Survey Research Methods (2020)
Vol. 14, No. 2, pp. 115-121
doi:10.18148/srm/2020.v14i2.7755

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

Unit nonresponse biases in estimates of SARS-CoV-2 prevalence

Julia C. Post
Faculty for Economic and Social Research
University of Potsdam, Germany

Fabian Class
Faculty for Economic and Social Research
University of Potsdam, Germany

Ulrich Kohler
Faculty for Economic and Social Research
University of Potsdam, Germany

Since COVID-19 became a pandemic, many studies are being conducted to get a better understanding of the disease itself and its spread. One crucial indicator is the prevalence of SARS-CoV-2 infections. Since this measure is an important foundation for political decisions, its estimate must be reliable and unbiased. This paper presents reasons for biases in prevalence estimates due to unit nonresponse in typical studies. Since it is difficult to avoid bias in situations with mostly unknown nonresponse mechanisms, we propose the maximum amount of bias as one measure to assess the uncertainty due to nonresponse. An interactive web application is presented that calculates the limits of such a conservative unit nonresponse confidence interval (CUNCI).

Keywords: COVID-19; prevalence; probability samples; unit nonresponse; conservative confidence limits; nonresponse bias

1 Introduction

Since the start of the lockdown policies, influential scientists and organizations from the medical sciences (e.g., Deutsches Netzwerk für evidenzbasierte Medizin, 2020; Ioannidis, 2020) asked for population estimates of the prevalence of SARS-CoV-2 infections. Such prevalence estimates are also requested for the decisions regarding policy changes. Since these policies affect the economy, public and private life, and above all the health of the population, it is of great importance that these prevalence estimates are highly reliable and as unbiased as possible.

In the meantime, several studies from medical science have been fielded. These studies are designed to estimate—among others—the prevalence of SARS-CoV-2 infections. In Germany, the most visible instances are the COVID-19 Case-Cluster-Study, commonly known as the "Gangelt-Study" (Streeck et al., 2020), and "KoCo19", the prospective COVID-19 cohort study (Radon et al., 2020).

Most of these studies emphasize the use of probability samples as a prerequisite for the estimation of prevalence. At the same time, however, they often seem to pay less attention to the many other sources for biases that survey methodologists discuss under the heading of the total survey error framework (Groves et al., 2009).

Population estimates of prevalence require a well-controlled sampling design. The general research design for such population estimates thus differs significantly from the design of the laboratory and clinical studies used for the estimation of treatment effects that are typical in medical science. It is therefore advisable that survey methodologists and specialists from medical science work together.

In this respect, we aim to increase the awareness of *one* of the other possible sources for biases: unit nonresponse. Unit nonresponse may or may not lead to biased estimates of population parameters. However, if a bias occurs, univariate statistics—such as the prevalence—tends to be particularly sensitive to it. We will show that there are several ways how nonresponse can bias the estimate of the prevalence. Moreover, it turns out that it is not possible to formulate a clear expectation about the direction of the bias. Thus, we do not know whether the estimates of the prevalence are too high or too low. Given this situation of great uncertainty, we propose a straightforward method to estimate a more conservative interval for the estimate. This interval contains the prevalence without unit nonresponse with certainty. We also provide access to an online tool that allows for fine-tuning the method to more realistic scenarios, leading to narrower intervals.

This paper, and the intervals proposed therein, focus solely on the bias due to unit nonresponse. As already mentioned, researchers need to consider several sources for error

Contact information: Julia C. Post, Faculty for Economic and Social Research, University of Potsdam, August-Bebel-Str. 89, 14482 Potsdam (jupost@uni-potsdam.de)

to get an unbiased estimate (e.g., measurement error, sampling error, coverage error, etc.).

2 Nonresponse Bias

Many scientists consider probability samples as a precondition to make inferences from a sample to the population (e.g., Bethlehem, 2015; Cornesse et al., 2020; Kohler, 2019; MacInnis, Krosnick, Ho, & Cho, 2018). But even the best sampling design may fail if some of the selected units do not participate in the study (known as "unit nonresponse"). Unit nonresponse is a common problem in surveys; there has consequently been a lot of research into the reasons for and impact of unit nonresponse, as well as solutions to deal with it (e.g., Bethlehem, Cobben, & Schouten, 2011; Groves & Cooper, 1998; Schnell, 1986, 1997). The theoretical amount of bias in the sample mean originating from unit nonresponse is well known from Bethlehem (1988, p. 254).[1] In practice the calculation of the amount of bias is not possible since the information for the nonrespondents is missing. Nevertheless, the following equation can be used to visualize the factors that influence the amount of bias. In terms of the prevalence P of SARS-CoV-2 infections, the amount of bias in the estimated prevalence is approximately

$$\text{Bias}(\hat{P}) \approx \frac{\sigma(\text{SARS-CoV-2}) \cdot \sigma(\pi) \cdot \rho(\text{SARS-CoV-2}, \pi)}{\bar{\pi}}, \quad (1)$$

with σ(SARS-CoV-2) being the standard deviation of an indicator variable for SARS-CoV-2 infections in the population and $\sigma(\pi)$ being the standard deviation of the individual response probabilities. ρ(SARS-CoV-2, π) is the correlation between the SARS-CoV-2 indicator and the response probability, and $\bar{\pi}$ is the average of the individual response probabilities.

The denominator in equation (1) can be considered as the response rate. If the true denominator is high, the response rate will be high, and thus the bias small. That is intuitive: the less unit nonresponse, the smaller the unit nonresponse bias.

However, the equation also shows that nonresponse does not necessarily create bias. If the numerator were zero, low response rates would not create bias. If the numerator remains very small, the bias will remain small even for low response rates.[2]

The numerator is a product of three factors. Thus, it will become zero if one of the three factors is zero. So, what do we know about each of the three factors for the case of SARS-CoV-2 prevalence studies?

The first factor, σ(SARS-CoV-2), is the standard deviation of an indicator variable for an actual or passed SARS-CoV-2 infection. This factor would be zero if either everyone or no one in the population were infected with SARS-CoV-2. Since a prevalence study that starts from such an assumption is pointless, we assume σ(SARS-CoV-2) $\neq 0$.

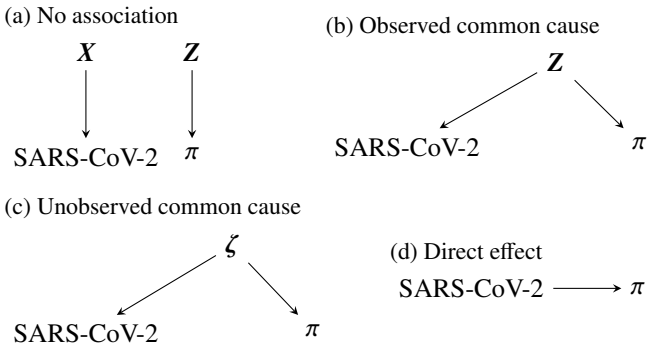

Figure 1. Missing data mechanisms

The second factor, $\sigma(\pi)$, would be zero if all sampled units had the same probability of participating in the study. The third factor, ρ(SARS-CoV-2, π), is the correlation between the response probability and an infection with SARS-CoV-2. The conditions for a zero or non-zero correlation will be discussed in section 3, depending on the assumed missing data mechanism. In section 4, we then give various practical examples to expect that ρ(SARS-CoV-2, π) $\neq 0$, and thus π is likely not a constant.

3 Missing Data Mechanisms

The concepts discussed here go back to Rubin (1976). In the methodological and statistical field, they are well known as "Missing Completely At Random" (MCAR), "Missing At Random" (MAR), and "Missing Not At Random" (MNAR). Since these terms are often misunderstood outside of the statistical community, we stopped short from using them here. For the sake of simplicity, we just give a rough idea of these mechanisms and introduce terms that are arguably less misleading for readers outside the field. For a detailed description see, for example, Enders (2010), Groves (2006), Rubin (1976).

Figure 1 shows simplified representations of four response mechanisms.[3] In each graph, arrows represent causal effects. The variable at the start of the arrow causally affects the variable at the tail of the arrow.[4] For the special case of this graphic, two variables are associated if it is possible to connect them with a path of subsequent arrows, no matter in which direction.

[1] Remember that an estimation of the prevalence is the sample mean of a variable with values 1 and 0 for having a SARS-CoV-2 infection, or not having one.

[2] Nonresponse is still a problem because it reduces the number of cases leading to inflated standard errors.

[3] The graphics are adapted from Groves (2006).

[4] See Elwert (2013) for an introduction to the formal graphical language used in the graphs.

In Figure 1a, some variables $\boldsymbol{X}$ cause a SARS-CoV-2 infection, and some other variables $\boldsymbol{Z}$ affect the response probabilities, π.[5] Since there is neither a connection between $\boldsymbol{Z}$ and $\boldsymbol{X}$ nor any other connection between SARS-CoV-2 and π, SARS-CoV-2 and the response probabilities are uncorrelated. Hence, this missing data mechanism would not cause any unit nonresponse bias in the prevalence estimate.

In Figures 1b and 1c, both the response probabilities and the SARS-CoV-2 infections are caused by one or more *common causes*, $\boldsymbol{Z}$ resp. $\boldsymbol{\zeta}$, which leads to a non-zero correlation of π and SARS-CoV-2. In Figure 1b, $\boldsymbol{Z}$ is assumed to be observed, so that it can be taken into account in the analysis in a way that removes the association, and thus the unit nonresponse bias. However, in Figure 1c, the common cause $\boldsymbol{\zeta}$ is considered to be *not* observed (or not observable). In this case, the common cause cannot be taken into account, which leads to a biased estimate of the prevalence. Observing the common causes could remove the bias. But therefore, it must be observed for both, the respondents and the nonrespondents.

Finally, Figure 1d shows the case when a SARS-CoV-2 infection directly affects the response probabilities. If this were the case, there would be a correlation between these two variables and therefore a bias in the prevalence estimate.[6]

4 Missing data mechanisms in COVID-19 studies

In the following, we discuss various arguments to sustain the assumption that the response probabilities are neither constant nor uncorrelated with SARS-CoV-2 infections for most, if not all, COVID-19 studies in the field. Generally, three main causes for unit nonresponse are considered:[7]

1. Refusal: The sampled unit (person or household) was contacted and would be able to participate but *refuses* to take part in the study.
2. Noncontact: The sampled unit could not be *contacted*.
3. Not-able: The sampled unit is *not able* to take part in the study.

Since the indicator variable for an actual or passed SARS-CoV-2 infection is determined by the infection probability, we use both concepts in the following.

For the discussion, we presume that COVID-19 studies take place in a lockdown situation, i.e., people are asked to stay home. Most shops are closed, alongside restaurants, cultural facilities, etc. Schools and kindergartens are also largely closed, and many people work from home. Under such conditions, it should be easier to contact the sampled units at home than under normal circumstances. Moreover, we assume that COVID-19 is considered an important topic by a large majority of the sampled units. Also, the question of whether one was already infected without experiencing symptoms may concern many of them. Hence, many people should be less averse to participating in such a study than in regular surveys. So the question is: Who are the nonrespondents?

Refusal

The cooperation probability which impacts the response probability evokes bias if people with a higher risk of a SARS-CoV-2 infection are more likely to participate than those with a lower risk. The same applies when people with a lower risk of a SARS-CoV-2 infection are more cooperative than people with a higher risk. If people with COVID-19 cases in their environment are likely to be more concerned about having an infection themselves, it is plausible that they have a higher cooperation probability.[8] To know someone with COVID-19 would thus be a common cause for the infection probability and the response probability. Depending on whether this cause is observed this would be mechanism 1b or 1c in Figure 1.

In the German KoCo19-study, the survey team is accompanied by police to emphasize the trustworthiness of the study (Radon et al., 2020). This certainly reduces the cooperation probability of people who are averse to the state, the government's recent decisions, or the police. At the same time, this cause of the cooperation probability may also affect the probability of a SARS-CoV-2 infection. This is the case when people who are averse to the state authorities are more likely to ignore the lockdown policies, which in turn increases the infection probability. This is another example of the common cause mechanism. While the first scenario would lead to a positive correlation, and thus to an overestimate of the prevalence, the second scenario suggests a negative correlation, and therefore an underestimate of the prevalence.

Another example is if some people have already been tested so that they know whether or not they have been infected with SARS-CoV-2. Hence, participating could be less attractive to them and therefore lead to refusals. So the (known) infection is a direct cause for the response probability. In this case, it is difficult to estimate the direction of the bias since we do not know if this is more likely for the infected or not-infected.

---

[5] We use bold upper case letters to refer to a set of variables. Every single variable in a set of variables should be thought of having its own arrow pointing to the respective variable at the arrow's tail.

[6] In these cases, the process that generated the missing data needs to be taken into account to try to get unbiased estimates. Such models are built on strong assumptions and therefore require knowledge about the missing data process (Enders, 2010; Groves, 2006; van Buuren, 2018).

[7] For a more detailed description see Groves et al. (2009).

[8] Christian Drosten, the director of the institute of virology at Charité mentioned this kind of mechanism for self-selection samples recently in his podcast https://tinyurl.com/ya8shkqj.

These are only some examples of cooperation scenarios, but others are conceivable. In many scenarios refusals will be determined by the attitude towards the topic or the survey circumstances. This will likely lead to a correlation between the response probability and the probability of a SARS-CoV-2 infection.

Noncontact

The contactability of sampled units for studies that rely on personal contact, crucially depends on the amount of time the sampled units stay at home. In the ideal-typical lockdown situation, most people are at home most of the time. Still, some people are not: not only medical staff, supermarket retailers and other jobs in which working from home is not possible, but also people who violate the contact restrictions or leave their homes a lot for other reasons. All these people have a higher infection probability than people who stay at home. There are again several common causes for the response probability and the infection probability. They all suggest a negative correlation, which leads to an underestimation.

But there are also reasons conceivable which suggest a positive correlation between the infection probability and the contactability of the sampled units. Especially people who live in flats in bigger cities could move to their allotments or weekend homes to escape from the density of the city. This gets even more likely due to new home office possibilities. If they do so, they are less accessible at home. Besides, moving to a more rural or spacious area could come along with a lower infection probability. In these cases, this would lead to an overestimate of the prevalence. These thoughts should also be taken into account when defining the target population and constructing the sampling frame.

Not-able

Not-able refers to people who are not able to take part in the study. This could be the case if someone requires care. In that situation, this person could have a higher infection probability because of close contact with caregivers who commute between many clients and thus could transmit the infection from one client to the other. The need for care is the common cause of the infection probability and the response probability. For those who are hospitalized due to COVID-19, there is a direct effect of a SARS-CoV-2 infection on the response probability. These scenarios would lead to an underestimation of the prevalence.

All these examples further show that it is not likely that the response probability is constant for all sampled units.

5 What can be done?

The previous section provided arguments to expect a correlation between SARS-CoV-2 infections and the response probabilities. Whether this causes bias in the estimate of the prevalence depends on the data collected, the methods used, and if the assumptions made are correct.

The first strategy to avoid unit nonresponse bias is to avoid nonresponse. Nonresponse can be decreased to some extent, although specific recommendations depend on many factors. Repeated contact attempts at different times, changing staff, and providing incentives certainly help—but there is more (Groves et al., 2009). All this needs to be well planned and takes a lot of knowledge and effort. Experienced survey methodologists are the right people to be asked for support.

Frequently, the reasons for correlations between SARS-CoV-2 infections and the response probabilities are a variant of the common cause mechanism. In these cases, the statistical remedies for the bias crucially depend on information about the common causes. Such information tends to be only available if plausible assumptions about the common causes have been developed *before* the actual data collection. If this has been done, nonresponse bias could be corrected by using missing data techniques such as weighting or imputation (Enders, 2010; Groves, 2006). However, applicants of those methods should stay aware that ritualized application will often not sufficiently reduce biases. Those methods require a sound theoretical understanding of the variable of interest as well as the nonresponse mechanism. For the case of the COVID-19 studies, they would require interdisciplinary cooperation between survey methodologists and epidemiologists.

If the common cause of nonresponse and a SARS-CoV-2 infection is not known or not observable, or if the infection probability directly affects the response probability and the missing data mechanism cannot be modeled in any other way, there is very little that can be done. Observing the necessary information to account for bias is crucial.

Since the COVID-19 pandemic is an entirely new situation, we have to acknowledge that the possible response behavior in the COVID-19 studies is fairly unknown. We thus propose a simple technique to estimate conservative boundaries of the minimum and maximum prevalence under a given amount of unit nonresponse.[9]

The idea of the minimum and maximum prevalence boundaries is to estimate the maximal possible change of the prevalence if all nonrespondents had participated in the study. The calculation of this value is simple, fast, and needs no assumptions about the missing data mechanism: First, calculate the prevalence under the assumption that none of the nonrespondents had a SARS-CoV-2 infection. Then, recalculate the prevalence under the assumption that all nonre-

[9]This idea is taken from Cochran (1977) who terms the boundaries "conservative confidence limits". Unlike the approach shown here, Cochran combines the idea with the uncertainty stemming from the sampling design. Since this combination is straightforward in principle, we only discuss the unit nonresponse part.

spondents had a SARS-CoV-2 infection. These two values generate an interval that contains the estimate of the prevalence without unit nonresponse with certainty.

The approach is illustrated in a software application that is online available on

https://lmes.shinyapps.io/maxbias/

The application runs in a browser. It is based on the situation of the COVID-19 Case-Cluster-Study conducted in the German community of Gangelt, a community of 12,529 inhabitants.[10] The gross sample of the study was 600 households. It observed around 1000 individuals from roughly 400 households, leading to an average household size of around $\frac{1000}{400} = 2.5$. Using the average household size, the number of nonrespondents in the 200 unobserved households are estimated to be $200 \cdot 2.5 = 500$. This implies a response rate of 1000/1500 or 67%.

The estimated prevalence of the Gangelt study was 15%. If none of the 500 nonrespondents were infected, the actual estimated prevalence would have been 10%. In contrast, if all nonrespondents were infected it would be 43%. Hence, the conservative unit nonresponse confidence interval (CUNCI) runs from 10% to 43%.

Of course, this interval is much wider than it would be under reasonable assumptions about the missing data mechanism (Schnell, 1997). Nevertheless, this approach is a good starting point. If there is no way to improve the estimation of the prevalence, one should at least talk about its uncertainty. Politicians and policy makers may ask themselves the question, whether they would make the same decisions for the entire range of uncertainty.

The CUNCI is very wide since its boundaries indicate the maximum impact of the nonrespondents. The web application can be used to change the assumed proportion of infections among the nonrespondents. This gives the user an impression of the prevalence under less extreme, and thus more likely, infection probabilities of the nonrespondents. Moreover, the application offers two additional sliders to change the response rate and the proportion of infections among the respondents. This can be used to adapt the situation to other studies.

References

Bethlehem, J. (1988). Reduction of nonresponse bias through regression estimation. *Journal of Official Statistics*, *4*(3), 251–260.

Bethlehem, J. (2015). Essay: Sunday shopping – the case of three surveys. *Survey Research Methods*, *9*(3), 221–230. doi:10.18148/srm/2015.v9i3.6202

Bethlehem, J., Cobben, F., & Schouten, B. (2011). *Handbook of nonresponse in household surveys*. New York: Wiley.

Cochran, W. G. (1977). *Sampling techniques* (third edition). John Wiley & Sons, New York.

Cornesse, C., Blom, A. G., Dutwin, D., Krosnick, J. A., De Leeuw, E. D., Legleye, S., … Wenz, A. (2020). A review of conceptual approaches and empirical evidence on probability and nonprobability sample survey research. *Journal of Survey Statistics and Methodology*, *8*(1), 4–36. doi:10.1093/jssam/smz041

Deutsches Netzwerk für evidenzbasierte Medizin. (2020). COVID-19: Wo ist die Evidenz? Retrieved from https://tinyurl.com/y8omeu3r

Elwert, F. (2013). Graphical causal models. In S. Morgan (Ed.), *Handbook of causal analysis for social research* (pp. 245–273). Dordrecht: Springer.

Enders, C. K. (2010). *Applied missing data analysis*. New York: Guilford Press.

Groves, R. M. (2006). Nonresponse rates and nonresponse bias in household surveys. *Public Opinion Quarterly*, *70*(5, Special Issue), 646–675.

Groves, R. M., & Cooper, M. (1998). *Nonresponse in household interview surveys*. New York: Wiley.

Groves, R. M., Fowler Jr, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2009). *Survey methodology*. John Wiley & Sons.

Ioannidis, J. P. (2020). A fiasco in the making? As the coronavirus pandemic takes hold, we are making decisions without reliable data. Statnews. Retrieved from https://tinyurl.com/uj539o4

Kohler, U. (2019). Possible uses of nonprobability sampling for the social sciences. *Survey Methods: Insights from the Field*, 1–12. doi:10.13094/SMIF-2019-00014

MacInnis, B., Krosnick, J. A., Ho, A. S., & Cho, M.-J. (2018). The accuracy of measurements with probability and nonprobability survey samples. *Public Opinion Quarterly*, *82*(4), 707–744. doi:10.1093/poq/nfy038

Mohan, K., Pearl, J., & Tian, J. (2013). Graphical models for inference with missing data. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Q. Weinberger (Eds.), *Advances in neural information processing system* (pp. 1277–1285). Curran Associates.

Radon, K., Saathoff, E., Pritsch, M., Guggenbuehl Noller, J. M., Kroidl, I., Olbrich, L., … Hoelscher, M. (2020). Protocol of a population-based prospective COVID-19 cohort study Munich, Germany (KoCo19). *medRxiv*. doi:10.1101/2020.04.28.20082743

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592.

[10] The numbers are taken from the following preliminary report: https://tinyurl.com/rlr7pc7. In the meantime another preliminary report with slightly different numbers has been published (see Streeck et al., 2020).

Schnell, R. (1986). *Missing-Data-Probleme in der empirischen Sozialforschung*. Bochum: Dissertation Universität Bochum.

Schnell, R. (1997). *Nonresponse in Bevölkerungsumfragen. Ausmaß, Entwicklung, Ursachen*. Opladen: Leske und Budrich.

Schuessler, J., & Selb, P. (2019). Graphical causal models for survey inference. SocArXiv. doi:10.31235/osf.io/hbg3m

Streeck, H., Schulte, B., Kümmerer, B. M., Richter, E., Höller, T., Fuhrmann, C., ... Hartmann, G. (2020). Infection fatality rate of SARS-CoV-2 infection in a German community with a super-spreading event. Retrieved from https://tinyurl.com/yblwgmtt

Thoemmes, F., & Mohan, K. (2015). Graphical representation of missing data problems. *Structural Equation Modeling*, *22*, 631–642.

van Buuren, S. (2018). *Flexible imputation of missing data* (second edition). Boca Raton: Chapman & Hall/CRC Press.

Commentary

We really enjoyed reading this paper and appreciate the idea of connecting survey research methods with epidemiological research. We agree with the authors that the interplay of both will increase the overall quality of studies.

From our point of view, the paper is well-structured with a good to follow argumentation. The authors illustrate the theoretical basics of nonresponse bias and the assumptions that are given in a way that enables interested readers that are not familiar with the whole nonresponse literature to get a sense of potential problems.

They start with the motivation of the paper and why they have chosen unit nonresponse as one central problem regarding COVID-19 prevalence estimation. Based on the formula of nonresponse bias, they are illustrating the elements and the assumptions that lead to a biased or unbiased estimate of the prevalence in the following section. Subsequently, they give a brief introduction to the logic of missing data mechanisms that need to be understood in order to evaluate whether bias in the estimate of prevalence can be expected and whether it is possible to correct for a potential bias post hoc. Those theoretical considerations are followed by practical examples for missing data mechanisms in COVID-19 studies based on nonresponse reasons due to noncontact, non- ability and noncooperation. Based on a few examples, the authors vividly depict that there are arguments for overestimation as well as underestimation of infection prevalence.

Given the uncertainty of nonresponse mechanisms, the authors propose an estimation of the minimum and maximum prevalence boundaries. It can be interpreted as one strategy to illustrate the uncertainty that comes along with all studies of COVID-19 prevalence. For illustrative purposes they provide a web application that intuitively shows how the variation of different parameters affects the uncertainty of the estimates. The short paper closes with a wrap-up of survey researchers' abilities to increase quality of prevalence studies.

In the following, we would like to focus to some points of the paper that could have been addressed in more detail and that we would expect to be discussed in a more comprehensive paper:

Post hoc bias estimation and causes of nonresponse

In the last section of their paper the authors state that "Observing the necessary information to account for bias is crucial." However, throughout the text they do not discuss the necessary prerequisites and limitations of post hoc bias estimation. It has been shown that bias correction is not trivial. It is only possible when common causes or potential confounder variables are known for respondents and nonrespondents and target variables and response mechanism are not strongly correlated (see Groves, 2006).[11] This fact is only implicitly mentioned in the paper. Thus, the paper would profit from the discussion about observability of common causes that the authors speculate about in section 4. For instance, whether someone knows someone with an infection is an information that is not available for the population. This fact has practical implications for the possibility of bias correction. As a logical next step, it is of great importance to learn more about causes of nonresponse in COVID-19 prevalence studies to be able to better quantify uncertainty of prevalence measures in future studies.

Application of DAGs on the nonresponse problem

Even though we only have limited expertise in the application of DAGs, we would like to discuss two statements in the paper. In the paper it is stated: "In Figure 1a, some variables X cause a SARS-CoV-2 infection, and some other variables Z affect the response probabilities, π". This is, in our opinion, where Groves and DAGs do not mix well. Following Elwert (2013, p. 248) "DAGs consist of three elements: variables (nodes, vertices), arrows (edges), and missing arrows. Arrows represent possible direct causal effects between pairs of variables and order the variables in time." It is our understanding that nodes are variables, not probabilities. Since DAGs can be translated into probability statements, we were wondering if it makes sense to speak of a probability of a probability (distribution). In Elwert's (2013: 246) words: "DAGs are visual representations of qualitative causal assumptions: They encode researchers' expert knowledge and beliefs about how the world works. Simple rules then map these causal assumptions onto statements about probability distributions". Next, DAGs are being used to illustrate bias in the prevalence estimate: "Finally, Figure 1d shows the

[11] References are shown in the bibliography of the main article.

case when a SARS-CoV-2 infection directly affects the response probabilities. If this were the case, there would be a correlation between these two variables and therefore a bias in the prevalence estimate." Here, we refer to a related approach, using m-graphs, which might be a useful alternative since they claim to represent missing data problems using graph theory, i.e., DAGs (see Mohan, Pearl, and Tian, 2013; Thoemmes and Mohan, 2015 and, for a recent working paper see Schuessler and Selb, 2019).

In sum, this article marks a starting point of an important discussion. It is of utmost importance to learn more about causes of nonresponse in studies of COVID-19 prevalence to be able to better quantify at least the direction of a potential nonresponse bias. This is even more important if prevalence is not only studied in a restricted regional context as in the cited studies but will be investigated in huge studies that aims at estimating prevalence for the whole country.

Ines Schaurer, and Bernd Weiß
GESIS – Leibniz Institute for the Social Sciences
Mannheim, Germany

Reply to Schaurer and Weiß

We thank Ines Schaurer and Bernd Weiß for the gentle evaluation of our paper. We also explicitly agree with their skepticism towards the possibilities of post hoc bias correction. Such corrections require carefully tested models of the data generating process and valid data about the variables of the assumed data generating process. It seems to us that such corrections are primarily promising for well-studied research topics, and probably not for research on the prevalence of SARS-CoV-2 infections. Thus, any post hoc bias correction should be complemented by a design based approach.

We would like to make our stand clear concerning the meaning of the probability π in the DAGs of Figure 1. First of all, while we welcome graph theoretical approaches to missing data processes, we used the DAGs primarily as an illustrative device for the theoretical arguments. At the same time, however, in our understanding π is, in fact, a variable. We assume that each research unit has an *individual* probability to respond to a request to participate in a specific sample, given the situation the request arrives. This *individual* probability might correlate with some known or unknown covariate-set, $\boldsymbol{Z}$. We consider the individual response probability to be a latent variable that affects the decision to actually participate in a survey.[12] We thus believe that there is no inconsistency to the cited approaches that use a (manifest) dichotomous selection indicator variable to formalize the missing data mechanism.

It might be the case that this "controversy" originated from our decision to notate the *individual* probability without an explicit subscript for the individual. We did this to simplify the notation. We apologize if this simplification created misunderstandings.

Julia C. Post, Fabian Class, and Ulrich Kohler

[12]This is similar to the underlying probability that affects the manifest response to a dichotomous survey question.