Survey Research Methods (2026)
Vol. 20, No. 1, pp. 97-114
https://doi.org/10.18148/srm/2026.v20i1.8447
European Survey Research Association

Survey Design and Quality During the COVID-19 Pandemic in Germany: An Assessment with 686 Social Science Surveys

Karolina von Glasenapp[1] · Thomas Skora[1] · Tobias Gummer[1,2] · Elias Naumann[1]
[1]GESIS – Leibniz Institute for the Social Sciences
[2]School of Social Sciences, University of Mannheim

During the COVID-19 pandemic, there was a high demand for readily available but also reliable survey data. Overall, a comprehensive assessment of how survey data were collected and what their quality was remains missing. In this study, we provide a multi-dimensional quality assessment of social science surveys conducted during the COVID-19 pandemic in Germany (*N*=686). We assess survey quality based on three dimensions: accuracy (proxied by survey design), interpretability (referring to the quality of documentation), and accessibility of data (referring to the timely publication of results and data). Our results show that surveys varied considerably in these three quality dimensions over time. We found that surveys followed different purposes at different times of the pandemic: whereas early surveys focused on quickly producing results and traded other aspects of survey quality for this goal, later surveys were more focused on operating better-designed surveys and producing shareable data.

Keywords: survey design; data quality; total survey quality; total survey error; COVID-19 pandemic; accuracy; interpretability; accessibility

1 Introduction

During the COVID-19 pandemic, survey data were an important source for social science research. These data enabled researchers to assess the attitudes and behaviour of the general population during these challenging times (Pierce et al., 2020). The insights gathered, particularly at the onset of the pandemic, have been instrumental in assisting scientists and policymakers in developing strategies to combat the spread of the virus (e.g., Adams-Prassl et al., 2020; Arpino et al., 2021; Huebener et al., 2021; Munzert et al., 2021).

The social sciences saw a significant rise in data collection activity as the pandemic unfolded as researchers rushed to adapt to the rapidly evolving situation. These data collection efforts included newly designed surveys specifically aimed at gathering information during the pandemic (e.g., Betsch et al., 2020; Busemeyer, 2023; Fetzer et al., 2021; Yamada et al., 2021). However, pre-existing or pre-planned surveys were also administered without necessarily being motivated by the COVID-19 outbreak (e.g., European Social Survey European Research Infrastructure [ESS ERIC] 2023; Kapteyn et al. 2020; Kühne et al. 2020).

The pandemic and political decisions such as quarantines and travel restrictions posed challenges to traditional face-to-face-based survey designs (e.g., Burton et al., 2020; Gummer et al., 2020). As a consequence, some surveys had to adapt their survey design to ensure that data collection could be conducted during COVID-19. Although some studies have acknowledged the adjustments made to their data collection protocols, these reports are sparse and pri-

Supplementary Information The online version of this article (https://doi.org/10.18148/srm/2026.v20i1.8447) contains supplementary material, which is available to authorized users.

Corresponding author: Karolina von Glasenapp, GESIS – Leibniz Institute for the Social Sciences, Mannheim, Germany (Email: Karolina.Glasenapp@gesis.org)

marily originate from existing infrastructure programs like the Survey of Health, Ageing, and Retirement in Europe (SHARE) (A. Scherpenzeel et al., 2020), Understanding Society Study in the UK (Burton et al., 2020), German Internet Panel, German Family Demographic Panel (FReDA), German Family Panel (pairfam) (Gummer et al., 2020), and various labor market surveys carried out by the German Institute for Employment Research (Sakshaug et al., 2020); see also the special issue published by Survey Research Methods (Kohler, 2020). These individual reports on data collection practices focus on single or a few studies and differ widely in the information they convey. A comprehensive overview of the data collection practices of social science surveys during the COVID-19 pandemic is largely lacking. Notable exceptions are two review reports on changing social research practices (Nind et al., 2021, 2022). However, the reports neither include a quantitative assessment of research practices, nor do they dive into the particularities of different survey design decisions and aspects of survey quality.

The existing research gap is unfortunate, since the survey designs used to collect data during the COVID-19 pandemic are unknown. As demonstrated by the Total Survey Error (TSE) framework (e.g., Groves et al. 2009; Groves and Lyberg 2010; Lyberg and Weisberg 2016), survey design determines the quality of the data obtained. Thus, without a systematic assessment and comparison of survey designs, the reliability of the data remains unclear on which impactful political actions were taken during the pandemic, and on which important social science research was based. We see merit in investigating the constraints under which social scientists were operating during the COVID-19 pandemic and what trade-offs regarding different dimensions of survey quality they were facing. Furthermore, it remains an open question as to whether data collection practices and, thus, survey quality changed over the course of the pandemic (e.g., due to changing demands for data). If designs changed, variations in substantive measures (e.g., vaccination preference) across the pandemic might be methodological artifacts.

To address these pressing issues and gain deeper insights into the quality of social science survey data, we used a dataset published by the Survey Data Collection and the COVID-19 Pandemic (SDCCP) project (Gummer et al., 2024) on the design and quality of 686 academic social science surveys conducted between March 2020 and December 2021 in Germany. In our study, we investigated two research questions (RQs):

RQ1. What was the quality of the social science surveys conducted during the COVID-19 pandemic?

RQ2. How were the different dimensions of survey quality associated with each other during the COVID-19 pandemic?

In the next section, we provide a background on survey data quality frameworks and derive a multi-dimensional perspective on quality for our study. We also detail how survey design impacts the accuracy of survey data. We then introduce our data, measures, and methods. After presenting our results, we conclude with closing remarks and an outlook for future research opportunities.

2 Background

In the literature, *survey quality* is a multi-dimensional concept covered by a number of theoretical frameworks such as the TSE framework (e.g., Biemer & Lyberg, 2003; Groves et al., 2009; Lyberg & Weisberg, 2016), Total Survey Quality (TSQ) framework (Biemer 2010), and FAIR framework (Wilkinson et al., 2016). The TSE is a well-established concept that provides a systematic overview of survey error sources. According to the TSE, high-quality survey data is characterized by minimized errors in survey statistics. The TSQ framework builds on the TSE framework and extends it in dimensions other than accuracy. According to Biemer (2010), the TSQ framework is not uniformly applied and, thus, may differ across institutions. However, the most commonly used dimensions (in addition to accuracy) are credibility, comparability, usability/interpretability, relevance, accessibility, timeliness/punctuality, completeness, and coherence. In addition to the TSQ framework, the perspective of data users is captured by the FAIR Principles (Wilkinson et al., 2016). These principles formulate the requirements for ensuring optimal data management that in turn facilitates the reuse of data. The four FAIR principles require data and metadata to be findable, accessible, interoperable, and reusable.

In practice, however, the empirical assessment of survey quality is limited due to the difficulties related to the operationalization of quality dimensions. Consequently, reports evaluating the survey quality of all surveys in a specific domain are rare. Moreover, they are published mainly by the data producers themselves and are restricted to data from a single/small number of surveys (e.g., for the European Social Survey: ESS ERIC Core Scientific Team 2022; for the Sample Survey of Income and Expenditure: Statistisches Bundesamt [Destatis] 2021).

For the purpose of our study, we derived an applied survey quality framework from the TSE, TSQ, and FAIR frameworks. Our framework covers three dimensions: accuracy, interpretability, and accessibility. We relied on quality dimensions (and measures, see below) that can be assessed using publicly available survey information. We decided to

include dimensions that cover the data producers' and data users' perspectives to acknowledge the constraints and demands imposed by the COVID-19 pandemic.

2.1 Accuracy

Defined as the difference between the survey estimate and the underlying "true" value, *accuracy of data* is a key quality aspect and a necessary condition for correct inferences drawn from the data. For this reason, data producers pay great attention to optimizing accuracy. According to the TSE framework (e.g., Biemer & Lyberg, 2003; Groves et al., 2009; Lyberg & Weisberg, 2016), accuracy is maximized by the minimization of survey errors arising throughout the data collection process. In practice, accuracy can be evaluated through the comparison of survey estimates with known true values (i.e., benchmarks). Examples of this approach include vote choice in election studies (e.g. Kennedy et al., 2018; Sohlberg et al., 2017; Sturgis et al., 2018), more recently the evaluation of the COVID-19 vaccination coverage estimates (Bradley et al., 2021; Nguyen et al., 2023), and other characteristics for which benchmarks exist, for instance, from population censuses. Studies applying this approach are described in the following subchapter. All the accuracy studies we list share a common emphasis on and confirmation of the crucial role that survey design plays in shaping accuracy.

2.1.1 Literature Review: The Relationship Between Survey Design and Accuracy

According to the TSE framework, accuracy is directly associated with survey design. While the complexity of the design may vary significantly across surveys, two important characteristics—sampling procedure and survey mode—constitute the integral components of every data collection. In the following, we provide a literature-based assessment of the impact of sampling and mode on accuracy.

Sampling procedures can be divided into two main groups: probability- and non-probability-based samples. As a result of the different selection approaches, probability and non-probability surveys may differ to the extent to which the sample enables drawing conclusions about the target population. The literature comparing both sampling procedures is based mainly on the comparison of survey estimates with external population benchmarks. Based on this approach, a number of studies have shown that probability surveys of different modes are more accurate than non-probability web surveys. More concretely, these studies included comparisons of non-probability surveys with probability personal surveys (Malhotra & Krosnick, 2007; Rohr et al., 2024; A. C. Scherpenzeel & Bethlehem, 2011; Sturgis et al., 2018), probability telephone surveys (Chang & Krosnick, 2009; MacInnis et al., 2018; Sohlberg et al., 2017; Yeager et al., 2011) as well as probability web surveys (Chang & Krosnick, 2009; MacInnis et al., 2018; Rohr et al., 2024; Yeager et al., 2011). In contrast, some studies found that probability surveys did not consistently outperform non-probability ones. Instead, the performance relative to the probability survey varied by the provider of the non-probability sample (Kennedy et al., 2016) and by the examined variables (Chan & Ambrose, 2011; Loosveldt & Sonck, 2008; Steinmetz et al., 2014).

Expanding on these individual studies, systematic studies have been conducted that synthesize findings concerning the accuracy of probability and non-probability surveys of previous research. For example, a meta-analysis of 110 surveys concluded that probability surveys are more accurate than non-probability surveys (Cornesse & Bosnjak, 2018). Furthermore, a comprehensive overview of empirical evidence confirmed that this conclusion holds across different countries and topics (Cornesse et al., 2020).

Finally, we would like to note that *weighting* is a commonly employed method with the potential to improve accuracy. Although a few studies have found a positive effect of weighting on the accuracy of non-probability samples (Steinmetz et al., 2014; Wang et al., 2015), in the majority of studies, weights did not consistently and sufficiently improve the accuracy of estimates (Chang & Krosnick, 2009; Loosveldt & Sonck, 2008; MacInnis et al., 2018; Malhotra & Krosnick, 2007; Yeager et al., 2011).

Overall, taking the above-listed literature into account, we conclude that probability surveys are more accurate than non-probability surveys.

Regarding the survey mode, today, non-probability surveys are conducted typically as web surveys. In contrast, probability surveys rely on a variety of modes (or their combinations) such as web, telephone, mail, and face-to-face surveys. The evaluation of how the survey mode affects accuracy is a complex issue, since the survey mode can cause both measurement- and representation-related survey errors. Consequently, the potential trade-offs between survey errors leave the overall effect of the mode unclear. To circumvent this difficulty, in the following, we describe the literature on the two branches of error that the TSE includes—representation and measurement—as well as on the effect of the survey mode on overall accuracy (i.e., the total error, not differentiating between representation and measurement).

On the representation branch of the TSE framework, one of the main issues related to the survey mode is the potential exclusion of certain groups of the target population. This exclusion leads to bias if the excluded group systemati-

cally differs from those included in the survey sample. One widely discussed specific group is the offline population that differs from the online population in various sociodemographic characteristics (Blom et al., 2017; Leenheer & Scherpenzeel, 2013; Revilla et al., 2016). This systematic difference between the offline and online population is of high relevance for web surveys, which need to ensure that the offline population is not excluded from the survey. This inclusion could be achieved by the provision of an online device or an alternative survey mode. However, Bach et al. (2024) found that the provision of a device did not significantly change the univariate and multivariate survey estimates. Cornesse and Schaurer (2021) concluded that the provision of an alternative mode has a comparatively larger effect on sample accuracy. The advantage of a mixed-mode survey as compared to a solely-web survey has been confirmed further by a study by Pforr and Dannwolf (2017) that shows the prevalence of bias in the latter survey design, which cannot be corrected with nonresponse weights.

On the measurement branch of the TSE framework, the presence of an interviewer plays a decisive role. Two measurement errors that are widely discussed in relation to the presence of an interviewer are *social desirability bias*, which is respondents' tendency to provide socially desirable answers (e.g., Krumpal, 2013; Phillips & Clancy, 1972; Tourangeau & Yan, 2007) and *satisficing*, which is respondents' behaviour aimed at reducing the required cognitive effort by providing satisfactory answers (e.g., Krosnick, 1991; Krosnick et al., 1996). With respect to social desirability bias, several studies have found a greater prevalence in interviewer-based surveys as compared to self-administered surveys (Cernat et al., 2016; Hope et al., 2022; Tourangeau et al., 2013). In addition, a meta-analysis by Dodou and de Winter (2014) did not find any difference in the level of social desirability bias between the self-administered modes of paper-based and computer-based surveys. Regarding satisficing, studies have shown that self-administered surveys are more prone to satisficing behaviour than interviewer-administered surveys (see Cernat and Revilla 2021 for a comparison of face-to-face and web interviews; Hope et al. 2022 for a comparison of computer-assisted personal interviews, telephone, and web).

Finally, the literature examining the effect of a survey mode on the total survey error is rare; to our knowledge, only one such study exists (Felderer et al., 2019). In this study, the authors evaluated the combined bias of nonresponse and measurement error in a survey that randomly assigned respondents to a telephone or web survey. Their results show that the combined bias is smaller in the telephone survey than in the web survey. In line with the above-described literature, the web survey showed a comparatively larger nonresponse bias (representation branch) that was not outweighed by the comparatively smaller measurement error.

In summary, the previous literature suggests that the survey mode has multiple effects on data quality. Regarding the representation-related errors, surveys conducted solely via the web risk bias due to the exclusion of an offline population. Regarding the measurement-related errors, self-administered modes (including web surveys) perform better than interviewer-administered modes in terms of social desirability bias, but worse in terms of satisficing. The comparison of different self-administered modes does not show significant differences between web and other modes. Regarding the overall effect of the survey mode on data quality, the literature indicates that the combined bias is lower in telephone surveys than web surveys. All findings considered, we conclude that among probability surveys, web surveys are less accurate than surveys conducted in other mode(s).

2.2 User-Centred Quality Dimensions

Moving from the perspective of data producers to data users, the second dimension incorporated into our framework is interpretability. According to the TSQ framework (Biemer, 2010), *interpretability* requires that the survey is clearly documented and that the metadata are well managed. This enables data users to comprehend the applied method and, consequently, evaluate the quality of the survey (Jedinger et al., 2018). Studies examining the data user-centred dimensions of survey quality are relatively scarce. Interpretability in form of the quality of methodological documentation has been assessed by two studies focusing on cross-national surveys in Europe (Jabkowski, 2023; Jabkowski & Kołczyńska, 2020) and a study encompassing multiple continents (Kołczyńska & Schoene, 2018). These studies have concluded that documentation quality varies across projects but has improved over time. Another study by Eder and Jedinger (2019) measured the quality of the methodological metadata of national election studies as part of the FAIR framework assessment, and found a generally satisfactory provision of metadata. Notably, all these studies focused on well-established large-scale surveys. To our knowledge, the only study that comprehensively evaluated the quality of reporting among diverse surveys is Stefkovics et al. (2024). This study, which examined the survey reporting in high-ranked journals between 2011 and 2021, concluded that the reporting level was generally high and stable over time. However, it also found that researchers tended to omit certain descriptors of survey design characteristics and performance (e.g., response or other outcome rates) (Stefkovics et al., 2024).

The third dimension of our framework is *accessibility*, which denotes the publication of results and data and their timeliness. The publication of results and data is an important quality indicator in the TSQ framework (Biemer, 2010) and in the FAIR framework (Wilkinson et al., 2016). The publishing of data and results in a swift and timely manner has gained further importance during the COVID-19 pandemic (Moretti & Santi, 2020). Regarding the accessibility of results, to the best of our knowledge, no studies have analysed the extent and speed of survey results publication. However, with regard to the publication of results in general, evidence exists that the pandemic accelerated the publication process of COVID-19 related journal articles in disciplines such as medicine (Forti et al., 2021; Horbach, 2020) and ecology (Forti et al., 2021). Regarding the publication of data in general, Eder and Jedinger (2019) investigated the case of election studies and showed that all the examined studies provided data access, usually upon registration. We are not aware of any studies that have evaluated the accessibility and timeliness of data for a large set of diverse surveys and with a focus on their association with survey design.

3 Data and Methods

3.1 Data

We used a dataset on survey designs and quality collected by the SDCCP project (Gummer et al., 2024). Our analyses are based on data release v1.0.0 of the SDCCP dataset (von Glasenapp et al., 2024). The SDCCP data aims at covering all social science and public health surveys of the general population conducted in Germany between March 1st, 2020 and December 31st, 2021. To identify eligible surveys, we systematically searched multiple national and international data archives, data research centers, and survey listings. In addition, in the screening, we included surveys we found through the web. In total, we screened 896 records and included 717 surveys in the final dataset. These surveys were conducted as part of 183 survey programs (e.g., waves of a panel) and were clustered within these programs. On average, a survey programme included 3.92 surveys. The distribution of surveys within survey programs was skewed (skewness = 4.65) with a median and mode of 1 (minimum = 1, maximum = 59). 55% of the survey programs included only one survey. The dataset includes 5 high-frequency survey programs with more than 30 surveys. See our methods section for how we handled the clustered data structure.

All surveys in the SDCCP dataset were manually coded by expert coders between November 2022 and the end of May 2023. The coding was based on information available on project websites, study documentations, and fieldwork reports. The coding scheme obtained extensive information on the survey design and the survey quality dimensions relevant here. Tests of double-coding subsets of the dataset indicated excellent intercoder reliability (categorical variables: average Brennan and Prediger's coefficient = 0.87; continuous variables: average intra-class correlation = 0.99). Further information on the sample selection, the coding process, and the variables included in the dataset can be found in the data descriptor (Gummer et al., 2024) and dataset (von Glasenapp et al., 2024).

3.2 Measures

3.2.1 Dependent Variables

We distinguish between different survey designs categorized with regard to sampling and survey mode. With respect to sampling, we differentiate between probability-based surveys (if each member of the population has a known and non-zero probability of being selected) and non-probability-based surveys (if the probability of selection is zero or unknown). Combining the categories of sampling with the survey mode, we define the following survey design groups: (1) non-probability & web, (2) probability & web, (3) probability & telephone, and (4) probability & mixed-mode. Due to their low prevalence during our observation window and the correspondingly low case numbers, we combined personal interviews and paper interviews into a fifth category: (5) probability & personal interview or (self-administered) paper interview.[1,2]

Regarding the dimension of interpretability, we created an additive index based on five binary variables that reflected the availability of information on key survey design features—target population, concrete sampling procedure, sample size, date of data collection (at the day level), and any outcome rate.[3] Each variable takes the value of 1 if

[1] This category comprises two very different modes and continues to have low case numbers. Therefore, the possibilities for interpreting this category separately in the context of the subsequent correlation analyses are very limited. Although it is included in the following (regression) analyses, we will not provide an analysis of its content.

[2] In addition to these 5 categories, one survey program (with three surveys) was conducted as a non-probability mixed-mode survey. We excluded this program from the analysis due to the small number of observations in this category.

[3] The variable "any outcome rate" comprises response rates, completion rates, break-off rates, and any general information regarding the proportion of completed interviews. Although not every survey design allows for the computation of all the rates listed, information on at least one can be calculated for each survey in our dataset. More detailed in-

information is available and 0 otherwise. As a result, the values of the additive interpretability index lie in the range between 0 (no information available) and 5 (information on all considered variables available).

To determine accessibility, we assessed whether and when first results and survey data were published. We recorded this variable as the time in months between the beginning of fieldwork and the publication of first results or of the dataset. Regarding the accessibility of first results, we considered any type of publication such as scientific articles, reports, and summaries on project websites.

3.2.2 Independent Variables

The month of the fieldwork start served as a measure for assessing calendar time effects and, together with the survey design (see above), was included as a central independent variable in the subsequent analyses.

3.3 Methods

To answer RQ1 ("What was the quality of the social science surveys collected during the COVID-19 pandemic?"), we provide an overview of each survey quality dimension and assess how they developed over time between March 2020 and December 2021. Given the different specifications of each dimension, we applied diverse approaches. To assess survey designs, we plotted the frequency of survey design choices across time. Regarding interpretability, we showed the development of the mean index score, and with respect to accessibility, we examined the proportion of surveys for which results or data have been published up to a certain point in time.

To answer RQ2 ("How were the different dimensions of survey quality associated with each other during the COVID-19 pandemic?"), we investigated the relationship between survey design and the user-centered quality dimensions interpretability and accessibility. For this purpose, we utilized different regression methods. In all the regression models, we used the respective data user-centred quality indicator as the dependent variable, and the calendar time (i.e., fieldwork start date) and survey design as independent variables. We built the models stepwise, first only including the fieldwork start dates and then adding survey design characteristics. We decided on stepwise modelling to be able to assess which parts of the variance of data quality was accounted for by design characteristics. Based on the properties of the dependent variable, we selected the regression method.

Regarding interpretability, we used linear models to estimate the effects of calendar time and survey design. To account for the clustering of single surveys in survey programs, we decided to rely on mixed models.

Regarding accessibility, we utilized (discrete-time) event analysis models (Singer & Willett, 2003) in which the dependent variable is binary and indicates whether the event 'publication' (of the results or data, respectively) occurred between two time points or not. We estimated the transition rate using a logit link function. The time that a survey has already been exposed to the risk of event occurrence corresponds to the so-called *process time*. As is typical for this approach, we included a measure of process time in all the event-history regression models in addition to the variables of interest (calendar time and survey design). As the starting point of process time, we chose the month in which the fieldwork began. This time ends as soon as a publication is recorded for the first time. If no publication was observed by the end of the observation window (in our case the end of May 2023), the respective survey observation is right-censored. Regarding the date of the data publication, we modelled the process time using a linear specification. With respect to the first publication of results, we modelled the process time by using a quadratic specification based on the logarithm of the process time.[4] Regarding all the event history regressions, we used mixed models with cluster-robust standard errors to account for the clustering of single surveys in survey programs.

3.3.1 Handling of Missing Data

Regarding the surveys included in the SDCCP data, information about all the relevant variables was not publicly available for all the surveys. Consequently, we applied listwise deletion and excluded all observations for which no information was available on the fieldwork start date, sampling, or mode. After deleting cases with missing data (4% of the original dataset), we were left with 686 surveys. Likely, this information is not missing at random. Assuming that missing information is linked to lower survey data quality, listwise exclusion of these cases from the analysis may lead to an overestimation of survey quality. Furthermore, the accessibility dimension was affected by missing information (usually concerning the month of publication), which further restricted the sample sizes for these analyses.

formation on the variables can be found in the data descriptor (Gummer et al., 2024).

[4] We made decisions in favor of these two modelling options based on a comparison with the unparameterized transition rate and on a comparison of the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) across different modelling alternatives.

Consequently, the analyses of the accessibility of results were based on 483 surveys (an exclusion of 33% of the original dataset), and the analyses of the accessibility of data were based on 590 surveys (an exclusion of 18% of the original dataset).[5] Survey design groups differ in their degree of missing information, and these differences seem to be driven by probability-based telephone surveys—that when compared to other survey designs—more frequently lack information on the publication date of results and less frequently lack information on the publication date of data.

To examine whether our approach of handling missing information on the date of publication affected our findings, we conducted robustness checks using additionally the full dataset. These checks suggest that the main findings are not significantly biased by the restriction of the dataset. We provide more details on the robustness check in the Results section.

3.3.2 Handling of Clustered Data Structure

In our main analytical sample, the 686 surveys are clustered in 165 survey programs. As in the full dataset, this distribution is skewed (skewness = 4.41) with a mean of 4.15, and median and mode of 1. The main analytical sample includes 5 high-frequency survey programs with more than 30 surveys each. Given our goal to evaluate the German survey landscape with all cases following our definition of a survey, we decided to not exclude any observations. Instead, we acknowledge the clustered structure of our data through mixed models with clustered standard errors.

4 Results

4.1 Survey Quality of Social Science Surveys During the Pandemic

Our data showed that among data collections conducted in Germany between March 2020 and December 2021, the most frequently used *survey design* was non-probability surveys conducted in the web mode (48% of all the surveys in the dataset). The second largest group of survey designs in our observation period consisted of probability surveys that utilized the telephone as a survey mode (38%). Only a comparatively small proportion of the surveys (less than 10% each) relied on other modes of probability-based sampling designs—web, personal/paper, and mixed-mode.

[5] The mean and standard deviation regarding the independent variables based on each of the three analytical samples is provided in the appendix (Table A1).

Interpretability refers to the information available on survey design characteristics and survey performance. For our entire dataset, we found a mean interpretability score of 3.78. Additionally, the majority of surveys (67%) scored 4 or 5 on the interpretability index (range 0 to 5). In other words, these surveys provided the larger part of the information required for external researchers and data users to evaluate the survey design features and performance. A notable exception to the availability of information is the outcome rate, which is the least-available component of our index with only 21% of surveys providing the information. This is surprising, since outcome rates (of any kind), in our view, are quite well-known and popular among survey researchers. Furthermore, there may be a difference in the reporting standards of outcome rates between probability and non-probability surveys. While 28% of the probability surveys provided information on at least one outcome rate, only 15% of the non-probability surveys did so. Both the overall finding on the comparatively low frequency of outcome rates reporting and the difference by sampling procedure resembled the results of a study on the quality of reporting on survey design in top journals (Stefkovics et al., 2024).

Regarding the *accessibility* of first results and survey data, whether and when (i.e., how quickly) the outputs are made available to researchers and the public are the decisive factors. Fig. 1 provides insights into the accessibility of results and survey data across our entire dataset by displaying the cumulative probability of the event occurrence (i.e., inverted survival functions) of both outcomes as a total and separately by survey design categories. These graphs reflect the probability that the event of interest has occurred up to a particular point in time.

With regard to results, our analyses point to a rapid publication practice in many cases: for more than half of the surveys (52%), first results already were published within two months after the start of fieldwork. Within one year after the fieldwork start, first results had been published for almost three-quarters of the surveys (74%). For 19% of all the surveys in the entire dataset (not excluding cases with known publication but an unknown date), however, no results were detectable at all during the coding process of the SDCCP dataset used for this analysis. Concerning survey design, we observed two different patterns of results publication. First, the group comprising probability telephone surveys, probability personal interview or (self-administered) paper interview, and non-probability web surveys resembled the overall development of rapid as well as an overall high share of publication. Second, probability web and mixed-mode surveys published their results at a considerably lower pace and extent. Overall, the rapid publication of first results for the majority of the sample is in line with our prior reasoning that early data collection efforts were stimulated

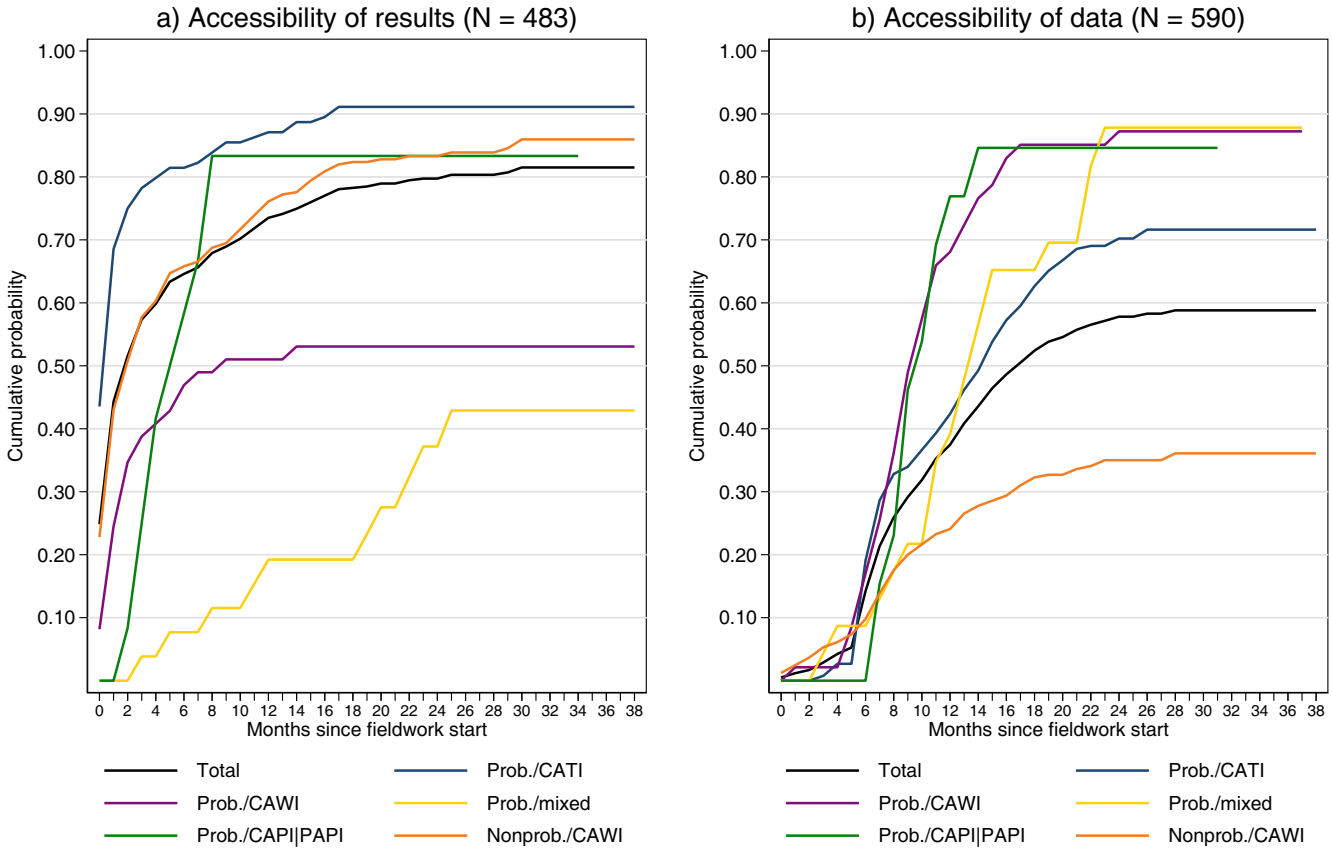

Fig. 1

Accessibility of results and data after fieldwork start (cumulative incidence function)

by a high demand for knowledge during an unprecedented crisis.

In contrast to the publication of results, the accessibility of survey data was comparatively slow and started at a lower level than publishing first results. About 6 months after the start of fieldwork, however, we observed a substantial increase in publication activity, which subsequently declined later. Overall, the data from 14% of all the surveys were released within 6 months after the start of fieldwork, while the proportion within 12 months after fieldwork start was 38%. It is noteworthy that only 64% of the surveys that started fieldwork between March 2020 and December 2021 had published their data by the time our SDCCP dataset was coded. Compared to the accessibility of first results (published by 81% of the surveys), this figure indicated that during the COVID-19 pandemic, survey data collection was more focused on generating results than collecting data for reuse and sharing. However, an analysis by survey design revealed that this trend was primarily driven by nonprobability web surveys, where only 36% published data. Among probability-based surveys, the rate of data publication ranged between 72% for telephone surveys and 88% for mixed-mode surveys.

Regarding our analysis of the development over time, the development of the number of surveys over time is particularly revealing (Fig. 2). The peak in the first months of our observed period suggests that a high number of surveys was fielded at the beginning of the COVID-19 pandemic in Germany. A possible explanation for this development was the high demand for data in the early phase of the crisis. As the analysis by survey design indicates, the initial surge in the number of fielded surveys was mainly driven by nonprobability web surveys. In contrast, the number of probability-based surveys remained relatively stable over the observed period.

Fig. 3 shows how the quality of the surveys developed over time. Panel (a) shows that the distribution of applied survey designs varied considerably across time between March 2020 and December 2021. On the one hand, the group of non-probability web surveys vastly dominated the survey landscape in the first months of the pandemic. This dominance may be related to the comparatively higher flexibility and speed of how such surveys can be implemented

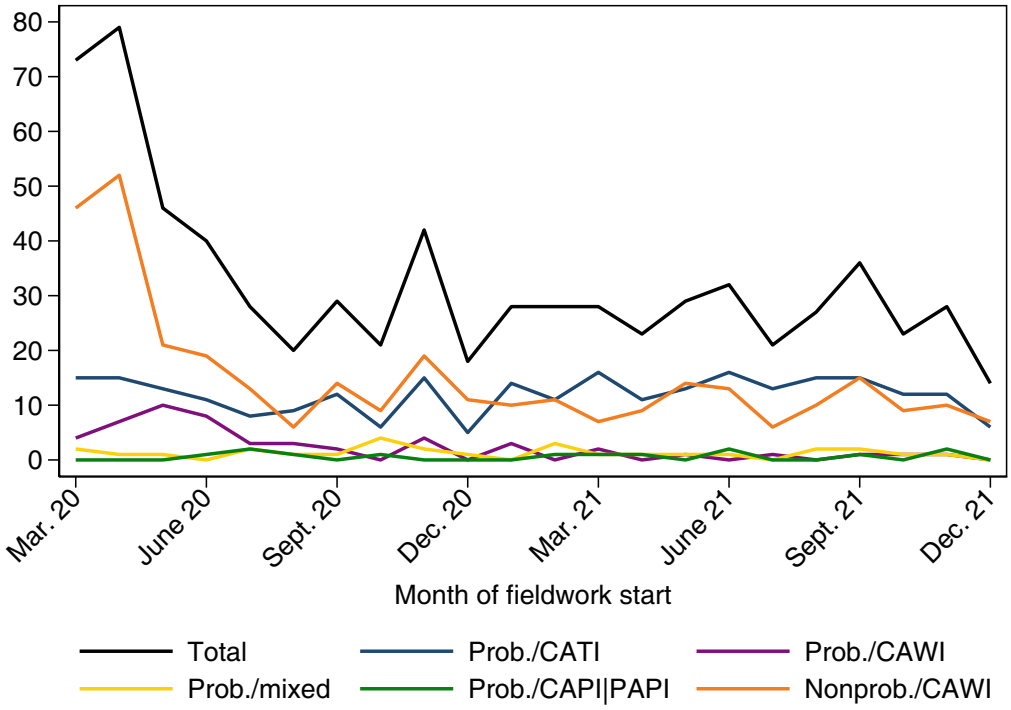

Fig. 2

Number of fieldwork starts over time, N = 686

and fielded. Furthermore, this group of designs also prevailed in December 2020 and 2021, indicating a possible seasonal effect. On the other hand, probability surveys gained in frequency over time and reached their highest prevalence in the first half of 2021. We suspect that the delayed emergence of the probability surveys is likely a result of the setup time required to field these kinds of surveys.

Panel (b) in Fig. 3 shows how the interpretability of survey data developed over time. Overall, our findings show that data collected early in the COVID-19 pandemic had, on average, a lower interpretability score compared to those collected later. More specifically, the interpretability score increased considerably from its lowest level in March 2020 to September 2020 and subsequently levelled off at a comparatively high level with some fluctuations. The slight inverted-U-shaped form that we observed might indicate that surveys collected at the end of the year also scored lower in interpretability, thus, providing less information for data users. In this regard, note that the low scores at the end of 2021 could merely reflect some variation or seasonal effects, so that the interpretation as a trend should be approached with some caution.

The lower panels in Fig. 3 showcase the development of accessibility regarding results (c) and data (d)[6]. The former shows a particularly high propensity for quickly publishing first results among surveys that started data collection shortly after the outbreak of the pandemic (i.e., March and April 2020), with around 50% of those publishing their results. Subsequently, the propensity to publish quickly decreased between May and October 2020. This pattern was consistent with the assumption that the onset of the COVID-19 pandemic and the related demand for rapid information influenced data producers to publish first results more quickly. After reaching a low point in November 2020, the propensity to publish results quickly increased again to a high level for surveys that started data collection in February to April 2021. This pattern could reflect a renewed increased need for information that came with the two successive pandemic infection waves in Germany, the so-called second wave (approximately October 2020 to February 2021) and third wave (approximately March 2021

[6] In both analyses, we examined the proportion of surveys for which first results or data have been published up to a certain point in time. We considered the second month after the start of fieldwork as the reference for analyzing the accessibility of first results, since the rapid availability of findings was particularly important during the COVID-19 pandemic. For the analysis of data accessibility, the 12th month served as the reference.

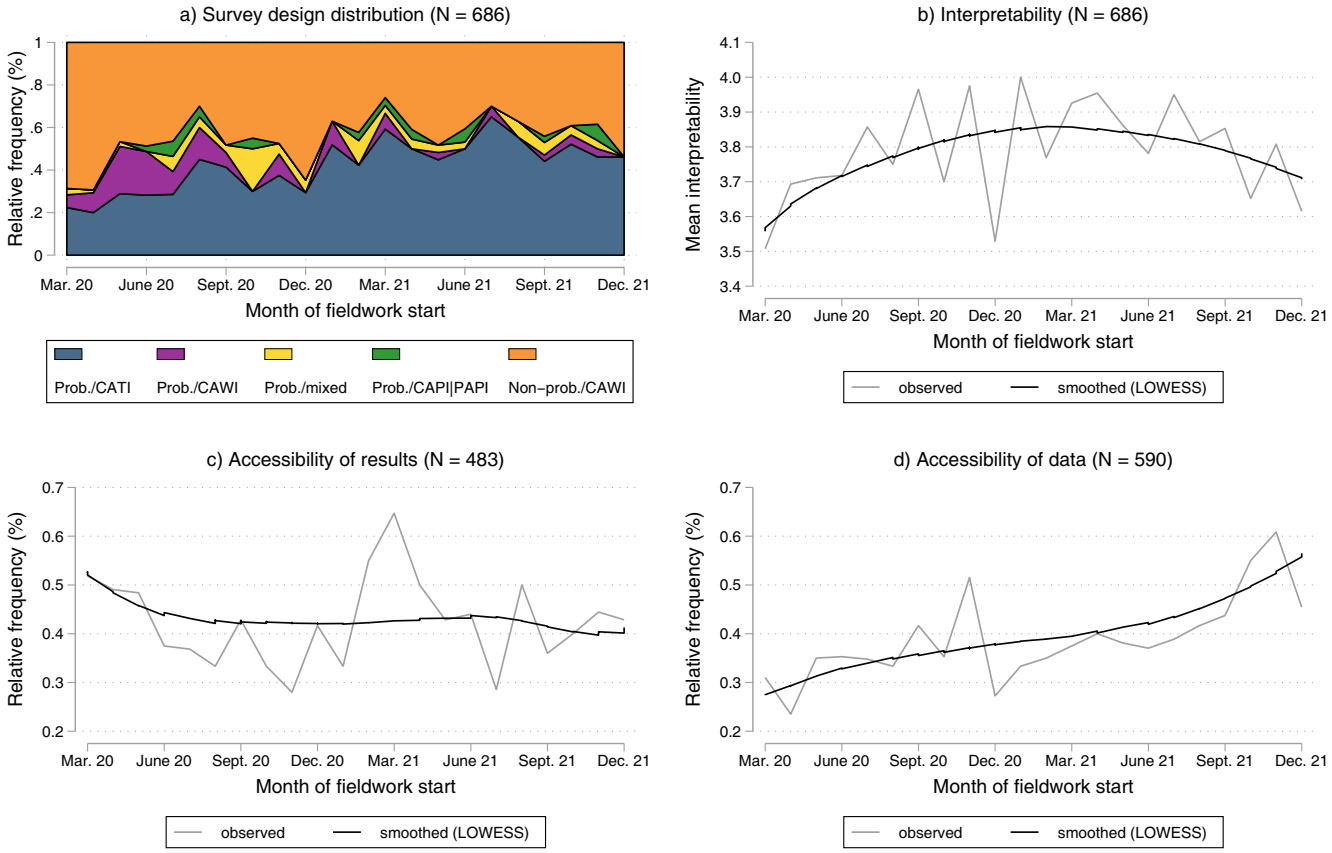

Fig. 3

Quality of social science surveys over time

to May 2021) (Schilling et al., 2022). Subsequently, the propensity to publish decreased towards the end of 2021.

Regarding the accessibility of survey data, Panel (d) in Fig. 3 shows a positive trend in the propensity to publish data as the pandemic progressed. Accordingly, surveys that began their field phase in March to August 2020 showed a comparatively low propensity to publish data within one year of data collection, with an overall minimum of 24% reached by the surveys fielded in April 2020. In contrast, the publication propensity of surveys that were fielded later was considerably higher in most months, with the proportion frequently exceeding the 40% mark from August 2021 onwards. Compared to our previous findings on the accessibility of first results, this finding indicates that surveys conducted early during the COVID-19 pandemic were devoted to gaining quick insights, while the later surveys were more dedicated to producing reusable data.

4.2 Associations Between Dimensions of Survey Quality

Table 1 details our results regarding the association between survey design and the user-centred quality dimensions (i.e., interpretability and accessibility). For this purpose, we used the interpretability of the survey data (Models 1 and 2), the accessibility of results (Models 3 and 4), and the accessibility of survey data (Models 5 and 6) as dependent variables. In addition to considering the time of data collection (month of the fieldwork start), we included the survey design in the models as an independent variable. Models 1, 3, and 5 served as null models assessing the effect of calendar time regarding the respective quality dimension.

Applying linear mixed models, we found the inverted-U-shaped relationship between fieldwork start and interpretability (Model 1), which already was unveiled in Fig. 3. Using event history models, we could replicate the non-linear wave-like relationship between fieldwork start and the accessibility of results (Model 3). Similarly, we were able to replicate the steadily increasing quadratic concave-down relationship between fieldwork start and the accessibility of

data (Model 5) (see Figure A1 in the appendix for a visualization of the predicted values conditional on calendar time for all the regression models).

Models 2, 4, and 6 provided additional information on the relationship between survey design and the other two data user-centred quality dimensions. With respect to interpretability (Model 2), we found that probability surveys were more likely to provide relevant information compared to non-probability surveys. These findings are in line with our expectation that collecting these more costly surveys is associated with more effort invested in making these data usable for third parties. Further, our results indicate that especially probability surveys that only relied on the web mode were able to provide thorough information. Since certain survey designs (e.g., those with non-probability sampling) are limited in the number of calculable outcome rates, we conducted a robustness check with a reduced interpretability index excluding the variable "any outcome rate" (see Table A2 in the appendix for detailed results). Although the regression coefficients became weaker in the model with the restricted index, the direction and conclusions were not affected. The only exception were probability mixed-mode surveys that no longer showed better interpretability than non-probability surveys in the restricted model.

Table 1

Effect of period and survey design on interpretability (Models 1 & 2), accessibility of results (Models 3 & 4), and accessibility of data (Models 5 & 6).

	Interpretability		Accessibility of results		Accessibility of data	
	1	2	3	4	5	6
Fixed effects (Survey level)						
(Fieldwork start month)	0.502	0.521	–0.015	–0.019	0.095***	0.097***
	(0.843)	(0.810)	(0.032)	(0.031)	(0.030)	(0.031)
(Fieldwork start month)^2	–0.000	–0.000	0.012**	0.011*	–0.002	–0.003
	(0.001)	(0.001)	(0.006)	(0.006)	(0.004)	(0.004)
(Fieldwork start month)^3	–	–	–0.001**	–0.001**	–	–
	–	–	(0.001)	(0.001)	–	–
Survey design (Ref. Non-probability)						
Probability & telephone	–	0.245*	–	0.241	–	0.633
	–	(0.145)	–	(0.384)	–	(0.488)
Probability & web	–	0.798**	–	–1.309*	–	–0.131
	–	(0.383)	–	(0.780)	–	(0.657)
Probability & mixed-mode	–	0.126	–	–2.003**	–	0.635
	–	(0.318)	–	(0.843)	–	(1.098)
Probability & personal or paper	–	0.189	–	–1.516**	–	2.190*
	–	(0.276)	–	(0.733)	–	(1.286)
Constant	–178.058	–185.913	–4.864***	–4.537***	–8.675**	–8.669**
	(308.172)	(296.251)	(0.508)	(0.479)	(3.403)	(3.809)
Random effects (Survey program level)						
Variance—constant	0.850	0.735	7.643***	6.775***	65.624	62.138
	(0.114)	(0.113)	(2.064)	(1.806)	(58.833)	(63.863)
Variance—residual	0.099	0.093	–	–	–	–
	(0.025)	(0.023)	–	–	–	–
AIC	851.194	806.325	1795.236	1792.978	2134.869	2134.369
BIC	873.848	847.103	1840.292	1863.781	2193.835	2222.818
N surveys	686	686	483	483	590	590
N observations	686	686	4613	4613	11,740	11,740

Models 1 & 2: b coefficients of linear mixed models. Models 3–6: mixed logit coefficients of time-discrete event history models. Cluster-robust standard errors in parentheses. Models 3–6 additionally account for process time (coefficients omitted)
***$p < 0.01$, **$p < 0.05$, *$p < 0.1$

In general, when comparing Model 1 (without survey design variables) and Model 2 (with survey design variables), we found that survey design explained parts of the variance in the interpretability scores. In other words, surveys with specific designs varied in how much relevant information they provided for data use. A Wald test highlighted the role of survey design as a contributing factor to the variation in the interpretability scores ($F(4,164) = 12.37$, $p = 0.0148$). The comparison of AIC and BIC information criteria confirmed the better predictive fit of Model 2 (with survey design variables) as compared to Model 1. However, the residual pseudo R^2 of 0.074 showed that a large proportion of variance remains unexplained by the survey-level indicators included in the model and indicated that other factors might be at play that were relevant for the data producers' ability or willingness to provide information on the data collection process.

In terms of accessibility, Model 4 showed that non-probability web surveys were associated with an increased propensity to publish results quickly compared to probability web and mixed-mode surveys. This finding is in line with our expectation that non-probability samples often are favoured for rapid result production, since they are easy to set up and field. However, with regard to publication propensity, no advantage was found compared to probability telephone surveys, which seem to provide a probability alternative with a potential for fast results delivery. Regarding the accessibility of data (Model 6), survey data had a higher probability of being published if they were collected based on a probability telephone or mixed-mode survey instead of a non-probability sample. However, with regard to publication propensity, no advantage was found compared to probability web surveys. In our view, these findings is a result of the purpose of costly probability surveys conducted (at least partially) in offline modes. Many of these large-scale efforts either require data publication due to funding policies or the data collection projects per se are aimed at collecting and sharing data as a service to the scientific community (e.g., General Social Surveys).

The finding that the development of results and data accessibility over time relates to the applied survey design was confirmed using a Wald test. Survey design proved to be a predictor variable for both outcomes: accessibility of results ($chi^2(4) = 15.13$, $p = 0.0044$) and accessibility of data ($chi^2(4) = 10.44$, $p = 0.0337$). As described above, our dataset included surveys with known publication of results or data, but unknown publication dates. These cases had to be excluded from the event history analysis of accessibility. As a robustness check of Models 4 and 6, we estimated mixed logit models to analyse the association between survey design and the general likelihood of publishing results and data (Table A3 in the Appendix). The conclusions regarding the association with survey design hold across models.

5 Discussion

We set out to answer two RQs to close a research gap on the quality of the social science survey data collected during the COVID-19 pandemic. Overall, our study showcases that the surveys conducted in Germany between March 2020 and December 2021 varied considerably in their survey design and consequently their quality.

Regarding RQ1 ("What was the quality of the social science surveys conducted during the COVID-19 pandemic?"), our study shows that the surveys strongly varied in their design and quality. Designs that the broader literature often (but not equivocally) relates to lower accuracy were used more frequently early in the pandemic compared to later. Interpretability was lower for surveys fielded early in the pandemic compared to those fielded later. Our findings for accessibility were ambiguous: surveys conducted early in the pandemic focused heavily on publishing first results quickly, but they were less focused on producing data to share. This relationship shifted, since surveys conducted later were more inclined to share data. Overall, the quality of the surveys conducted later during the pandemic can be considered higher than those of the surveys conducted early in the pandemic.

Regarding RQ2 ("How were the different dimensions of survey quality associated with each other during the COVID-19 pandemic?"), we found that trade-offs were made with respect to the three quality dimensions we investigated in our study. Early in the pandemic, researchers seemed to be focused heavily on quickly obtaining results and fielding surveys that were operatable under the restrictions imposed as a reaction to COVID-19. In consequence, non-probability web-based surveys were the most prominent survey design used during these times. These survey projects quickly published results but traded this benefit for a likely decrease in the accuracy of survey estimates, and they were less inclined to make the extra efforts of publishing their data for reuse. Later during the pandemic, more laborious probability surveys were fielded more frequently. The data obtained by these surveys were published more frequently and can be assumed to produce estimates of higher quality. However, these surveys achieved these beneficial properties at the cost of later fieldwork starts.

Our findings have several implications. First, the general body of findings on COVID-19 obtained from the data collected early during the pandemic is likely based on surveys that operated with designs that might have impaired accuracy. The lower accessibility of data for replication purposes of these surveys further worsens the situation. Therefore,

we recommend that researchers investigating COVID-19-related topics carefully choose the data of the highest available quality. For every research project, we further recommend screening the applied research design and reflecting on its impact on the examined research questions. Moreover, we strongly recommend conducting extensive robustness checks of findings published early in the pandemic using different existing datasets. Our study shows that data of good quality is usually available during all periods of the pandemic.

Second, our study showed that design and data quality strongly differed across surveys. In our view, it is reasonable to expect that substantive findings also will vary across surveys depending on their research design. To explore this issue further, we urge future research to investigate the differences in the substantive measures across the different surveys. For this purpose, the SDCCP dataset can serve as a source of information on survey design that can be linked with individual-level data.

Third, based on our results, we advocate for further advancing probability-based survey infrastructures that are quickly available to react to ad hoc demands for data collection (e.g., in term of crises). According to our data, only a relatively small number of probability-based data collection efforts were able to be fielded in the months following the outbreak of the pandemic as compared to the non-probability surveys. The majority of these probability data collections were existing longitudinal surveys that benefitted from the established infrastructure and did not have to require funding or go through the laborious sampling required for a probability survey sample. Examples of such studies in Germany are the German Family Panel (pairfam; Gummer et al., 2020), the Mannheim Corona Study based on the German Internet Panel (Blom et al., 2020), and the SOEP-CoV survey based on the Socio-Economic Panel (SOEP; Kühne et al., 2020). Examples from other countries include the UK Household Longitudinal Study (Burton et al., 2020) and the Panel Study of Income Dynamics (PSID) in the USA (Sastry et al., 2020).

Our study is not without limitations that pose opportunities for future research or warrant caution when interpreting our findings. First, we focused on surveys fielded in Germany. We made this decision to keep the context of all surveys similar. Countries will differ in how surveys can be and are operated (e.g., due to the availability of population registers to draw probability survey samples), the general survey climate of the country, and which restrictions were put in place as a reaction to the COVID-19 pandemic. We would welcome a cross-national comparison generalizing our research design to investigate differences in how survey practices during COVID-19 differed among countries. However, we would like to caution that such a research project will be demanding in terms of coding all the relevant information in different languages. In Germany alone, as part of the SDCCP project, 686 surveys had to be manually coded.

Second, in our study, we relied on the data available to the public via data and fieldwork documentations as well as project websites. We did not interview the data producers for their intentions or ask why they made certain trade-offs. Consequently, we were restricted to making assumptions on their motives or simply describing the status quo of data collection practices. In our view, detailing what was done during the pandemic is important by itself, and our findings also highlight that researchers seem to be driven by time-varying goals when conducting surveys. We see merit in investigating the individual researchers' reasons for design decisions in more detail, especially if they concern trade-offs that end in quickly publishing results of low accuracy. For this purpose, we would recommend surveying data producers and appending metadata datasets such as ours with information obtained from the data producers.

Third, our sample selection procedure relied heavily on searching archives and lists of surveys; consequently, studies that published data and results are likely to be overrepresented, and studies that simply collected data and produced unpublished results remain underrepresented. Nevertheless, 44% of our dataset did not publish results. Considering our sampling procedure, we argue that our study likely overestimates the quality of survey data during the COVID-19 pandemic, and we expect that the actual quality of all surveys conducted during this period is even lower.

Fourth, our study did not directly measure one of the key quality dimensions—accuracy. Although we provided a summary of the findings on the association between certain survey design features (survey mode and sampling procedure) and accuracy, we encourage additional analyses with an empirical assessment of accuracy. Based on our insights about the different surveys, this endeavour, however, will involve a fundamental challenge of harmonizing measures between surveys. Also, the number of similar constructs that should be harmonized and compared will likely reduce case numbers and limit statistical power for analyses. Moreover, we see merit in expanding this analysis and by differentiating the sampling and recruitment methods even further. In particular, non-probability surveys differ in their procedures how to recruit respondents (e.g., non-probability river sampling, non-probability online access panels, non-probability snowball sampling) and these differences might have important implications for data quality. For such a research project, however, more data points would be required, since convenience samples made up the majority (84%) of our data. To obtain more data points, we would recommend either expanding the observation period by additional years or including data from other countries.

Fifth, our study did not distinguish surveys by their substantive focus. Although we assumed all surveys in the dataset to be broadly pandemic-related (given the eligibility criteria for fieldwork period and survey type), demand for data collection and, thus, survey design and data quality, might differ depending on the substantive focus of surveys. Future research could examine how survey design and data quality differed between survey topics (e.g., political attitudes and behaviour, values, family surveys, health and wellbeing) across the pandemic. However, we would like to note that such fine-grained analyses likely would require statistical techniques to deal with limited case numbers (i.e., low statistical power).

Sixth, we set out to investigate how surveys were done during the COVID-19 pandemic and how this affected quality. We did not investigate how survey results affected policy decisions and whether some surveys had a higher impact than others on political actors or the public debate. Based on our findings that survey practices and most likely data quality changed throughout the pandemic, we see merit in investigating the data-driven decision-making processes during the COVID-19 pandemic.

Funding We gratefully acknowledge funding from the German Federal Ministry of Education and Research (BMBF) as part of the SDCCP project (grant number 01UP2100).

Data Availability Statement Data presented in this study are available in the Gesis archive: von Glasenapp, Karolina, Skora, Thomas, Gummer, Tobias, & Naumann, Elias (2024). SDCCP 1—Survey Design and Quality During the Covid-19 Pandemic. *GESIS, Köln. Datenfile Version 1.0.0,* https://doi.org/10.7802/2652.

Conflicting Interests The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

Adams-Prassl, A., Boneva, T., Golin, M., & Rauh, C. (2020). Inequality in the impact of the coronavirus shock: Evidence from real time surveys. *Journal of Public Economics*, *189*, 104245. https://doi.org/10.1016/j.jpubeco.2020.104245.

Arpino, B., Pasqualini, M., Bordone, V., & Solé-Auró, A. (2021). Older people's nonphysical contacts and depression during the COVID-19 Lockdown. *The Gerontologist*, *61*(2), 176–186. https://doi.org/10.1093/geront/gnaa144.

Bach, R.L., Cornesse, C., & Daikeler, J. (2024). Equipping the Offline population with Internet access in an Online panel: does it make a difference? *Journal of Survey Statistics and Methodology*, *12*(1), 80–93. https://doi.org/10.1093/jssam/smad003.

Betsch, C., Wieler, L.H., & Habersaat, K. (2020). Monitoring behavioural insights related to COVID-19. *The Lancet*, *395*(10232), 1255–1256. https://doi.org/10.1016/S0140-6736(20)30729-7.

Biemer, P.P. (2010). Total survey error: design, implementation, and evaluation. *Public Opinion Quarterly*, *74*(5), 817–848. https://doi.org/10.1093/poq/nfq058.

Biemer, P.P., & Lyberg, L.E. (2003). *Introduction to survey quality* (1st edn.). Wiley. https://doi.org/10.1002/0471458740.

Blom, A.G., Herzing, J.M.E., Cornesse, C., Sakshaug, J.W., Krieger, U., & Bossert, D. (2017). Does the recruitment of Offline households increase the sample representativeness of probability-based Online panels? Evidence from the German Internet panel. *Social Science Computer Review*, *35*(4), 498–520. https://doi.org/10.1177/0894439316651584.

Blom, A.G., Cornesse, C., Friedel, S., Krieger, U., Fikel, M., Rettig, T., Wenz, A., Juhl, S., Lehrer, R., Möhring, K., Naumann, E., & Reifenscheid, M. (2020). High frequency and high quality survey data collection: The Mannheim Corona Study. *Survey Research Methods*, *14*(2), 2. https://doi.org/10.18148/srm/2020.v14i2.7735.

Bradley, V.C., Kuriwaki, S., Isakov, M., Sejdinovic, D., Meng, X.-L., & Flaxman, S. (2021). Unrepresentative big surveys significantly overestimated US vaccine uptake. *Nature*, *600*(7890), 7890. https://doi.org/10.1038/s41586-021-04198-4.

Burton, J., Lynn, P., & Benzeval, M. (2020). How understanding society: the UK household longitudinal study adapted to the Covid-19 pandemic. *Survey Research Methods*. https://doi.org/10.18148/srm/2020.v14i2.7746.

Busemeyer, M.R. (2023). Financing the welfare state in times of extreme crisis: public support for health care spending during the Covid-19 pandemic in Germany. *Journal of European Public Policy*, *30*(1), 21–40. https://doi.org/10.1080/13501763.2021.1977375.

Cernat, A., & Revilla, M. (2021). Moving from face-to-face to a web panel: impacts on measurement quality. *Journal of Survey Statistics and Methodology*, *9*(4), 745–763. https://doi.org/10.1093/jssam/smaa007.

Cernat, A., Couper, M.P., & Ofstedal, M.B. (2016). Estimation of mode effects in the health and retirement study using measurement models. *Journal of Survey Statistics and Methodology*, *4*(4), 501–524. https://doi.org/10.1093/jssam/smw021.

Chan, P., & Ambrose, D. (2011). *Canadian online panels: Similar or different? Vue, 16–20*. https://www.websm.org/db/12/15980/Web%20Survey%20Bibliography/Canadian_online_panels_Similar_or_different/

Chang, L., & Krosnick, J.A. (2009). National surveys via Rdd telephone interviewing versus the Internet: comparing sample representativeness and response quality. *Public Opinion Quarterly*, *73*(4), 641–678. https://doi.org/10.1093/poq/nfp075.

Cornesse, C., & Bosnjak, M. (2018). Is there an association between survey characteristics and representativeness? A meta-analysis. *Survey Research Methods*, *12*(1), 1. https://doi.org/10.18148/srm/2018.v12i1.7205.

Cornesse, C., & Schaurer, I. (2021). The long-term impact of different Offline population inclusion strategies in probability-based Online panels: evidence from the German Internet panel and the GESIS panel. *Social Science Computer Review*, *39*(4), 687–704. https://doi.org/10.1177/0894439320984131.

Cornesse, C., Blom, A.G., Dutwin, D., Krosnick, J.A., De Leeuw, E.D., Legleye, S., Pasek, J., Pennay, D., Phillips, B., Sakshaug, J.W., Struminskaya, B., & 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. https://doi.org/10.1093/jssam/smz041.

Dodou, D., & de Winter, J.C.F. (2014). Social desirability is the same in offline, online, and paper surveys: a meta-analysis. *Computers in Human Behavior*, *36*, 487–495. https://doi.org/10.1016/j.chb.2014.04.005.

Eder, C., & Jedinger, A. (2019). FAIR national election studies: how well are we doing? *European Political Science*, *18*(4), 651–668. https://doi.org/10.1057/s41304-018-0194-3.

ESS ERIC Core Scientific Team (2022). *Quality report for the European social survey, round 9*. https://www.europeansocialsurvey.org/docs/round9/methods/ESS9_Quality_Report.pdf

European Social Survey European Research Infrastructure (ESS ERIC). (2023). *ESS Round 10: European Social Survey Round 10 Data (2020). Data file edition 3.2. Sikt—Norwegian Agency for Shared Services in Education and Research, Norway—Data Archive and distributor of ESS data for ESS ERIC* [Dataset]. https://doi.org/10.21338/ess10e03_2

Felderer, B., Kirchner, A., & Kreuter, F. (2019). The effect of survey mode on data quality: disentangling nonresponse and measurement error bias. *Journal of Official Statistics*, *35*(1), 93–115. https://doi.org/10.2478/jos-2019-0005.

Fetzer, T., Hensel, L., Hermle, J., & Roth, C. (2021). Coronavirus perceptions and economic anxiety. *The Review of Economics and Statistics*, *103*(5), 968–978. https://doi.org/10.1162/rest_a_00946.

Forti, L.R., Solino, L.A., & Szabo, J.K. (2021). Trade-off between urgency and reduced editorial capacity affect publication speed in ecological and medical journals during 2020. *Humanities and Social Sciences Communications*, *8*(1), 1–9. https://doi.org/10.1057/s41599-021-00920-9.

von Glasenapp, K., Skora, T., Gummer, T., & Naumann, E. (2024). *SDCCP 1—survey design and data quality during the Covid-19 pandemic [Dataset]*. https://doi.org/10.7802/2652.

Groves, R.M., & Lyberg, L. (2010). Total survey error: past, present, and future. *Public Opinion Quarterly*, *74*(5), 849–879. https://doi.org/10.1093/poq/nfq065.

Groves, R.M., Fowler, F.J. Jr., Couper, M.P., Lepkowski, J.M., Singer, E., & Tourangeau, R. (2009). *Survey methodology* (2nd edn.). Wiley.

Gummer, T., Schmiedeberg, C., Bujard, M., Christmann, P., Hank, K., Kunz, T., Lück, D., & Neyer, F.J. (2020). The impact of Covid-19 on fieldwork efforts and planning in pairfam and FReDA-GGS. *Survey Research Methods*, *14*(2), 2. https://doi.org/10.18148/srm/2020.v14i2.7740.

Gummer, T., Skora, T., von Glasenapp, K., & Naumann, E. (2024). A dataset on survey designs and quality of social and behavioral science surveys during the COVID-19 pandemic. *Scientific Data*, *11*(1), 619. https://doi.org/10.1038/s41597-024-03475-x.

Hope, S., Campanelli, P., Nicolaas, G., Lynn, P., & Jäckle, A. (2022). The role of the interviewer in producing mode effects: results from a mixed modes experiment comparing face-to-face, telephone and web administration. *Survey Research Methods*, *16*(2), 2. https://doi.org/10.18148/srm/2022.v16i2.7771.

Horbach, S.P.J.M. (2020). Pandemic publishing: medical journals strongly speed up their publication process for COVID-19. *Quantitative Science Studies*, *1*(3), 1056–1067. https://doi.org/10.1162/qss_a_00076.

Huebener, M., Waights, S., Spiess, C.K., Siegel, N.A., & Wagner, G.G. (2021). Parental well-being in times of Covid-19 in Germany. *Review of Economics of the Household*, *19*(1), 91–122. https://doi.org/10.1007/s11150-020-09529-4.

Jabkowski, P. (2023). Increase in the quality of methodological documentation of cross-national pan-European multi-wave surveys over the last 40 years—a research note. *International Journal of Social Research Methodology*, *26*(6), 817–824. https://doi.org/10.1080/13645579.2022.2097394.

Jabkowski, P., & Kołczyńska, M. (2020). Sampling and fieldwork practices in Europe: analysis of methodological documentation from 1,537 surveys in five cross-national projects, 1981–2017. *Methodology*, *16*(3), 3. https://doi.org/10.5964/meth.2795.

Jedinger, A., Watteler, O., & Förster, A. (2018). Improving the quality of survey data documentation: a total survey error perspective. *Data*, *3*(4), 4. https://doi.org/10.3390/data3040045.

Kapteyn, A., Angrisani, M., Bennett, D., de Bruin, W.B., Darling, J., Gutsche, T., Liu, Y., Meijer, E., Perez-Arce, F., Schaner, S., Thomas, K., & Weerman, B. (2020). Tracking the effect of the Covid-19 pandemic on the lives of American households. *Survey Research Methods*. https://doi.org/10.18148/srm/2020.v14i2.7737.

Kennedy, C., Mercer, A., Keeter, S., Hatley, N., McGeeney, K., & Gimenez, A. (2016). *Evaluating Online nonprobability surveys*. Pew Research Center. https://assets.pewresearch.org/wp-content/uploads/sites/12/2016/04/Nonprobability-report-May-2016-FINAL.pdf

Kennedy, C., Blumenthal, M., Clement, S., Clinton, J.D., Durand, C., Franklin, C., McGeeney, K., Miringoff, L., Olson, K., Rivers, D., Saad, L., Witt, G.E., & Wlezien, C. (2018). An evaluation of the 2016 election polls in the United States. *Public Opinion Quarterly*, *82*(1), 1–33. https://doi.org/10.1093/poq/nfx047.

Kohler, U. (2020). Survey research methods during the Covid-19 crisis. *Survey Research Methods*, *14*(2), 2. https://doi.org/10.18148/srm/2020.v14i2.7769.

Kołczyńska, M., & Schoene, M. (2018). Survey data harmonization and the quality of data documentation in cross-national surveys. In *Advances in comparative survey methods* (pp. 963–984). Wiley. https://doi.org/10.1002/9781118884997.ch44.

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

Krosnick, J.A., Narayan, S., & Smith, W.R. (1996). Satisficing in surveys: Initial evidence. *New Directions for Evaluation*, *1996*(70), 29–44. https://doi.org/10.1002/ev.1033.

Krumpal, I. (2013). Determinants of social desirability bias in sensitive surveys: A literature review. *Quality & Quantity*, *47*(4), 2025–2047. https://doi.org/10.1007/s11135-011-9640-9.

Kühne, S., Kroh, M., Liebig, S., & Zinn, S. (2020). The need for household panel surveys in times of crisis: the case of SOEP-coV. In *Survey Research Methods* (pp. 195–203). https://doi.org/10.18148/SRM/2020.V14I2.7748.

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

Loosveldt, G., & Sonck, N. (2008). An evaluation of the weighting procedures for an online access panel survey. *Survey Research Methods*, *2*(2), 2. https://doi.org/10.18148/srm/2008.v2i2.82.

Lyberg, L.E., & Weisberg, H.F. (2016). Total survey error: a paradigm for survey methodology. In C. Wolf, D. Joye, T.W. Smith & Y. Fu (Eds.), *The SAGE handbook of survey methodology*. SAGE.

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

Malhotra, N., & Krosnick, J.A. (2007). The effect of survey mode and sampling on inferences about political attitudes and behavior: comparing the 2000 and 2004 ANES to Internet surveys with Nonprobability samples. *Political Analysis*, *15*(3), 286–323. https://doi.org/10.1093/pan/mpm003.

Moretti, A., & Santi, C. (2020). *The need for reliable and timely data to contrast COVID-19: what went wrong?* SSRN Scholarly Paper No. 3633827. https://doi.org/10.2139/ssrn.3633827.

Munzert, S., Selb, P., Gohdes, A., Stoetzer, L.F., & Lowe, W. (2021). Tracking and promoting the usage of a COVID-19 contact tracing app. *Nature Human Behaviour*, *5*(2), 2. https://doi.org/10.1038/s41562-020-01044-x.

Nguyen, K.H., Lu, P.-J., Meador, S., Hung, M.-C., Kahn, K., Hoehner, J., Razzaghi, H., Black, C., & Singleton, J.A. (2023). *Comparison of COVID-19 vaccination coverage estimates from the Household Pulse Survey, Omnibus Panel Surveys, and COVID-19 vaccine administration data, United States, March 2021*. https://www.cdc.gov/vaccines/imz-managers/coverage/adultvaxview/pubs-resources/covid19-coverage-estimates-comparison.html

Nind, M., Coverdale, A., & Meckin, R. (2021). *Changing Social Research Practices in the Context of Covid-19: Rapid Evidence Review*. Working Paper. NCRM. https://doi.org/10.5258/NCRM/NCRM.00004458.

Nind, M., Coverdale, A., & Meckin, R. (2022). *Changing social research practices in the context of Covid-19: updated rapid evidence review—synthesis of the 2021 literature*. Working Paper. National Centre for Research Methods. https://eprints.ncrm.ac.uk/id/eprint/4602/

Pforr, K., & Dannwolf, T. (2017). What do we lose with Online-only surveys? Estimating the bias in selected political variables due to Online mode restriction. *Statistics, Politics and Policy*, *8*(1), 105–120. https://doi.org/10.1515/spp-2016-0004.

Phillips, D.L., & Clancy, K.J. (1972). Some effects of 'social desirability' in survey studies. *American Journal of Sociology*. https://doi.org/10.1086/225231.

Pierce, M., McManus, S., Jessop, C., John, A., Hotopf, M., Ford, T., Hatch, S., Wessely, S., & Abel, K.M. (2020). Says who? The significance of sampling in mental health surveys during COVID-19. *The Lancet. Psychiatry*, *7*(7), 567–568. https://doi.org/10.1016/S2215-0366(20)30237-6.

Revilla, M., Cornilleau, A., Cousteaux, A.-S., Legleye, S., & de Pedraza, P. (2016). What is the gain in a probability-based Online panel of providing Internet access to sampling units who previously had no access? *Social Science Computer Review*, *34*(4), 479–496. https://doi.org/10.1177/0894439315590206.

Rohr, B., Silber, H., & Felderer, B. (2024). Comparing the accuracy of Univariate, bivariate, and multivariate estimates across probability and nonprobability surveys with population benchmarks. *Sociological Methodology*. https://doi.org/10.1177/00811750241280963.

Sakshaug, J.W., Beste, J., Coban, M., Fendel, T., Haas, G.-C., Hülle, S., Kosyakova, Y., König, C., Kreuter, F., Küfner, B., Müller, B., Osiander, C., Schwanhäuser, S., Stephan, G., Vallizadeh, E., Volkert, M., Wenzig, C., Westermeier, C., Zabel, C., & Zins, S. (2020). Impacts of the COVID-19 pandemic on labor market surveys at the German institute for employment research. *Survey Research Methods*, *14*(2), 2. https://doi.org/10.18148/srm/2020.v14i2.7743.

Sastry, N., McGonagle, K., & Fomby, P. (2020). Effects of the Covid-19 crisis on survey fieldwork: experience and lessons from two major supplements to the U.S. Panel study of income dynamics. *Survey Research Methods*, *14*(2), 241–245. https://doi.org/10.18148/srm/2020.v14i2.7752.

Scherpenzeel, A.C., & Bethlehem, J.G. (2011). How representative are online panels? Problems of coverage and selection and possible solutions. In *Social and behavioral research and the Internet: Advances in applied methods and research strategies* (pp. 105–132). Routledge/Taylor & Francis.

Scherpenzeel, A., Axt, K., Bergmann, M., Douhou, S., Oepen, A., Sand, G., Schuller, K., Stuck, S., Wagner, M., & Börsch-Supan, A. (2020). Collecting survey data among the 50+ population during the COVID-19 outbreak: the survey of health, ageing and retirement in europe (SHARE). *Survey Research Methods*. https://doi.org/10.18148/srm/2020.v14i2.7738.

Schilling, J., Buda, S., & Tolksdorf, K. (2022). *Zweite Aktualisierung der „Retrospektiven Phaseneinteilung der COVID-19-Pandemie in Deutschland"*. https://doi.org/10.25646/9787.

Singer, J.D., & Willett, J.B. (2003). *Applied longitudinal data analysis: modeling change and event occurrence*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195152968.001.0001.

Sohlberg, J., Gilljam, M., & Martinsson, J. (2017). Determinants of polling accuracy: the effect of opt-in Internet surveys. *Journal of Elections, Public Opinion and Parties*, *27*(4), 433–447. https://doi.org/10.1080/17457289.2017.1300588.

Statistisches Bundesamt (Destatis) (2021). *Einkommens- und Verbrauchsstichprobe EVS 2018 Qualitätsbericht*

Stefkovics, Á., Eichhorst, A., Skinnion, D., & Harrison, C.H. (2024). Are we becoming more transparent? Survey reporting trends in top journals of social sciences. *International Journal of Public Opinion Research*, *36*(2), edae13. https://doi.org/10.1093/ijpor/edae013.

Steinmetz, S., Bianchi, A., Tijdens, K., & Biffignandi, S. (2014). Improving web survey quality. In M. Callegaro, R. Baker, J. Bethlehem, A.S. Göritz, J.A. Krosnick & P.J. Lavrakas (Eds.), *Online panel research: data quality perspective* (pp. 273–298). Wiley. https://doi.org/10.1002/9781118763520.ch12.

Sturgis, P., Kuha, J., Baker, N., Callegaro, M., Fisher, S., Green, J., Jennings, W., Lauderdale, B.E., & Smith, P. (2018). An assessment of the causes of the errors in the 2015 UK General election opinion polls. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *181*(3), 757–781. https://doi.org/10.1111/rssa.12329.

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

Tourangeau, R., Conrad, F.G., & Couper, M.P. (2013). *The science of web surveys*. Oxford: University Press. https://doi.org/10.1093/acprof:oso/9780199747047.001.0001.

Wang, W., Rothschild, D., Goel, S., & Gelman, A. (2015). Forecasting elections with non-representative polls. *International Journal of Forecasting*, *31*(3), 980–991. https://doi.org/10.1016/j.ijforecast.2014.06.001.

Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L.B., Bourne, P.E., Bouwman, J., Brookes, A.J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C.T., Finkers, R., & Mons, B. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, *3*(1), 1. https://doi.org/10.1038/sdata.2016.18.

Yamada, Y., Ćepulić, D.-B., Coll-Martín, T., Debove, S., Gautreau, G., Han, H., Rasmussen, J., Tran, T. P., Travaglino, G. A., & Lieberoth, A. (2021). COVIDiSTRESS Global survey dataset on psychological and behavioural consequences of the Covid-19 outbreak. *Scientific Data*, *8*(1), 1. https://doi.org/10.1038/s41597-020-00784-9.

Yeager, D. S., Krosnick, J. A., Chang, L., Javitz, H. S., Levendusky, M. S., Simpser, A., & Wang, R. (2011). Comparing the accuracy of RDD telephone surveys and Internet surveys conducted with probability and non-probability samples. *Public Opinion Quarterly*, *75*(4), 709–747. https://doi.org/10.1093/poq/nfr020.