Linking Surveys and Digital Trace Data: Insights From two Studies on Determinants of Data Sharing Behaviour

Silber, Henning; Breuer, Johannes; Beuthner, Christoph; Gummer, Tobias; Keusch, Florian; Siegers, Pascal; Stier, Sebastian; Weiß, Bernd

Gefördert durch die Deutsche Forschungsgemeinschaft (DFG) - Projektnummer 491156185 / Funded by the German Research Foundation (DFG) - Project number 491156185

Empfohlene Zitierung / Suggested Citation:
Silber, H., Breuer, J., Beuthner, C., Gummer, T., Keusch, F., Siegers, P., ... Weiß, B. (2022). Linking Surveys and Digital Trace Data: Insights From two Studies on Determinants of Data Sharing Behaviour. *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, *185*(Suppl. 2), 387-407. https://doi.org/10.1111/rssa.12954

Received: 2 March 2021 | Accepted: 17 August 2022

DOI: 10.1111/rssa.12954

ORIGINAL ARTICLE

ROYAL STATISTICAL SOCIETY | Series A Statistics in Society

Linking surveys and digital trace data: Insights from two studies on determinants of data sharing behaviour

Henning Silber[1] | **Johannes Breuer**[2,3] | **Christoph Beuthner**[1] | **Tobias Gummer**[1,4] | **Florian Keusch**[4] | **Pascal Siegers**[2] | **Sebastian Stier**[2] | **Bernd Weiß**[1]

[1]GESIS – Leibniz Institute for the Social Sciences, Mannheim, Germany

[2]GESIS - Leibniz Institute for the Social Sciences, Cologne, Germany

[3]Center for Advanced Internet Studies (CAIS), Bochum, Germany

[4]University of Mannheim, Mannheim, Germany

Correspondence
Henning Silber, GESIS – Leibniz Institute for the Social Sciences, Mannheim, Germany.
Email: henning.silber@gesis.org

Abstract

Combining surveys and digital trace data can enhance the analytic potential of both data types. We present two studies that examine factors influencing data sharing behaviour of survey respondents for different types of digital trace data: Facebook, Twitter, Spotify and health app data. Across those data types, we compared the relative impact of four factors on data sharing: data sharing method, respondent characteristics, sample composition and incentives. The results show that data sharing rates differ substantially across data types. Two particularly important factors predicting data sharing behaviour are the incentive size and data sharing method, which are both directly related to task difficulty and respondent burden. In sum, the paper reveals systematic variation in the willingness to share additional data which need to be considered in research designs linking surveys and digital traces.

KEYWORDS

consent, data donation, data linkage, data sharing rates, incentives, social network sites

1 | INTRODUCTION

The widespread use of digital devices and online platforms produces vast amounts of data. These data are often subsumed under the category of digital trace data and can be a rich source of information for the social and behavioural sciences (King, 2011; Ledford, 2020; Shlomo & Goldstein, 2015). For social scientists, digital traces are especially interesting when they are available on the individual level and can be linked with data on person-level variables, such as attitudes, values, personality and personal characteristics.

A fruitful approach for increasing the analytic potential of digital trace data is to combine them with data from surveys so that the two data sources can complement and enrich one another (Al Baghal et al., 2020; Amaya et al., 2021; DiGrazia et al., 2013; Harari et al., 2017; Stier et al., 2020). While researchers using surveys usually rely on self-reported behaviour, digital trace data allow researchers to record and track many different types of behaviour over time and with high granularity. On the other hand, surveys allow getting more profound insights into personal belief systems, which helps to understand behaviour and investigate causal relationships (Stier et al., 2020). While the combination of those two data types holds great promise for the social sciences, the practicalities of linking surveys and digital trace data and their implications have not yet been systematically studied. Understanding the different ways in which surveys and digital trace data can be linked is important to assess the nature and quality of the resulting combined data. The options for linking surveys and digital trace data and their outcomes depend on a variety of factors, including technological developments and privacy considerations (Beuthner et al., 2021; Boeschoten et al., 2020; Nissenbaum, 2018; Oberski & Kreuter, 2020). A key challenge, hence, is choosing the appropriate linking option(s) for answering specific research questions.

Against this background, the paper presents results from two studies exploring different ways of combining surveys and digital trace data. The main objective of both studies was to gain insights into the data sharing process of survey respondents when additional digital trace data are requested within a survey interview. In Study 1, we explored the combination of surveys with different types of online platform data. In Study 2, we experimentally investigated the combination of a survey with social media and health app data. By identifying which factors determine survey participants' data sharing decisions in studies that link surveys and digital trace data, our aim is to derive recommendations and best practice advice for future research. Specifically, this research explores how five factors affect data sharing decisions: (RQ1) the data type, (RQ2) the data sharing method, (RQ3) respondent characteristics, (RQ4) incentives and (RQ5) the sampling method.

2 | LINKING SURVEY DATA AND DIGITAL TRACE DATA

A straightforward way to link survey and digital trace data is to ask respondents within a survey whether they are willing to share additional data (Kreuter et al., 2020; Sakshaug, 2020; Sloan et al., 2020). However, such questions can constitute an additional response burden (Eckman & Haas, 2017), leading to high item non-response or even to break-off.

Since a data sharing request introduces a new demand into the survey interview, respondents must evaluate whether they approve the request. As for any behavioural choice, cost and benefit considerations are likely to guide the decision-making process (Biner & Kidd, 1994;

Dutwin et al., 2015; Esser, 1986; Leeper, 2019; Porter & Whitcomb, 2003). Benefits of an affirmative answer can include congruence with conversational norms, time savings and possible incentives. First, conversation norms of human interactions suggest that it is perceived as impolite to say no and decline a request. Second, time savings may include a shorter questionnaire because information that would have otherwise been collected through self-reports can be derived from digital traces. Third, incentives are often monetary stimuli that will be received if a person agrees to the request for data. Costs of an affirmative answer can include the effort necessary to share the data and the data's potential sensitivity.

Additional considerations that may guide the decision process are attitudes and norms towards privacy, data sharing and scientific research (Keusch et al., 2020; Oberski & Kreuter, 2020; Sloan et al., 2020). In a situation where costs are high, for example, if a person is asked to go through multiple complicated steps to share their data or if the requested data are especially sensitive, attitudes and norms are unlikely to strongly affect the decision-making process (Best & Kroneberg, 2012; Riker & Ordeshook, 1968; Stern, 1992). However, in the context of data sharing requests where the costs are low, attitudes regarding privacy or conversational norms are more likely to influence the decision-making process. For example, in a low-cost situation people with fewer privacy concerns might be more likely to share their data and conversational norms may guide persons to answer the data sharing request affirmatively to avoid creating disagreement by selecting 'no'.

Responses to data sharing requests within a survey can be considered through different theoretical perspectives, such as the social exchange theory (e.g., Dillman, 1978), the leverage salience theory (e.g., Groves et al., 2000) and the theory of contextual integrity (e.g., Nissenbaum, 2018). The social exchange theory describes survey interviews as transactions between the respondent and the interviewer (or the researcher in case of a self-administered survey mode) and emphasises the social component and role of social trust in the data sharing process. The leverage-salience theory emphasises the situational impact of the various survey characteristics (e.g., interviewer, incentive, survey topic, mode and length) and, thus, puts emphasis on both economic and social components. The situational impact of those survey characteristics can be very different, depending on a respondent's personal characteristics (e.g., age, gender, education or privacy attitudes). Finally, the theory of contextual integrity suggests that contextual informational norms define which data flows are appropriate and which are not. Situational parameters include the data type, the involved actors—such as the data sender and recipient—and the transmission principles, that is, the 'rules' under which the data are transferred. For example, people might find it acceptable to provide Twitter data when they are asked to share their Twitter handle for research purposes but not when they are asked to download and upload their data.

Regarding the practicalities of sharing digital trace data, typically, three conditions must be met before respondents share their data (Elevelt et al., 2019; Keusch et al., 2019; Revilla et al., 2018, 2019). First, only users of the platforms, services or devices that generate the data can be asked to share them (*usage*). For instance, for social media data, usage rates can vary dramatically between age groups, occupations, platforms and countries (e.g., among the German online population: 44% use Facebook, 29% use Instagram and 12% use Twitter, see Newman et al., 2021). Second, researchers have to obtain *informed consent* from participants to use their data. Third, users who have given informed consent have to successfully complete the data sharing procedure to share their data (*data sharing behaviour*). This is, for instance, necessary when respondents have to download and instal an app or a browser plug-in or are asked to export and share data from the platforms or devices under study.

While there are some previous studies that compared willingness to share additional data in surveys across different *data types* (Jenkins et al., 2006; Revilla et al., 2019; Wenz et al., 2019), it remains largely an open question which factors determine the sharing of digital trace data and which similarities and differences in sharing behaviour exist across platforms/data types. Specifically, we identified five factors that likely affect survey respondents' willingness to share digital trace data (see Column 1 in Table 1 for an overview). There are likely base rate differences between platforms regarding their users' willingness to share digital trace data because the platforms are used for different purposes and the generated data can be more or less sensitive. For example, Twitter is largely viewed as a platform for public communication, whereas Facebook is usually used for and regarded as a platform for personal and private communication.

Firstly, the way in which people are asked to share their digital trace data (*data sharing method*) is likely to influence the sharing decision (Boeschoten et al., 2020; Settanni et al., 2018). There are various ways of collecting digital trace data that differ, among other things, in the type of data they generate as well as in the amount of effort the data sharing requires (Breuer et al., 2020). Many of these data collection options can also be used in studies where researchers partner with users to access digital trace data (Halavais, 2019). For example, participants can be asked to simply share their Twitter handle, allowing researchers to collect their Twitter data through the platform's API. Or researchers can ask participants to use an app or browser-plugin that records (parts of) their digital traces (de Haan & Hendriks, 2013; Haim & Nienierza, 2019; Kosinski et al., 2013). Another option is to ask participants to export (parts of) their digital trace data themselves, which is a functionality that most platforms and services offer, and then share these 'data download packages' (Boeschoten et al., 2020) with the researchers, for example, by uploading them through a web tool. This relatively new approach is often called 'data donation' due to the active role of the respondents within the data sharing process (see, e.g., Araujo et al., 2021). Notably, those different data sharing procedures vary with regard to task difficulty and respondent burden, which translates into perceived costs. Specifically, more active data sharing procedures, such as data donation, are usually more burdensome for respondents than passive ones, such as providing consent and a username (Araujo et al., 2021; Keusch et al., 2019). At the same time, more active data sharing procedures provide more control, which can increase the willingness to share data (Juga et al., 2021).

It is uncertain in which situations attitudes and norms (*respondent characteristics*) regarding privacy, data sharing or science guide data sharing behaviour (Keusch et al., 2019). Using the distinction between low-cost and high-cost situations introduced above (Best & Kroneberg, 2012), it is an open question under which circumstances a data sharing situation is considered a low-cost or high-cost situation, and which attitudes and norms are more likely to affect decision-making in those scenarios. In addition, socio-demographic attributes of respondents may also be relevant for data sharing decisions. For example, older people are typically less knowledgeable when it comes to digital media (Kuru et al., 2017; Smith & Page, 2015) which might be especially important for data sharing methods that require a substantial amount of effort and/or technical expertise from the participants. Other respondent characteristics that have been shown to influence data sharing rates are education (higher percentages for more educated participants: Elevelt et al., 2019; Revilla et al., 2021), gender (higher percentages for male participants: Keusch et al., 2019; Revilla et al., 2021), income (mixed effects: Revilla et al., 2021), attitudes towards surveys (higher percentages with more positive attitudes: Keusch et al., 2019), device usage (higher percentages when more usage: Elevelt et al., 2019;Jäckle et al., 2019; Keusch et al., 2019), and privacy concerns (higher percentages when less concerns: Jäckle et al., 2019; Keusch et al., 2019).

TABLE 1 Conceptual design of factors influencing data sharing for the two studies

	Study 1			
Factor	**Survey 1**	**Survey 2**	**Study 2**	**Comparison type**
(1) Data type	Twitter	Facebook, Spotify	Twitter, Health App	Observational
(2) Sharing method	API	Browser plug-in, app	API, data donation[a]	Experimental and observational
(3) Respondent characteristics and device	Socio-demographics, privacy, network usage, device	Socio-demographics, survey evaluation, network usage, device	Socio-demographics, personality, survey evaluation, survey attitudes, privacy, technology, network/app usage, device[b]	Quasi-experimental and observational
(4) Incentives	5€ (pre- vs. post- paid)[c]	5€, 2.5€	0€, 2.5€, 5€, 10€[d]	Experimental and observational
(5) Sampling method	Web tracking sample	Web tracking sample	Online panel, only Twitter/device users	Observational

[a] In Study 2, respondents in the Twitter data group, who agreed to share were data, were randomly assigned to the data sharing method (API or data donation).
[b] In Study 2, respondents using an iPhone or Samsung smartphone received a device-specific health app data sharing request. Since the group assignment was not random, the comparison is classified as quasi-experimental.
[c] In Study 1, respondents were randomly assigned to receive the 5€ incentive either pre- or post-paid.
[d] In Study 2, respondents were randomly assigned to one of four incentive conditions (0€, 2.5€, 5€ or 10€). This experiment was implemented for both data types.

Although there is a large body of experimental literature on the types of *monetary incentives* optimal to increase survey participation (e.g., Göritz, 2006; Singer, 2018), much less is known about how incentives influence the likelihood of data sharing within a survey. Specifically, while it is evident that incentives can increase the data sharing probability, it is not clear which amounts may be ideal (Jäckle et al., 2019; Keusch et al., 2019) and whether they should be offered as pre- or post-paid incentives. The literature on incentives regarding survey participation would suggest that the effect of incentives is non-linear with a certain point of saturation and that pre-paid incentives work better than post-paid incentives (Singer, 2018). However, since respondents have already agreed to participate in the survey, the situation and the request are different in the case of sharing digital trace data. Thus, investigating the generalisability of previous findings on incentives for data sharing decisions is relevant from a cost perspective since researchers typically do not have unlimited funds and would like to implement an efficient incentive strategy. Moreover, investigating the use of incentives for data sharing decisions is also relevant from another practical perspective. If an unusually high incentive is offered, respondents might even be less likely to share their data because they might regard the data as very valuable or suspect a possibly harmful use.

Data sharing behaviour may also vary depending on the *sampling method and sample composition* (Brosnan et al., 2017; Elevelt et al., 2019; Jäckle et al., 2019; Keusch et al., 2019). For example, respondents from a cross-sectional, general population sample might be less likely to share additional data than respondents from special populations, such as participants of commercial online access panels, who might be more familiar with requests to share digital content.

3 | STUDY 1: DATA SHARING BEHAVIOUR OF FACEBOOK, SPOTIFY AND TWITTER USERS

3.1 | Methods

3.1.1 | Data

The data for Study 1 came from a non-probability panel of German Internet users who agreed to use software that tracks their web browsing behaviour on desktop computers and/or smartphones. The panel is managed by a German market research company and contains around 2000 participants (the sample size fluctuated due to dropout and consecutive sample refreshing; additional information about the panel can be found in Data S1 – Section D). For our study, we acquired access to the web tracking data from June 2018 to May 2019. During that period, participants of the tracking panel were invited to complete two web surveys which included the data sharing requests. All 2042 individuals, who participated in the web tracking panel in July 2018, were invited to participate in the first online survey, of which 1411 followed the invitation and started the survey (participation rate = 69.1%). One thousand and three hundred fifty-five panellists completed the first survey (completion rate = 96.0%). For the second survey, all 2007 individuals, who were part of the web tracking panel in March 2019, were invited and 1325 took part in the survey (participation rate = 66%), of which 1240 completed the survey (completion rate = 93.6%). Of those 1240 respondents in the second survey, $n = 804$ (64.8%) had also participated in the first survey.

In the first survey, we asked respondents whether they were willing to share their Twitter data, and in the second survey, whether they were willing to share Facebook and Spotify data

(see Table 1 for how the factors influencing data sharing map on the two studies presented in this paper). Survey 1 also included an incentive experiment, in which respondents were randomly assigned to receive a 5€ data sharing incentive after (post-paid) or before (pre-paid) the completion of the data sharing procedure for their Twitter data (notably, the panellists were used to receiving post-paid incentives for their participation in the web tracking as well as in online surveys). In Survey 2, for the sharing of their Facebook data, respondents received a 5€ post-paid incentive, and for the sharing of their Spotify data, respondents received a 2.50€ post-paid incentive; there was no experimental variation of incentive conditions in the second survey. These incentives were higher than the amount participants typically get paid for only answering a survey (for completing the first survey, participants received 1.50 Euros, and for the second survey 2 Euros, as it was expected to be slightly longer). The overall incentive amount was limited by budget constraints and the specific amount for each platform was meant to reflect the assumed response burden as well as the perceived sensitivity of the data, which we assumed to be higher for Facebook than Spotify data. The median response time for those who completed the first survey was 15 min and 2 s. For the second survey, it was 13 min and 34 s.

Since our study was based on a non-probability sample, we compared key demographic variables to those of the probability-based GESIS Panel (Bosnjak et al., 2018), limiting the comparison only to GESIS Panel respondents who use the Internet (see Table S1). We selected the GESIS Panel because it includes the relevant measures and Bosnjak et al. (2018) showed that the panel represents the German population. In addition, we compared the demographics of Facebook and Twitter users in our data to those in the GESIS Panel (see Tables A2 and A3). These three comparisons showed that the samples were quite similar with respect to most demographic variables. One of the main differences was that the social media users in the GESIS Panel were more highly educated than the ones in our study.

3.1.2 | Data sharing procedure

Different methods were used to collect the social media data in Study 1. First, respondents were asked to provide informed consent for each of the data requests in the two surveys (see Appendix C in Data S1). In addition to the information provided in the survey questions, participants had the opportunity to read further information about privacy and data handling via a website whose URL was prominently placed in the survey text. However, web tracking data showed that a maximum of four (Spotify) to seven (Facebook) respondents who consented to sharing their data accessed this extended information for at least one of the three data types. Twitter data were collected using the platform's public APIs and included profile information and up to 3200 past tweets (collected via the REST API) as well as new tweets by the participants from the end of the field phase of the second online survey in August 2018 until the end of the project's overall data collection phase in May 2019 (via the Streaming-API). The Twitter data could be linked with the survey data via the username/handle. Respondents who indicated that they have a personal Facebook account were asked whether they are willing to instal a browser plugin that collects public posts from their Facebook news feed whenever users login to their Facebook account and see or scroll through their feed (for details see Haim & Nienierza, 2019). The plugin was available for the desktop versions of the Firefox and Chrome browsers and could be downloaded and installed through the respective official plugin stores (Haim & Nienierza, 2019). Participants were able to deactivate the plugin and could also delete the data that it collected. Notably, none of the participants made use of the latter option. The Facebook data could be linked with the survey data through an anonymised

ID code that participants were asked to generate in the survey and during the plugin installation process. Respondents who reported that they have a Spotify account were asked to provide data on the last 50 songs played, their playlists and music preferences as defined by Spotify through a web app accessible via a link in the survey. To use the web app, participants were asked to log in with their Spotify account. The app then collected the data via the Spotify API. Participants could review the data and decide whether they want to share it or not. The Spotify data and the survey data could be linked via a numeric participant ID that was passed on as a URL parameter from the market research company to the online survey platform and then to the Spotify web app.

3.1.3 | Measures

Regarding data sharing behaviour of Twitter, Facebook and Spotify data, we generated a dichotomous measure for each data type, indicating whether a respondent shared the respective data. Through the web and app tracking data, platform usage for Twitter, Facebook and Spotify was tracked. Specifically, we used the number of website visits and app usage for the tracked period before the surveys. Survey 1 included a survey evaluation measure and Survey 2 included a measurement of privacy concerns. In addition, the device respondents used to answer the surveys (smartphone/tablet vs. PC) and demographic information about respondents' age, gender, education and income were collected. We assumed that the distinction between device types (smartphone and tablet vs. desktop computer) would be especially important for the Facebook data as the plugin was only available for desktop browsers, which may lead to lower data sharing rates on smartphones and tablets compared to desktop computers. Yet, the survey could be taken on a smartphone, while the data could be shared on a desktop computer, so that it is meaningful to include both in the analyses. For the platforms included in the second survey (Facebook and Spotify), we computed participation in the first survey as an additional binary measure. See Table B1 in Data S1 for the description, source and coding of all measures used in Study 1, and Tables B2 and B3 in Data S1 for the descriptive statistics of those variables.

3.1.4 | Analyses

The results for Study 1 are based on descriptive analyses and logistic regression models predicting data sharing behaviour. For the descriptive analyses, our main goal was to learn how many participants drop out during each of the three steps of the data sharing process (Step 1: platform usage; Step 2: informed consent; Step 3: data sharing). In the regression models, the dependent variable is dichotomous with 'shared the respective data' coded as 1 and 'did not share data' coded as 0. For those models, the category 'did not share data' includes respondents who agreed to share the data but eventually did not share it either because they decided against it during the data sharing process or because of technical difficulties. Additional models predicting 'informed consent' versus 'no informed consent' are displayed in Data S1 (Table A6 for Study 1 and Table A7 for Study 2). We also computed additional regression models that include platform as well as the predictors that are available for all platforms (see Table A8 for Study 1 and Table A9 for Study 2 in Data S1). All data processing and analyses were performed with R 4.1.2 (R Core Team, 2020).

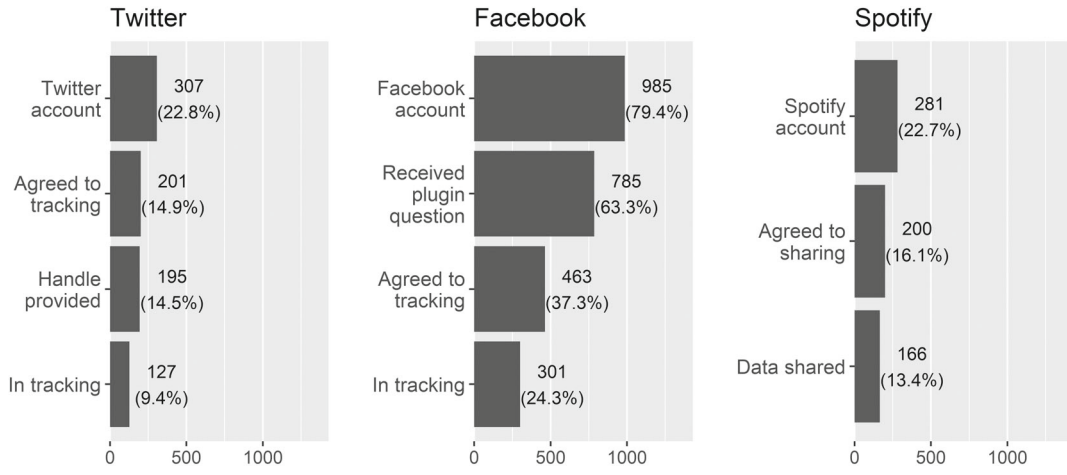

FIGURE 1 Number of respondents at different recruitment stages for Study 1. Percentages are based on the full samples. Survey 1 (Twitter): $N = 1347$, Survey 2 (Facebook and Spotify): $N = 1240$

3.2 | Descriptive results

Regarding our RQ1, the descriptive results from Study 1 (see bar charts in Figure 1) show that the sample included 79.4% Facebook users, 22.8% Twitter users and 22.7% Spotify users. These percentages are substantively higher than those from the *Reuters Digital News Report* (Newman et al., 2021), which reports 44% Facebook users and 12% Twitter users among the German online population. Of the users, 31.2% shared their Facebook data, 41.4% their Twitter data and 59.1% their Spotify data. Considering the full samples, 24.3% of the respondents shared Facebook data, 13.4% Spotify data and 9.4% Twitter data. Considering only those respondents who agreed to the data sharing request, 83.0% shared their Spotify data, 65.0% their Facebook data and 63.2% their Twitter data. The consent rate of 65.5% for Twitter data in our study was clearly higher than the consent rates between 27.1% and 36.8% reported by Al Baghal et al. (2020).

Notably, only 785 of the 985 Facebook users (79.7%) received the consent request for the Facebook plugin. The reason for this was a technical issue during the data collection. Specifically, Facebook changed the newsfeed during our data collection phase, which caused the plugin to freeze the browser tab when users clicked on a picture from their feed to enlarge it. With respect to Twitter data, we manually checked whether the Twitter handles were valid and whether they likely belonged to the respondents (and not a celebrity or some other person/institution). This resulted in the exclusion of $n = 68$ (34.9%) of provided handles. Mirroring these descriptive results, the combined regression model predicting data sharing behaviour across data types showed that the lower data sharing rates for Facebook and Twitter compared to Spotify data were statistically significant (see Table A8 in Data S1). While the data type (RQ1) and the sharing method (RQ2) were confounded in this study, the differences in dropout at the various stages suggest that both of these factors matter for data sharing decisions.

3.3 | Predicting data sharing

Regarding respondent characteristics (RQ3), we found several relevant predictors. Higher platform usage, as measured by the tracking data, significantly increased the likelihood of

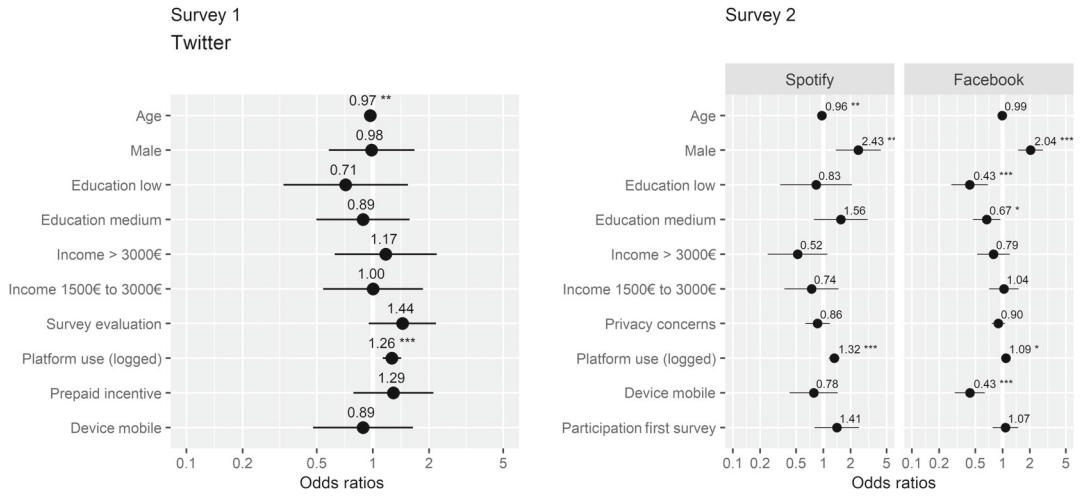

FIGURE 2 Coefficient plot for logistic regressions predicting data sharing behaviour in Study 1. Results are based on separate logistic regression models for each data type. Full regression tables are provided in Data S1. ***$p < 0.001$ **$p < 0.01$ *$p < 0.05$

data sharing for all three data types (see Figure 2 and Table A4 in Data S1). Some of the socio-demographic variables were also significant predictors of the data sharing probability for all three data types. Specifically, for Twitter and Spotify, younger respondents were more likely to share their digital data. For Facebook and Spotify, the odds of male respondents regarding data sharing were twice as high as for female respondents, and for Facebook, respondents with a high level of education had twice the odds to share their data compared to respondents with low education. The device used to answer the survey only affected the likelihood of data sharing significantly for Facebook. Respondents who answered the survey on a desktop computer had twice the odds to share their Facebook data compared to respondents who answered the survey on a mobile device (tablet or phone). Respondents' survey evaluation and privacy concerns did not significantly affect the likelihood of data sharing, and neither did participation in the first survey. Regarding the incentive experiment in the first survey (RQ4), pre-paid and post-paid incentives led to the same sharing rates for Twitter data.

4 | STUDY 2: DATA SHARING BEHAVIOUR OF TWITTER AND HEALTH APP USERS

4.1 | Methods

4.1.1 | Data

The data for Study 2 come from a German non-probability online panel ($N = 3136$; invitations = 26,339; participation rate = 39.8%; completion rate = 86.2%). The field period of the web survey in Study 2 was from October 2019 to December 2019, and respondents received a post-paid incentive of 1.50€ for their participation. To increase the number of eligible respondents for the data sharing requests, Twitter usage and smartphone type (iPhone or Samsung) were used as

screening variables (screen-out rate = 65.5%) in this study. iPhones and Samsung smartphones were selected because they are the two most frequently used smartphone types in Germany. Respondents were asked whether they were willing to share their Twitter or health app data (the introductory texts are reported in Appendix C in Data S1). Each respondent received only one of the two data sharing requests (Twitter or health app data) within the survey. All respondents who reported having a Twitter account received the Twitter data sharing request (about one third of the sample). Respondents who did not have a Twitter account were asked whether they have an iPhone or Samsung smartphone, and received the health app data sharing request if they answered affirmatively (about two thirds of the sample). For both data sharing requests, an incentive experiment with four groups was implemented: Respondents were randomly assigned to receive 0€, 2.50€, 5€ or 10€ for sharing their digital traces. The specific incentive amounts were selected to be similar to Study 1. The median response time of the survey was 12.4 min (additional information about the panel can be found in Data S1 – Section D).

As for Study 1, we compared demographic variables to those of the probability-based GESIS Panel, limiting the comparison only to GESIS Panel respondents who use the Internet (see Table S1). In addition, we compared the demographics of Twitter users in our data to those in the GESIS Panel (see Table A3). These comparisons showed some important differences across the samples. Specially, the sample of Study 2 was younger, and included more people with lower levels of education and income than the GESIS Panel sample. Those sample differences were most likely to due to screening for platform users in our study. Similarly, Twitter users in Study 2 had less formal education and lower income compared to those in the GESIS Panel. Finally, the respondents in Study 2 had more positive attitudes towards surveys with respect to survey enjoyment than GESIS Panel respondents.

4.1.2 | Data sharing procedure

For Twitter data, respondents were randomly assigned to share their data either by providing their user handle (passive procedure) or through data donation (active procedure). Health app (only for iPhone or Samsung devices) data were shared via data donation. To donate their data, respondents were first asked to export their data from the respective application or via the website (Twitter). In a second step, they were asked to upload the data through a secure web tool (Cryptshare), which is regularly used for sharing research data. To lower the burden, we provided instructions for each service, describing each step of the download and upload process (these instructions can be found in Data S2 – Section E). Notably, however, completing the data sharing procedure for the health app was more effortful on Samsung devices (due to the need to identify and upload only certain files).

4.1.3 | Measures

With respect to data sharing behaviour for Twitter and health app data, we included a dichotomous measure for each data type, indicating whether a respondent shared the respective data. For the health app data, we included a variable indicating a respondent's smartphone type (iPhone or Samsung). Platform usage was measured via self-reports for both Twitter and health app usage. Moreover, the survey included measures of privacy concerns, earlier experiences with privacy intrusion, technological affinity, perceptions of surveys in general (value, enjoyment and burden;

see De Leeuw et al., 2019), and the evaluation of the survey (see Gummer & Daikeler, 2020). In addition, data on the device (smartphone vs. PC/tablet) used for answering the survey and demographic information about respondents' age, gender, education, income and personality characteristics (conscientiousness and openness) were collected. Device was coded as smartphone (1) versus PC/tablet (0) in Study 2 because donating data from the Apple and Samsung health apps was only possible using a smartphone. This was done because answering the questionnaire on a device other than a smartphone required using an additional device to complete the data sharing task. See Table B4 in Data S1 for the description and coding of all measures of Study 2, and Table B5 for the descriptive statistics of those variables.

4.1.4 | Analyses

As for Study 1, the results for Study 2 are based on descriptive analyses and logistic regression models predicting data sharing behaviour.

4.2 | Descriptive results

In accordance with the screening procedure (RQ5), the descriptive results from Study 2 (see Figure 3) show that the survey included 33.8% Twitter users and 66.2% health app users (32.6% iPhone users and 33.6% Samsung phone users). With respect to data type (RQ1), 24.0% of Twitter users shared their data, and 9.4% of iPhone users and 2.7% of Samsung phone users shared their health app data. Considering the full sample, 7.0% of the respondents shared Twitter data through providing their user handle and 1.1% via data donation; 3.1% shared iPhone health app data and 0.9% shared Samsung health app data. Considering the sharing method (RQ2), of the respondents

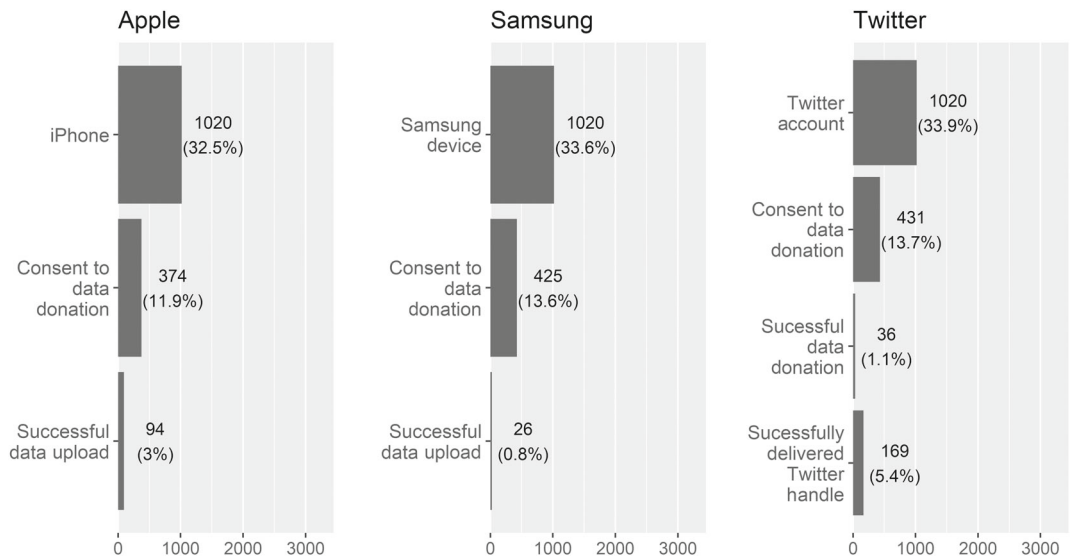

FIGURE 3 Number of respondents at different recruitment stages for Study 2. Percentages are based on the full sample ($N = 3136$).

who agreed to the respective data sharing requests, 95.6% shared their Twitter data by providing a user handle, and 17.3% via data upload. By comparison, 22.2% shared their iPhone health app data and 5.6% shared their Samsung health app data. Also, in this study, the consent rate of 42.3% for Twitter users was higher than the consent rates reported by Al Baghal et al. (2020).

The lower data sharing rate for Samsung users compared to iPhone users was most likely due to the more burdensome sharing procedure for Samsung smartphones. Regarding the provided Twitter handles, we did not verify their validity in this study. However, if we would assume a similar percentage of invalid Twitter handles as in Study 1 (35%), a substantial difference between the two Twitter data sharing methods would remain. The combined regression models predicting data sharing across data types, which were limited to (a) the data donation data sharing method and (b) respondents who gave data sharing consent to increase comparability, showed a significant lower data sharing rate (about 50% lower odds) for Twitter data than for Health data via iPhones (see Table A9).

4.3 | Predicting data sharing

Higher incentives (RQ4) significantly increased the odds ratios for data sharing for both data types up to 3.5 times (see Figure 4 and Table A5 in Data S1). However, while for Twitter data an incentive of 2.50€ already significantly increased the sharing likelihood compared to not providing an incentive, for health app data, only the 10€ incentive condition led to a statistically significant increase in the data sharing rate. Similar to Study 1, we also found several respondent characteristics (RQ3) to be predictive of data sharing. Higher health app usage led to a significant decrease in the likelihood of data sharing, while self-reported platform usage did not show a significant effect for Twitter data. In addition, earlier privacy intrusion experiences significantly decreased the likelihood of data sharing for health app data, and the personality trait of openness significantly decreased the likelihood of data sharing for Twitter data. In contrast, an affinity for technology significantly increased the likelihood of data sharing for Twitter but not for health app data. With respect to demographics, younger respondents were significantly more likely to share their health app data compared to older respondents and respondents with lower education were significantly less likely to share their Twitter and health app data than respondents with high levels of education. Moreover, respondents with higher income were less likely to share Twitter and health app data than respondents with lower income levels. Respondents who answered the survey on their smartphones had only half the odds to share health app data compared to respondents on PCs and tablets, and respondents who owned a Samsung smartphone had 64% lower odds of sharing health app data than iPhone owners. These two device effects were statistically significant. A positive survey evaluation significantly increased the likelihood of data sharing for both data types, while general attitudes towards surveys and privacy concerns were not statistically significant in Study 2.

5 | DISCUSSION

5.1 | Summary of results

Taken together, our studies show that data sharing rates can vary dramatically between samples, data types and data sharing methods. Important predictors of survey respondents' willingness to

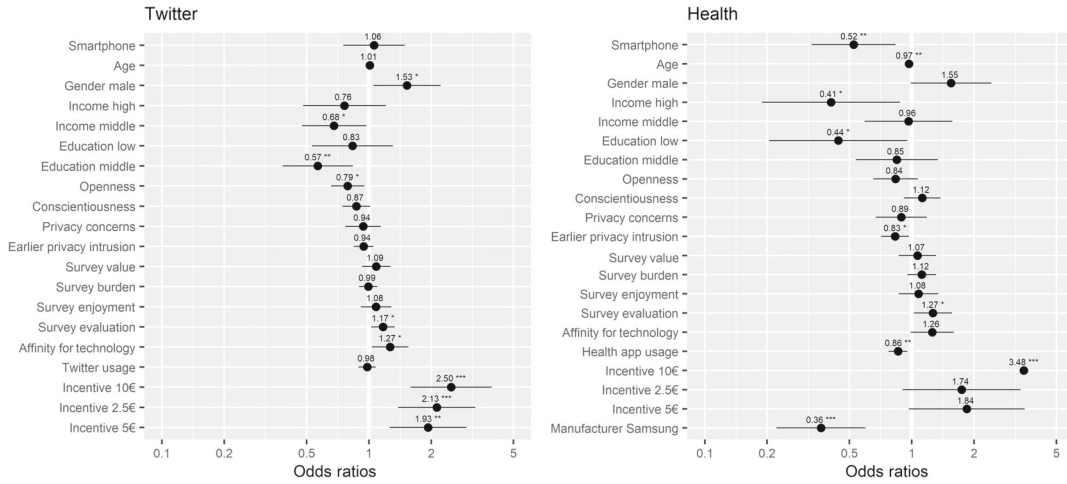

FIGURE 4 Coefficient plot for logistic regressions predicting data sharing behaviour in Study 2. Results are based on separate logistic regression models for each data type. Full regression tables are provided in Data S1. ***$p < 0.001$ **$p < 0.01$ *$p < 0.05$

share different types of digital trace data were incentive size, socio-demographic characteristics, such as gender, education and age, and the device used for completing the survey as this was connected to the ease of the data sharing. Since a crucial factor during the data sharing process appears to be respondent burden, data sharing requests generally seem to represent a high-cost situation (Best & Kroneberg, 2012). As a consequence, attitudes and values only affect sharing behaviour to a small degree. Hence, factors that directly relate to the data sharing difficulty, such as the data sharing method (Araujo et al., 2021; Boeschoten et al., 2020) should be tested and optimised for research that seeks to combine surveys and digital trace data. Providing survey respondents with additional incentives (Keusch et al., 2019) also seems beneficial for increasing the willingness to share data from digital platforms.

For *data type* (RQ1), our research shows that the users were most willing to share Spotify data (59.1%), while health app data, in contrast, had the lowest overall data sharing rates (6.1%). Our results confirm findings from previous research comparing different data types (Revilla et al., 2019; Wenz et al., 2019), which also showed that differences in sharing rates can be large (e.g., between 73.7% for receiving a product at home and 5.5% for requesting that respondents' children wear a small device that delivers real-time information about stress levels, see Revilla et al., 2019). Apart from the ease-of-use of the sharing methods, one likely reason for the differences in sharing rates for the various platforms in our study is how private and sensitive the platforms and their data are perceived by the users. For example, Twitter is typically viewed as a space for public communication, while Facebook is used more for personal and private communication.

With respect to the *data sharing method* (RQ2), our experimental design in Study 2 showed that respondents were more likely to share their data when they were asked to provide their user handle compared to exporting and uploading their Twitter data themselves. Since respondents only had to type in their Twitter handle for the API data collection, this sharing procedure was considerably less effortful. A similar effect was found for health app data sharing, which was especially effortful on Samsung devices, resulting in a data sharing rate

below 1%. These results indicate that, while the active data sharing method of data donation (Araujo et al., 2021; Boeschoten et al., 2020) is a promising tool for social research, researchers need to improve automatization and reduce respondent burden to increase the likelihood of a donation.

Looking at *respondent characteristics* (RQ3), we found that demographics and devices used were more likely to influence data sharing behaviour than respondents' attitudes and values. Specifically, we found that younger respondents, male respondents and respondents with a high level of education were more likely to share their digital trace data. Potential reasons for the demographic differences are technological affinity and usage behaviour. In fact, when not including technological affinity in Study 2 (Figure 3), male respondents were more likely than female respondents to share both Twitter and health app data. This effect, however, disappeared when a measure for technological affinity was included. With respect to user behaviour, previous studies have shown gender differences in the use of social media platforms. For example, a study by Muscanell and Guadagno (2012) found that men use Facebook more often for forming new relationships, whereas women use it more often to maintain existing relationships. Another study found that women are more concerned about privacy and engage more often in privacy-protecting behaviour on Facebook than men (Hoy & Milne, 2010). We also found mean differences in privacy concerns ($t[1230.5] = 3.01$, $p = 0.003$, $d = 0.17$) between women ($M = 3.75$, SD = 0.93) and men ($M = 3.59$, SD = 0.92) in Study 1. Notably, while privacy concerns were included in the regression model for sharing Facebook data, gender was a significant predictor, with men being more likely to share their Facebook data than women. An explanation for this finding might be that higher privacy concerns among women are associated with different usage behaviours (e.g., private vs. professional or bridging vs. bonding social capital), which may have caused the gender differences in willingness to share the data.

Interestingly, we found that the likelihood of sharing digital trace data increased with more regular platform usage for Facebook, Twitter and Spotify data (Study 1), but decreased the probability of data sharing for health app data. Possibly, health information might be considered more sensitive compared to social media data that is often freely available for other users or at least directly visible to the users' contacts on these networks. As more use means more data, it may be that more frequent users of health apps are more concerned about the sensitivity of their data. The type of device used for answering the survey affected data sharing for Facebook and health app data. Both were particularly effortful (health app) or impossible (Facebook desktop application) to engage in on a smartphone. Similar to Elevelt et al. (2019), we did not find substantial effects of survey attitudes and privacy considerations on data sharing behaviour in our studies.

The lacking influence of privacy attitudes is noteworthy since one might assume that they are one of the main determinants of data sharing behaviour. In this context it is important to distinguish between the different steps in the data sharing procedure. While our additional models (see Tables A6 and A7 in Data S1) indicate that privacy concerns do play a role for the second step of the data sharing process (*informed consent*), this effect does not carry over to the third step (*data sharing behaviour*). This finding suggests that the second step (consent) constitutes a low-cost situation, while the third (actual sharing) induces a situational change that constitutes a high-cost situation, in which privacy norms and attitudes are less impactful.

Our *incentive* experiments (RQ4) within the surveys confirmed the assumption that respondents were more willing to share additional data when they received higher incentives (Study 2).

However, whether the incentives were pre- or post-paid did not influence data sharing behaviour (Study 1). Compared to not offering incentives, an incentive of 2.50€ increased the likelihood that respondents agreed to share their Twitter data but for the health app data only an incentive of 10€ made a difference with respect to data sharing behaviour. This difference between Twitter and health app data suggests that if the sharing process is more effortful and/or the data are more sensitive, only high incentives can motivate respondents to share their data. However, even with high incentives, the data sharing rates remained relatively low for health app data, especially for respondents with Samsung devices where the data sharing process was most difficult. This finding is in line with the previous research showing that higher incentives do not increase the data sharing likelihood for all recipients (Jäckle et al., 2019) and that the time when incentives are given does not influence the likelihood of data sharing (Keusch et al., 2019). Our results regarding incentives also illustrate that some findings from the literature on incentives for survey participation (e.g., Singer, 2018) generalise to data sharing behaviour (i.e., the non-linear relationship between incentives and participation), while others are not or less applicable (i.e., pre-paid work better than post-paid incentives).

With respect to the *sample composition* (RQ5), we obtained higher linkage rates when we recruited respondents from the web tracking sample (Study 1), compared to the regular online access panel sample (Study 2). Participants in the tracking study had already agreed to provide more information than regular panel respondents so that it appears logical that they were also more willing to provide additional digital trace data. In Study 2, by contrast, we successfully implemented a screening procedure to increase the number of Twitter users as well as iPhone and Samsung device users.

5.2 | Recommendations

Requests for sharing digital trace data are likely to represent a high-cost situation in which respondents need to see their benefits clearly to be willing to share their data. Based on our two studies, we can derive four recommendations regarding the sample composition, data sharing method, incentives and the devices used to answer a survey. These recommendations might be especially important for sensitive data where respondents might be hesitant to share them with the researchers.

Recommendation 1: *When designing studies, researchers should consider the percentage of users of their targeted platforms or devices in the sample. If the target population consists of users of a particular platform, a possibility for increasing the percentage of users in the sample is implementing a screening procedure as we have done in our second study. If, however, the target population is the general population (of Internet users), screening would counteract the purpose of representing the general population. Hence, in such cases, we would rather recommend focusing on optimising the study design for increasing the data sharing rate.*

Recommendation 2: *In general, it is recommendable to minimise the effort required for sharing the data. This is especially important for data sharing methods in which respondents are asked to actively share their digital trace data ('data donation') since these are often more burdensome than, for example, merely asking for a username and informed consent.*

Recommendation 3: *Small incentives can increase the likelihood of data sharing if the data sharing task is not too burdensome. While large incentives can help to increase the likelihood of data sharing for very burdensome tasks, we recommend simplifying the data sharing process wherever possible so that small incentives are sufficient.*

Recommendation 4: *We recommend pretesting the feasibility of the data sharing process on multiple devices, especially on smartphones and tablets. An increasing number of respondents prefer completing questionnaires on mobile devices, so that researchers need to consider this during the study planning and design phase. The same is true for the use of the platforms and tools that generate digital trace data (e.g., social media or health apps). These apps are increasingly used on mobile devices, so the sharing methods should ideally be independent of the used device.*

5.2.1 | Data-specific recommendations

With respect to *Twitter data*, the optimal design seems to be asking for the Twitter handle and providing a small incentive of about 2.50€. While this necessitates extra effort for the researchers to collect the data (via the platform API), it significantly reduces respondent burden and, hence, leads to higher sharing rates. *Health app data* seem to be perceived as particularly sensitive so that frequent app users may be (more) hesitant to share their data. To increase the data sharing rate in this case, higher incentives of about 10€ seem advisable, and the data sharing process should be designed in a way that minimises respondent burden. For *Facebook data*, we found clear gender differences that were likely due to usage patterns. Hence, it may be necessary to consider different ways of user behaviour and data privacy perceptions for this platform. For example, it may be worthwhile to implement an adaptive design that adjusts incentive sizes based on respondents' reported usage and privacy concerns. *Spotify data* appear to be not considered particularly sensitive and a small incentive of around 2.50€ seem to be sufficient to elevate data sharing rates. Similar to Twitter data, a challenging task for this data type will likely be recruiting enough users (depending on the diffusion rates of the platforms in the respective country/population being studied).

Notably, those recommendations are based on the two studies presented here, so their generalisability needs to be tested in future research. For example, respondents of both studies were part of an online panel in which incentives were usually post-paid. Thus, the effect of pre-paid incentives might be different for other respondents. Similarly, the recommended incentive amount of 2.50€ for Twitter and Spotify data should be seen in the context of the surveys for which respondents received 1.50€ or 2€, respectively. In surveys for which respondents receive higher incentives for their participation, the incentives for additional data requests may also need to be higher.

5.3 | Limitations

While our research included several experimental and quasi-experimental (iPhone vs. Samsung device) designs, most comparisons are observational (see Table 1). For example, our study did not include an experimental design that directly compares different types of digital trace data. Accordingly, further carefully designed experiments are required for gaining additional insight into respondents' data sharing behaviour. Given that linking surveys and digital trace data is a relatively new approach, there may well be other factors influencing data sharing behaviour that we did not consider in our studies. Hence, the findings from our study need to be tested further for other samples, sharing methods and types of data in replication studies.

The data for both of our studies came from non-probability samples, which are adequate for explorative and experimental studies. However, additional evidence is needed from samples drawn with probability-based methods to test whether the results are generalizable. The

comparison of our samples and sub-samples of the respective platform users to the general population shows some relevant differences (see Tables A1–A3), which should be considered when interpreting the results. Given that our respondents were part of commercial access panels, we expect data sharing rates to be lower in general population samples. For such samples, following our Recommendations 2–4 will likely be even more important for increasing consent and data sharing rates. If researchers are interested in one or more groups of users (of specific platforms or devices), a screening procedure (Recommendation 1) may be a suitable way of increasing the likelihood of data sharing.

The bias towards sharing was likely even more pronounced for the participants of the web tracking panel as they already agreed to have their browsing behaviour tracked. Also, while it provides more reliable information than self-reported data, the tracking data itself has limitations. There are three main reasons that these data may not give the complete picture of participants' web and app use: (1) participants were able to pause the tracking; (2) although the participants were instructed that they should only register devices that they use alone, it may be that some of the tracked use comes from other individuals; (3) participants can use other untracked devices.

Finally, the browser plugin we used to collect Facebook data in Study 1 is a web scraping tool. While previous studies also used this approach (e.g., Mancosu & Vegetti, 2020), the Terms of Service (ToS) of Facebook do not allow automated data collection by means of web scraping. However, recent legislation from the United States (Sandvig, 2020) indicates that academic researchers should not be faced with civil or even criminal liability when violating platform ToS. A legal opinion included in a publication by the German Data Forum (German Data Forum (RatSWD), 2020) which is relevant for the European context also comes to the conclusion that web scraping for academic purposes should be possible and treated differently than scraping for commercial purposes from a legal standpoint. Nevertheless, compared to the data donation approaches used in Study 2, the use of such tools carries more risk for the researchers as, for example, recent actions of Facebook against NYU researchers have shown (Vincent, 2021).

DATA AVAILABILITY STATEMENT

A replication dataset and all analyses code for reproducing the results and supplementary results of this manuscript will be made available at a public repository.

ACKNOWLEDGEMENT

Open Access funding enabled and organized by Projekt DEAL.

ORCID

Henning Silber https://orcid.org/0000-0002-3568-3257
Johannes Breuer https://orcid.org/0000-0001-5906-7873
Tobias Gummer https://orcid.org/0000-0001-6469-7802
Florian Keusch https://orcid.org/0000-0003-1002-4092
Pascal Siegers https://orcid.org/0000-0001-7899-6045
Sebastian Stier https://orcid.org/0000-0002-1217-5778
Bernd Weiß https://orcid.org/0000-0002-1176-8408

REFERENCES

Al Baghal, T., Sloan, L., Jessop, C., Williams, M.L. & Burnap, P. (2020) Linking twitter and survey data: the impact of survey mode and demographics on consent rates across three UK studies. *Social Science Computer Review*, 38(5), 517–532.

Amaya, A., Bach, R., Keusch, F. & Kreuter, F. (2021) New data sources in social science research: things to know before working with Reddit data. *Social Science Computer Review*, 39(5), 943–960.

Araujo, T., Ausloos, J., van Atteveldt, W., Loecherbach, F., Moeller, J., Ohme, J. et al. (2021) *OSD2F: an open-source data donation framework* [Preprint]. SocArXiv.

Best, H. & Kroneberg, C. (2012) Die low-cost-hypothese. *KZfSS Kölner Zeitschrift Für Soziologie Und Sozialpsychologie*, 64(3), 535–561.

Beuthner, C., Breuer, J. & Jünger, S. (2021) Data linking—linking survey data with geospatial, social media, and sensor data. *GESIS Survey Guidelines*. Available at: https://www.gesis.org/fileadmin/upload/SDMwiki/2021_Beuthner_Linking_1.pdf

Biner, P.M. & Kidd, H.J. (1994) The interactive effects of monetary incentive justification and questionnaire length on mail survey response rates. *Psychology & Marketing*, 11(5), 483–492.

Boeschoten, L., Ausloos, J., Moeller, J., Araujo, T. & Oberski, D.L. (2020) Digital trace data collection through data donation. *ArXiv Preprint ArXiv:2011.09851*.

Bosnjak, M., Dannwolf, T., Enderle, T., Schaurer, I., Struminskaya, B., Tanner, A. et al. (2018) Establishing an open probability-based mixed-mode panel of the general population in Germany: the GESIS panel. *Social Science Computer Review*, 36(1), 103–115.

Breuer, J., Bishop, L. & Kinder-Kurlanda, K. (2020) The practical and ethical challenges in acquiring and sharing digital trace data: negotiating public-private partnerships. *New Media & Society*, 22(11), 2058–2080.

Brosnan, K., Grün, B. & Dolnicar, S. (2017) Use, preference and completion rates for web surveys. *International Journal of Market Research*, 59(1), 35–55.

DiGrazia, J., McKelvey, K., Bollen, J. & Rojas, F. (2013) More tweets, more votes: social media as a quantitative indicator of political behavior. *PLoS One*, 8(11), e79449.

Dillman, D.A. (1978) *Mail and telephone surveys: the total design method*. New York: John Wiley and Sons.

Dutwin, D., Loft, J.D., Darling, J.E., Holbrook, A.L., Johnson, T.P., Langley, R.E. et al. (2015) Current knowledge and considerations regarding survey refusals: executive summary of the AAPOR task force report on survey refusals. *Public Opinion Quarterly*, 79(2), 411–419.

Eckman, S. & Haas, G.-C. (2017) Does granting linkage consent in the beginning of the questionnaire affect data quality? *Journal of Survey Statistics and Methodology*, 5(4), 535–551.

Elevelt, A., Lugtig, P. & Toepoel, V. (2019) Doing a time use survey on smartphones only: what factors predict nonresponse at different stages of the survey process? *Survey Research Methods*, 13(2), 195–213.

Esser, H. (1986) Über die Teilnahme an Befragungen. *Zuma Nachrichten*, 10(18), 38–47.

German Data Forum (RatSWD). (2020) *Big data in social, behavioural, and economic sciences: Data access and research data management*. *RatSWD Output Paper Series*.

Göritz, A.S. (2006) Incentives in web studies: methodological issues and a review. *International Journal of Internet Science*, 1(1), 58–70.

Groves, R.M., Singer, E. & Corning, A. (2000) Leverage-saliency theory of survey participation: description and an illustration. *Public Opinion Quarterly*, 64(3), 299–308.

Gummer, T. & Daikeler, J. (2020) A note on how prior survey experience with self-administered panel surveys affects attrition in different modes. *Social Science Computer Review*, 38(4), 490–498.

de Haan, J. & Hendriks, R. (2013) Online data, fixed effects and the construction of high-frequency price indexes. *Proceedings of the economic measurement group workshop*, pp. 28–29.

Haim, M. & Nienierza, A. (2019) Computational observation: challenges and opportunities of automated observation within algorithmically curated media environments using a browser plug-in. *Computational Communication Research*, 1(1), 79–102.

Halavais, A. (2019) Overcoming terms of service: a proposal for ethical distributed research. *Information, Communication & Society*, 22(11), 1567–1581.

Harari, G.M., Müller, S.R., Aung, M.S. & Rentfrow, P.J. (2017) Smartphone sensing methods for studying behavior in everyday life. *Current Opinion in Behavioral Sciences*, 18, 83–90.

Hoy, M.G. & Milne, G. (2010) Gender differences in privacy-related measures for young adult facebook users. *Journal of Interactive Advertising*, 10(2), 28–45.

Jäckle, A., Burton, J., Couper, M.P. & Lessof, C. (2019) Participation in a mobile app survey to collect expenditure data as part of a large-scale probability household panel: coverage and participation rates and biases. *Survey Research Methods*, 13(1), 23–44.

Jenkins, S.P., Cappellari, L., Lynn, P., Jäckle, A. & Sala, E. (2006) Patterns of consent: evidence from a general household survey. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 169(4), 701–722.

Juga, J., Juntunen, J. & Koivumäki, T. (2021) Willingness to share personal health information: impact of attitudes, trust and control. *Records Management Journal*, 31(1), 48–59.

Keusch, F., Struminskaya, B., Antoun, C., Couper, M.P. & Kreuter, F. (2019) Willingness to participate in passive mobile data collection. *Public Opinion Quarterly*, 83(S1), 210–235.

Keusch, F., Struminskaya, B., Kreuter, F. & Weichbold, M. (2020) Combining active and passive mobile data collection: a survey of concerns. In: Hill, C.A., Biemer, P.P., Buskirk, T., Japec, L., Kirchner, A., Kolenikov, S. & Lyberg, L.E. (Eds.) *Big data meets survey science: a collection of innovative methods*. Hoboken, NJ: Wiley, pp. 657–682.

King, G. (2011) Ensuring the data-rich future of the social sciences. *Science*, 331(6018), 719–721.

Kosinski, M., Stillwell, D. & Graepel, T. (2013) Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, 110(15), 5802–5805.

Kreuter, F., Haas, G.-C., Keusch, F., Bähr, S. & Trappmann, M. (2020) Collecting survey and smartphone sensor data with an app: opportunities and challenges around privacy and informed consent. *Social Science Computer Review*, 38(5), 533–549.

Kuru, O., Bayer, J., Pasek, J. & Campbell, S.W. (2017) Understanding and measuring mobile Facebook use: who, why, and how? *Mobile Media & Communication*, 5(1), 102–120.

Ledford, H. (2020) How Facebook, twitter and other data troves are revolutionizing social science. *Nature*, 582(7812), 328–330.

Leeper, T.J. (2019) Where have the respondents gone? Perhaps we ate them all. *Public Opinion Quarterly*, 83(S1), 280–288.

de Leeuw, E., Hox, J., Silber, H., Struminskaya, B. & Vis, C. (2019) Development of an international survey attitude scale: measurement equivalence, reliability, and predictive validity. *Measurement Instruments for the Social Sciences*, 1(1), 1–10.

Mancosu, M. & Vegetti, F. (2020) What you can scrape and what is right to scrape: a proposal for a tool to collect public Facebook data. *Social Media + Society*, 6(3), 205630512094070.

Muscanell, N.L. & Guadagno, R.E. (2012) Make new friends or keep the old: gender and personality differences in social networking use. *Computers in Human Behavior*, 28(1), 107–112.

Newman, N., Fletcher, R., Schulz, A., Andi, S., Robertson, C.T. & Nielsen, R.K. (2021) *Reuters institute digital news report 2021*. Reuters Institute for the Study of Journalism.

Nissenbaum, H. (2018) Respecting context to protect privacy: why meaning matters. *Science and Engineering Ethics*, 24(3), 831–852.

Oberski, D.L. & Kreuter, F. (2020) Differential privacy and social science: an urgent puzzle. *Harvard Data Science Review*, 2(1), 1–21.

Porter, S.R. & Whitcomb, M.E. (2003) The impact of lottery incentives on student survey response rates. *Research in Higher Education*, 44(4), 389–407.

R Core Team. (2020) *R: a language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. Available at: https://www.R-project.org/

Revilla, M., Couper, M.P. & Ochoa, C. (2018) Giving respondents voice? The feasibility of voice input for mobile web surveys. *Survey Practice*, 11(2), 2713.

Revilla, M., Couper, M.P. & Ochoa, C. (2019) Willingness of online panelists to perform additional tasks. *Methods, Data, Analyses*, 13(2), 29.

Revilla, M., Couper, M.P., Paura, E. & Ochoa, C. (2021) Willingness to participate in a metered online panel. *Field Methods*, 33(2), 202–216.

Riker, W.H. & Ordeshook, P.C. (1968) A theory of the calculus of voting. *The American Political Science Review*, 62(1), 25–42.

Sakshaug, J.W. (2020) Linking surveys with big data. In: *Qualität bei zusammengeführten Daten*. New York: Springer, pp. 163–173.

Sandvig V.B. (2020) Civil Action No. 16–1368 (JDB) (United States District Court for the District of Columbia March 27, 2020). Available at: https://www.aclu.org/sandvig-v-barr-memorandum-opinion.

Settanni, M., Azucar, D. & Marengo, D. (2018) Predicting individual characteristics from digital traces on social media: a meta-analysis. *Cyberpsychology, Behavior and Social Networking*, 21(4), 217–228.

Shlomo, N. & Goldstein, H. (2015) Big data in social research. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 178(4), 787–790.

Singer, E. (2018) Survey incentives. In: *The Palgrave handbook of survey research*. Cham: Palgrave Macmillan, pp. 405–415.

Sloan, L., Jessop, C., Al Baghal, T. & Williams, M. (2020) Linking survey and twitter data: informed consent, disclosure, security, and archiving. *Journal of Empirical Research on Human Research Ethics*, 15(1–2), 63–76.

Smith, A. & Page, D. (2015) *US smartphone use in 2015*. Pew Research Center. Available at: https://www.pewresearch.org/internet/2015/04/01/us-smartphone-use-in-2015.

Stern, P.C. (1992) Psychological dimensions of global environmental change. *Annual Review of Psychology*, 43(1), 269–302.

Stier, S., Breuer, J., Siegers, P. & Thorson, K. (2020) Integrating survey data and digital trace data: key issues in developing an emerging field. *Social Science Computer Review*, 38(5), 503–516.

Vincent, J. (2021) Facebook bans academics who researched ad transparency and misinformation on Facebook. *The Verge*, 4 August. Available at: https://www.theverge.com/2021/8/4/22609020/facebook-bans-academic-researchers-ad-transparency-misinformation-nyu-ad-observatory-plug-in.

Wenz, A., Jackle, A. & Couper, M.P. (2019) Willingness to use mobile technologies for data collection in a probability household panel. *Survey Research Methods*, 13(1), 1–22.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Silber, H., Breuer, J., Beuthner, C., Gummer, T., Keusch, F., Siegers, P. et al. (2022) Linking surveys and digital trace data: Insights from two studies on determinants of data sharing behaviour. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 185(Suppl. 2), S387–S407. Available from: https://doi.org/10.1111/rssa.12954