary to our expectations, the prevalence estimated from direct (multivalued) question is a whopping 19 percentage points higher than that based on the item counts.

To analyze whether the prevalence levels estimated via different instruments vary by respondent characteristics, we turn to a multivariate regression estimator from Imai (2011). Figure 1 displays the effects of gender (baseline: males), age (baseline: 18-29 year olds), and education (baseline: low level of education) on the probability of the sensitive item by measurement instrument. We find that the three estimates vary more or less consistently across social groups, with the effects based on the list experiment being estimated with relatively low precision. Most notably, older (and thus more vulnerable) age groups seem more compliant than younger (and socially more connected) cohorts – a result in line with suspicions raised in Goldstein and Lipsitch (2020). Interestingly, only the list experiment reveals somewhat higher non-compliance among highly educated compared to those with a low level of education. This seems to indirectly support the finding from a related literature (i.e., vote validation) that the overreporting of socially desirable behaviors in surveys is bound to social status, presumably since high-status individuals are more susceptible to social pressure (e.g., An-

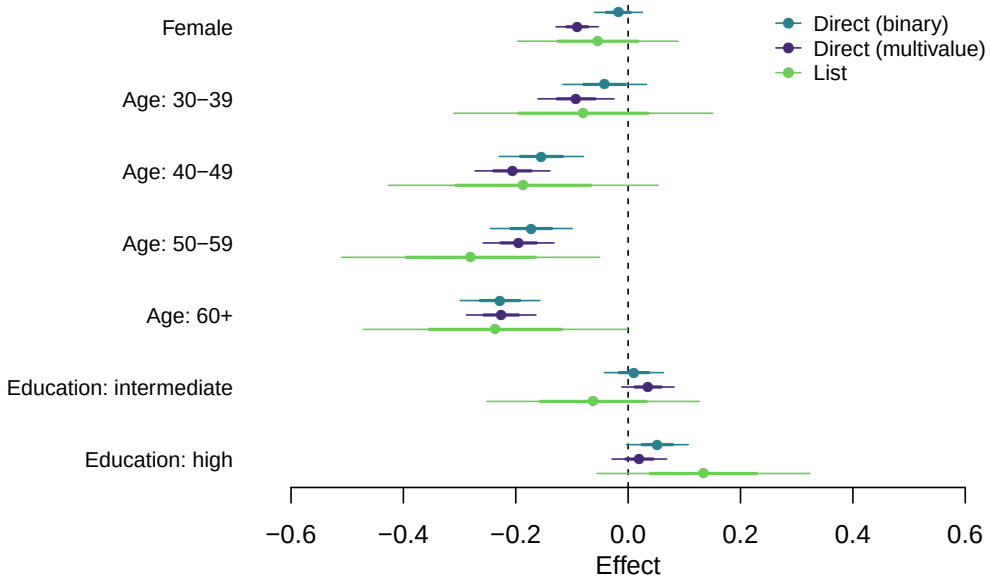

Figure 1. Estimated difference in the probability of sensitive item (meeting with friends or acquaintances) relative to a baseline group by measurement instrument. Point estimates along with 83% and 95% confidence intervals reported.

solabehere & Hersh, 2012). Otherwise, the differences between social groups and methods of measurement seem negligible.

4 Discussion

To sum up, our empirical comparison of direct and indirect survey measures of adherence to NPIs does not yield any conclusive evidence in support of the list experiment. While the prevalence estimate based on item counts is somewhat higher (and thus closer to the unobserved ground truth, so the presumption) than an estimate from a simple binary direct question embedded in the experiment, a more permissively phrased multivalued direct question included later in the survey suggests much higher prevalence rates. Moreover, the measurement variance of the item count measure is a multiple of the variances of both direct measurements, and the method does not provide an immediate measure of the sensitive item for each respondent (see Glynn, 2013).

To be sure, the devil may well be in the details of our implementation. For instance, the explicit framing of the list experiment and the choice of control items in terms of NPI compliance may have inhibited rather than encouraged truthful answers even in the treatment condition. Additionally, despite our effort to get rid of speeders, some satisficing respondents may have simply picked the first response category on offer ('daily') in the drop-down menu of the multivalued direct question. While this would have artificially inflated the prevalence estimate, Figure A3 in the SI suggests that prior exposure to the sensitive item in the experimental condition and control group II reduced affirmative responses to the multivalued direct question. On the other hand, the self-administered online mode may have granted the respondents enough privacy to truthfully answer even direct questions (e.g., Roger Tourangeau & Yan, 2007), so that the list experiment has been superfluous in this setting. Finally, perhaps the respondents did not consider the presumably sensitive item of interest (meeting friends and acquaintances) so sensitive after all.

For the time being, however, our findings corroborate the results from a recent Danish study (Larsen, Petersen, & Nyrup, 2020), and caution against the use of list experiments in survey-based measurements of NPI adherence. Another recent survey experiment conducted in Canada apparently achieved good results with simple face-saving item formulations (Daoust et al., 2020). Given the complications the list experiment carries with it for both the researcher and the respondent, this might be a valuable alternative.

Acknowledgement

Funded by the Deutsche Forschungsgemeinschaft (DFG - German Research Foundation) under the Excellence Strategy of the German federal and state governments - EXC-2035/1 - 390681379. OSF preregistration available at https://osf.io/s4f85. We are grateful to Thomas Wöhler and Konstantin Mozer from surveyLab (https://www.soziologie.uni-konstanz.de/hinz/surveylab/) for programming the experiment and organizing the field work.

Both authors contributed equally to this study.

References

Aiello, A. E., Coulborn, R. M., Aragon, T. J., Baker, M. G., Burrus, B. B., Cowling, B. J., … Ferng, Y.-h., et al. (2010). Research findings from nonpharmaceutical intervention studies for pandemic influenza and current gaps in the research. *American Journal of Infection Control*, *38*(4), 251–258.

Anderson, R. M., Heesterbeek, H., Klinkenberg, D., & Hollingsworth, T. D. (2020). How will country-based mitigation measures influence the course of the covid-19 epidemic? *The Lancet*, *395*(10228), 931–934.

Ansolabehere, S., & Hersh, E. (2012). Validation: What big data reveal about survey misreporting and the real electorate. *Political Analysis*, *20*(4), 437–459.

Betsch, C., Wieler, L. H., & Habersaat, K. (2020). Monitoring behavioural insights related to covid-19. *The Lancet*, *395*(10232), 1255–1256.

Blom, A., Möhring, K., Naumann, E., Reifenscheid, M., Lehrer, R., Juhl, S., … Axenfeld, J. (2020). *Mannheimer corona-studie*. Retrieved from https://www.uni-mannheim.de/gip/corona-studie/ [Accessed 11 May 2020].

Daoust, J.-F., Nadeau, R., Dassonneville, R., Lachapelle, E., Bélanger, É., Savoie, J., & van der Linden, C. (2020). How to survey citizens' compliance with covid-19 public health measures? evidence from three survey experiments.

Eady, G. (2017). The statistical analysis of misreporting on sensitive survey questions. *Political Analysis*, *25*(2), 241–259.

Glynn, A. N. (2013). What can we learn with statistical truth serum? design and analysis of the list experiment. *Public Opinion Quarterly*, *77*(S1), 159–172.

Goldstein, E., & Lipsitch, M. (2020). Temporal rise in the proportion of younger adults and older adolescents among coronavirus disease (covid-19) cases following the introduction of physical distancing measures, germany, march to april 2020. *Eurosurveillance*, *25*(17), 2000596.

Holbrook, A. L., & Krosnick, J. A. (2010). Measuring voter turnout by using the randomized response technique: Evidence calling into question the method's validity. *Public Opinion Quarterly*, *74*(2), 328–343. doi:10.1093/poq/nfq012. eprint: https://academic.oup.com/poq/article-pdf/74/2/328/5457918/nfq012.pdf

Imai, K. (2011). Multivariate regression analysis for the item count technique. *Journal of the American Statistical Association*, *106*(494), 407–416.

Larsen, M. V., Petersen, M. B., & Nyrup, J. (2020). Do survey estimates of the public's compliance with covid-19 regulations suffer from social desirability bias?

Schwarz, N., Hippler, H.-J., Deutsch, B., & Strack, F. (1985). Response scales: Effects of category range on reported behavior and comparative judgments. *Public Opinion Quarterly*, *49*(3), 388–395.

Tourangeau, R. [R.], & Yan, T. [T.]. (2007). Sensitive questions in surveys. *Psychological Bulletin*, *133*(5), 859–883. doi:10.1037/0033-2909.133.5.859

Tourangeau, R. [Roger], & Yan, T. [Ting]. (2007). Sensitive questions in surveys. *Psychological Bulletin*, *133*(5), 859.

Commentary

Measuring behaviors that are positively or negatively sanctioned poses a problem that is most often thought of in terms of social desirability bias. One approach designed to address social desirability bias is the list or item count method (R. Tourangeau & Yan, 2007)[1]. Munzert and Selb found that using this method in a web survey conducted during the Coronavirus pandemic did not affect admissions of a socially undesirable behavior. This is consistent with Holbrook and Krosnick's (2010) finding that the approach had an effect only in an interviewer-administered survey not in a comparable self-administered web survey.

By contrast, Munzert and Selb found that changing item wording—another approach to addressing social desirability—was related to reports of the stigmatized behavior. They found that asking a frequency question (How often have you met with friends . . . ?), as opposed to a dichotomous yes/no question, was associated with a large increase in reports of the proscribed behavior. However, the authors' experimental design confounded wording and context, thereby making the interpretation of this result uncertain. The dichotomous question was asked in the context of social distancing interventions to reduce the spread of the Corona virus and therefore "met with friends" clearly conveyed in-person meetings. But the analogous frequency question was asked in a non-pandemic context much later in the questionnaire and thus "met with friends" may have been interpreted as including Zoom, Face-Time or other nonface-to-face meetings.

Munzert and Selb say they used the same wording as the Mannheim Corona study. It would be useful for the readers to see the frequencies of the respondi study compared to the frequencies of the Mannheim study (for comparable weeks).[2] Using the Mannheim study data over time could also help to get a sense of changes in this behavior. If for example the baseline for the elderly is fewer friends visits, or no friends visits, then of course there is no need to misreport.

Munzert and Selb note that Daoust et al. (2020) found that the combination of a face-saving introduction with adding

[1] References are listed among the references for the main article.

[2] For details of the Mannheim Corons study see Blom et al. in this issue. Also see https://www.uni-mannheim.de/media/Einrichtungen/gip/Corona_Studie/29-05-2020_Result_Tables_for_the_Daily_Report.pdf.

"Only when necessary/occasionally" to Yes/No options increased reporting of stigmatized behaviors. To our knowledge, this approach has never been used before, possibly because it violates an elementary rule of question-wording: response options should be mutually exclusive. However, if replicated by others, this might constitute an important exception to an otherwise sensible rule.

Frauke Kreuter
University of Maryland, University of Mannheim, and IAB
Stanley Presser
University of Maryland