The blind spot in data donations

Who is (not) willing to donate digital data in social scientific research

Strycharz, J.; Meppelink, C.; Zarouali, B.; Araujo, T.; Voorveld, H.

DOI
10.5117/CCR2024.2.3.STRY

Publication date
2024

Document Version
Final published version

Published in
Computational Communication Research

Amsterdam University Press

COMPUTATIONAL COMMUNICATION RESEARCH 6.2 (2024) 1–27
HTTPS://DOI.ORG/10.5117/CCR2024.2.3.STRY

The blind spot in data donations: who is (not) willing to donate digital data in social scientific research

Joanna Strycharz
Amsterdam School of Communication Research, U. of Amsterdam, NL

Corine Meppelink
Amsterdam School of Communication Research, U. of Amsterdam, NL

Brahim Zarouali
Institute for Media Studies, KU Leuven, BE

Theo Araujo
Amsterdam School of Communication Research, U. of Amsterdam, NL

Hilde Voorveld
Amsterdam School of Communication Research, U. of Amsterdam, NL

Abstract
The use of online media has led to an increase in digital footprints of human behavior, which has resulted in a growing interest in the collection and analysis of such data in non-intrusive ways. One such approach is digital data donation, which involves requesting participants to share data that they have requested from digital platforms with researchers. While this approach promises to provide an unprecedented level of detail for computational communication research, it also raises concerns about the representativeness and validity of the data. This study investigates the issue of potential selection bias and non-response bias in data donation samples. It aims to identify subparts of the population that might be underrepresented (or even absent) in data donation sampling methods, which might lead to inaccurate and biased research conclusions. Utilizing a survey with a sample frame of 1178 and 289 participants from the Netherlands, the study investigates the relation between demographics, knowledge, privacy and trust factors, and donation behavior. The results show biases in donation behavior in terms of age and digital and algorithmic efficacy, while privacy and trust factors are not related to it. It suggests that individuals decide not to donate to academic research due to skills and not due to concerns or lack of trust. This offers possibilities for improvement in study design so that all willing individuals are equally able to participate.

Keywords: data donations, social media, digital traces, bias in data, representativeness, digital literacy skills

Introduction

The centrality and omnipresence of digital media has led to a culture in which people are permanently connected, being exposed to, and exchanging information (Vorderer et al., 2017). As much of this exchange is happening online, individuals also increasingly leave digital footprints of their communication activities: when communicating through emails, sharing opinions and personal details on social media, being persuaded by brands through digital ads or looking for products and information (Acquisti et al., 2015). This has resulted in an increased academic interest in the collection and analysis of digital traces of human behavior in a nonintrusive way allowing for higher precision than self-reports (Stier et al., 2020). Digital traces can be collected unobtrusively in different ways, for example through the Application Programming Interfaces (APIs) of social media platforms or web scraping tools. In recent years, novel methods have surfaced, such as letting participants install a browser plug-in on their computers or smartphones that can track and collect the personalized ads they were exposed to (e.g., Bol et al., 2020), or install software that takes automated screenshots of content seen on participants' smartphones every few seconds (Reeves et al., 2021). Although these methods are valuable in certain contexts, they also come with pitfalls such as access to a limited number of platforms or technical challenges (van Driel et al., 2022).

More recently, an additional promising collaborative approach to capturing digital trace data has been introduced that has a great potential for computational communication research: digital data donations. A data donation approach involves requesting participants to download their data that has been collected by digital platforms and to actively share this with researchers (Boeschoten et al., 2022). Such donated data includes, among others and depending on the platform, interactions with other users such as sharing and liking, posting, exposure to advertising and information and more. It promises to allow communication scholars to study human behavior and interaction at an unprecedented level of detail in collaboration with individual users, but more knowledge is needed about representativeness and thus indirectly the validity of measures constructed from such data (Lazer et al., 2021). Hence, this study investigates the representativeness and validity of data collected through donations.

Research involving digital trace data (for example using tracking plugins) can be vulnerable to biases related to sampling, as specific groups might be less able or willing to take part in this type of research (Boeschoten et al., 2022). This concerns academic research in general but is even more prob-

lematic in research that requires considerable effort and advanced digital skills. Hence, data subjects from vulnerable groups face a dual risk of being either made invisible or misrepresented. Regarding who is affected by these issues, past research has shown that demographic barriers may prevent individuals accessing their information in ways that are necessary to participate in data donation research (Anderson & Kumar, 2019). Donating data may also demand a certain level of technical expertise to access, download and upload data to the research platform in appropriate formats. Other concerns, like privacy or trust issues, may result in unequal participation in data donation projects (Bietz et al., 2019).

As data donation procedures can deter people for many different reasons, this research aims to investigate the generalizability of data collected through donations by focusing on the issue of potential selectivity and bias in the recorded donation responses (i.e., the issue of nonresponse bias and a lack of representativeness in the data donation sample). Our study diverges from existing research that has used a hypothetical dataset to present an error framework for data donations (Boeschoten et al., 2022), used an actual dataset to demonstrate the potential and challenges of data donations (van Driel et al., 2022), or asked respondents about their intention to donate data (Kmetty et al., 2024), but did not empirically investigate the profile of participants and systematically compared them to non-participants. By doing this, the study identifies subparts of the population that are underrepresented (or even absent) in data donation sampling methods. Uncovering the determinants of non-willingness to donate social media data can directly help researchers improve their recruitment strategies and research design, eventually improving representativeness of data donation samples.

Theoretical framework

Participation at different stages of a data donation study

While a data donation approach makes it possible to explore and empirically test ideas that researchers were unable to test with classical methods and data (e.g., experiments, cross-sectional surveys), much is still to be discovered about who is represented in such datasets and who is not. The reasons behind the (lack of) willingness to participate in data donation studies are partly similar to general research participation, but some aspects are unique due to the novelty of technology involved (Keusch et al., 2019; Struminskaya et al., 2020). Past research on willingness to share digital traces using vignette studies shows varying levels of participation inten-

tion: Keusch et al. (2019) in their investigation of mobile device tracking conducted in Germany concluded that 35% of respondents were willing to participate in such a study, while 39% were not willing to do so regardless of study conditions. Similarly, Skatova and Goulding (2019) in their study on likelihood of donating personal data conducted in the UK concluded that 54% of participants were likely to donate their personal data, while 31% were not likely to do so. Other studies showed higher willingness: Kmetty et al. (2024)) in a vignette study on social media data donation measured how willing individuals were to donate data (1 – not willing, 10 – very willing) and concluded that in Hungary the average willingness to donate was 4.2, while only 18% refused any kind of data sharing. In the context of medical data in the US, Seltzer et al. (2019) concluded that 65% of participants were willing to share their digital trace data. However, to what extent the willingness or intention result in actual (behavioral) donation remains unclear. Ample research on the intention-behavior gap has shown that intentions only partly explain the variance in behavior (Conner & Norman, 2022). This is particularly relevant in the context of personal data when potential privacy issues are at play. Past research suggests that privacy related behaviors are often paradoxical -so-called privacy paradox, Barnes (2006) - as individual intentions and behavior do not always align.

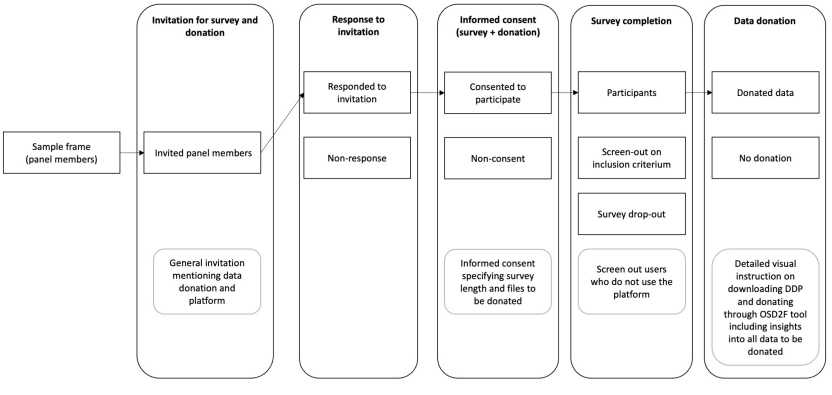
Past research has identified four different types of non-responses that affect studies that combine surveys and with collection of digital traces: 1) unit non-response in the survey (recruitment stage), 2) non-use of the online platform or service under investigation(self-report stage), 3) non-consent to the tracking or donating data (consent stage), and 4) non-response to the tracking (donation stage) (Stier et al., 2020). Hence, nonresponse can affect studies involving digital trace data at different stages (beyond non-response issues common to traditional survey research). Struminskaya et al. (2020) in their investigation of willingness to share smartphone-based tracking data differentiated the difficulty at the recruitment stage between the type of sample (nonprobability- vs. probability-based). While participants recruited through a non-probability route (e.g., flyers) express their willingness to participate, the probability route involves asking respondents to participate in the data donation study and collecting data from those who adhere to the request, sign informed consent and who can complete donation steps (see 1 for data donation procedure from a probability sample). At all these steps (recruitment, informed consent, self-reported data, data donation), participants can potentially decide to stop the study, which will result in a non-representative sample. Jürgens et al. (2020) introduced three stages at

which non-response can result in bias in data collection of digital traces: sampling bias (not only related to drawing a random sample, but also to willingness to participate in such a study), selection bias (related to the composition of the self-selected group that participates in the study), response bias (related to the errors that occur in the generation and collection of the trace data itself). Hence, considering the importance of drop-out at different stages of the procedure, we pose the following research question:

RQ1a: To what extent do individuals decide to participate at different stages of a social media data donation study?

RQ1b: In which stages of a social media data donation study do participants decide to drop-out from it?

Figure 1: The process of donating data from a probability sample and information provided to respondents



Demographic characteristics and donating data

Regarding who is affected by non-participation in data donation studies, different barriers can be identified that affect response rates at different stages. First, past research has shown that individuals with lower income may experience barriers when accessing their information in ways that are necessary to participate in data donation research (Anderson & Kumar, 2019). Hence, while some individuals may be willing to participate in data donation studies, they may not be able to do so as they are not able to download their

data donation package (DDP). Furthermore, previous research has shown that socio-demographic characteristics (e.g., older people) may prevent individuals from participating in data donation research (Bietz et al., 2019). Along these lines, age was found to be negatively associated with people's level of knowledge and skills about big data and social media algorithms (e.g., Zarouali, Boerman, & de Vreese, 2021), as well as how to protect their data and digital devices (e.g., de Vries et al., 2022). This could indicate not lack of willingness but lack of ability to donate data in a digital landscape. Also generally, research about digital divides often focuses on sociodemographic determinants of internet skills (Scheerder et al., 2017). As demographic characteristics might influence data donation behavior directly, but also through other variables such as skills or trust, it is important to examine the relative impact of both groups of variables on data donation behavior. Hence, the second research question focuses on demographic characteristics of non-participants:

RQ2: To what extent do demographic characteristics (age, gender, income, education) affect individuals' decision to donate their social media data?

Dispositional characteristics and donating data

Beyond demographic barriers, past research points to both digital efficacy as it are required for participation in data donation studies as well as privacy-related issues as digital trace data can be experienced as highly personal (the DDPs include individual-specific information that is generally experienced as more personal by internet users, Walrave et al. (2018)).

Digital efficacy

The act of donating one's own data can be a complex task. For example, it requires downloading and uploading skills, and the ability to navigate unfamiliar digital environments to store and locate folders. Even though Dutch citizens are found to belong to the most digitally skilled inhabitants of Europe (Eurostat, 2022), it has been shown that 20% of the Dutch citizens aged between 16 and 65 does not meet the level of basic digital skills (Non et al., 2021). Basic digital skills are for example, finding information on the internet or opening an e-mail. Generally, people who have a lower level of education, literacy skills, or an older age are overrepresented within the group of people having inadequate basic digital skills, however, over 23% of the people within this group are younger than 40 years of age (Non

et al., 2021). Furthermore, a study on digital competences among the Dutch population by de Vries et al. (2022) showed that a much bigger group of citizens finds it difficult to identify and cope with artificial intelligence used on websites and social media platforms. As digital skills and competences among a substantial part of the population are likely not sufficient to successfully complete the data donation task, this can be expected to impact data donation behavior.

Furthermore, we explore the extent to which people's awareness of the use of algorithms in the curation of messages on social media platforms (based on one's own past online behavior) is associated with data donation behavior. One possible scenario is that people with higher algorithmic awareness and digital efficacy realize the need for donating their platform data, as it is a highly valuable source of data for researchers (hereby providing them more insights into the black boxes that platforms are) (Beer, 2009; Zarouali et al., 2022). Thus, individuals with higher levels of algorithmic awareness and digital efficacy may be more aware of this and may perceive donating their platform data as an added value (leading to a higher likelihood of data donation). However, a second scenario involves individuals with higher levels of algorithmic awareness and digital efficacy being more critical regarding algorithmic-curated platforms (Silva et al., 2024; Zarouali, Helberger, & De Vreese, 2021), leading them to take a more cautious and protective approach and refrain from donating their data to researchers. Given these seemingly opposing scenarios, it is important to explore the (yet unclear) relationship between algorithmic awareness and digital efficacy and donation behavior. Regarding algorithmic awareness, recent studies found that a significant part of the Dutch population holds erroneous beliefs about algorithms (Zarouali, Boerman, & de Vreese, 2021), and that algorithmic awareness among Dutch citizens is rather low (Zarouali, Helberger, & De Vreese, 2021). Hence, the current study investigates the relation between different types of efficacy related to both general awareness and specifically data collection and algorithmic processing, and the decision to donate data. RQ3: To what extent do digital efficacy (i.e., algorithmic and digital efficacy,) affect individuals' decision to donate their social media data?

Dispositional factors

Privacy and data security concerns generally negatively influence willingness to self-disclose online (Baruh et al., 2017) and also to share personal data for research purposes. For example, individuals concerned about their privacy have a lower willingness to download or accept apps that track their

mobile media use (Keusch et al., 2019; Zarouali et al., 2022) and are less likely to actually download tracking apps and participate in such studies (Jäckle et al., 2019). In a similar vein, a recent study by Ohme et al. (2021) found that next to digital literacy, privacy concerns are a crucial factor for data donation behavior for mobile log data. Based on this, the authors concluded that the methodology of data donations may create a "digital divide in research", with a skew towards digitally skilled people and people with certain dispositional characteristics (e.g., low privacy concerns). Compared to e.g., tracking software, data donations offer higher transparency of what data is shared with researchers. This transparency may soothe concerns of respondents, but Ohme et al. (2021) argue that it is also possible that being confronted with the data (e.g., seeing how many files get donated in the tools used in this study and what they contain, see Figure 1) may have the reverse effect and lead to an even greater concern about privacy and lead to participants dropping out at the last stage.

Finally, across multiple disciplines, trust has been examined as a critical relational factor influencing various human behaviors in the presence of risks and uncertainty. As downloading DDPs is not widely known to nor used by the general population (5% of respondents to a survey on the General Data Protection Regulation that introduced the right to request DDPs have ever exercised this right, Strycharz et al. (2020)), requesting a DDP and donating it can be seen as a risky and uncertain behavior. McKnight et al. (2002) argued that the greater the perceived risk, the greater role trust plays in human behavior. Multiple studies have shown significant influence of trust in the web and in online companies on data sharing with commercial parties (Bol et al., 2018; McKnight et al., 2002; Metzger, 2006). A recent study highlighted that trust played an important role in people's willingness to accept an intervention where their (mobile) data is monitored and tracked (Zarouali et al., 2022). Along these lines, in the context of longitudinal data collection for research purposes, respondents' trust was consistently associated with higher consent (Sala et al., 2014). Considering the potential central role of individual characteristics, we ask:

RQ4: To what extent do dispositional variables (i.e., privacy concerns and trust in online companies and the web) affect individuals' decision to donate their social media data?

Methods

Procedure and participants

This data donation study focused on Facebook and is part of a larger project on the consequences of targeting on social media that ran between December 2021 and June 2022. In total, 1178 participants were invited to participate in the study. Participants were recruited via a panel company in the Netherlands. In the invitation panel members were informed about the inclusion criterium (Facebook use) and that the participation will require both filling in a survey and donating their data (they were informed that they would only receive compensation if they donated their data successfully). After clicking on the link provided in the invitation, the respondents were presented with informed consent in which they were informed (among others) about the specific categories from the Data Download Package (DDP) that they will be asked for (for example, Your_interactions_on_facebook). After consenting to take part in both parts of the study (survey and donation), participants were screened for Facebook use and only Facebook users proceeded to fill in a 15-minute survey. Upon survey completion, participants received detailed instructions on how to request and download their DDP from Facebook. Then, each participant was redirected to a data donation website that runs on the OSD2F infrastructure (Araujo et al., 2022). On this website, participants were asked to upload their DDP and then could explore the data and decide what parts of it to donate. When starting the uploading procedure, participants were informed that our data donation software automatically only selects specific categories of data named in the informed consent they agreed to earlier, i.e. the sections 'your profile information', your posts, comments, interactions and topics', your groups', 'pages you liked', your reactions to posts and comments', 'advertisers you have interacted with' and 'advertisers who uploaded your contact info'. In the OSD2F tool, they could then select themselves which parts of the data they want to donate and which parts not (they could select not to donate specific information e.g., donate only some of the advertisers named in the section 'advertisers you have interacted with'). They were also informed that personal messages will never be donated. If a participant did not complete a donation, they were sent two reminder emails directing them to the data donation website. If a participant indicated not to be able to donate, they were asked in an open-ended question to explain reasons for it. Once donated, the data was linked to the survey answers by a unique identifier. Figure 1 shows an overview of the data collection process. Ethical approval of this study was

provided by the research institute (number: 2021-CS-13824).

Measures

For all panel members who were invited to the study, the panel company provided information on demographics (age, gender, education level (six levels, CBS (2019a)), place of residence (five urbanization levels, CBS (2019b)). Education levels were recoded into three (low, medium, high) categories following CBS (2019a) standards. In the survey, additional questions were included to assess all other variables. All questions were asked on a 7-point Likert-scale. All items used can be found in the Appendix.

Algorithmic awareness in social media was measured using the 13-item algorithmic media content awareness (AMCA) scale (Zarouali, Boerman, & de Vreese, 2021). More specifically, this scale measures users' awareness of content filtering, automated decision-making, human-algorithm interplay, and ethical considerations. The items were for example, "Algorithms are used to tailor certain messages to me on my social media" and "The messages that algorithms recommend to me on my social media depend on my online behavioral data" (1 = I am not at all aware of this,7 = I am completely aware of this; *Cronbach's* α = 0.97; M = 4.44; SD = 1.66)

Self-reported knowledge about algorithms was asked with the question "How much do you know about algorithms?" (1 = Absolutely nothing, 7 = I am an expert; M = 2.97; SD = 1.64).

Digital self-efficacy was measured using eight items based on the Internet self-efficacy scale developed by Eastin and LaRose (2000). Items were slightly adapted, for example by including an item about using apps on mobile phones. A Likert-type scale was used to assess the participants' confidence that they could use the Internet in each of the ways specified, where 1 corresponded to "not at all confident" and 7 to "very confident." Participants were for example asked how much confidence that had to "talk about internet hardware, like networks and routers", and "to use apps on my mobile". An exploratory factor analysis suggests dropping two items and constructing two factors that measure digital self-efficacy: 1) Internet self-efficacy (*Cronbach's* α = .88; M = 5.53; SD = 1.22) and 2) digital self-efficacy (*Cronbach's* α = .93; M = 4.18; SD = 1.56).

Privacy concerns were measured using a five-item scale based on Baek and Morimoto (2012) and Kruikemeier et al. (2020) five-item scale. We asked participants the extent (1 = totally disagree, 7 = totally agree) to which they agreed with statements such as "I am concerned that my personal data (such as my surf and search behavior, name, and location) are misused by

others" and " I feel fear that personal information shared online may not be safe while stored" (*Cronbach's* α = .95.; M = 4.65; SD = 1.47).

Trust in online companies was measured with six items based on the 'trusting beliefs' items developed by Malhotra et al. (2004). Answer options ranged from (1 = totally disagree, 7 = totally agree). For example: "Companies are honest when it comes to using the personal information that I provide" and "I trust that companies would keep my best interests in mind when dealing with my personal information" (*Cronbach's* α = .92; M = 3.78; SD = 1.14).

Trust in the Web assessed individual trust in structural assurance of the internet and was measured by five 7-point Likert scale items (McKnight et al., 2002), including "I am comfortable sharing my personal information on the internet" (*Cronbach's* α = .91; M = 3.36; SD = 1.23).

Analysis

To gain insights into profiles of respondents who dropped out at the different stages of the study, first, the six respondent groups (non-response, no consent, no Facebook use, survey drop-out, no donation, donation) is compared on demographic and dispositional characteristics using ANOVA and χ^2 tests. The likelihood of remaining in the study at different stages is tested with a series of logistic regressions. Second, a series of logistic regressions with donation behavior as depend variable and independent variables entered at three stages (demographics + digital skills + individual differences) is conducted. Model fits are compared using analysis of variance.

Results

Drop-out and likelihood of participation

To answer RQ1a and RQ1b, Figure 2 is included to give an overview of how many individuals decided to actually participate and donate their social media data. This figure gives an overview of the number of participants and drop-out rates per research stage. Participants dropped out from the study at different stages and drop-out rate was highest at the donation stage (with 72% of survey completes not donating their data) and the survey stage (with 39% of participants who gave their consent for participation failing to complete the survey). Table 1 shows average age and gender distribution among all invited respondents who dropped out of the study at the different stages. Individuals who completed the survey, but did not donate were significantly younger than earlier drop-outs, while respondents who donated the data

were even younger ($F(5, 1172) = 37.26$, $p < .001$). Regarding gender, more men dropped out due to not using Facebook or not completing the survey compared to other stages ($F(5, 1172) = 6.65$, $p < .001$). Geographical location (urban vs. rural) or level of education did not differ between individuals dropping out at the different stages of the data donation study.

Regarding likelihood of remaining in the study at the different stages (RQ1a), Table 2 shows detailed results of logistic regression. Higher age was weakly associated with a decrease in the likelihood of starting the study indicating that one year increase in age would decrease the odds of starting the study by 2.70% (*OR*=-0.03, 95% CI [-0.04, -0.02]), while gender was moderately associated the likelihood of starting the study so that women had 30.51% lower odds of participating (*OR*=-0.36, 95% CI [-0.62, -0.10]). Regarding likelihood of consenting, similarly higher age was weakly associated with a decrease in the likelihood of providing informed consent indicating that one year increase in age would decrease the odds of consenting by 3.16% (*OR*=-0.03, 95% CI [-0.04, -0.02]), while gender was moderately associated the likelihood of starting the study so that women had 51.45% lower odds of participating (*OR*=-0.72, 95% CI [-1.04, -0.41]). For consent, urbanization was weakly associated with a decrease in the likelihood of consenting indicating that living in a more urbanized area would decrease the odds of consenting by 12.73% (*OR*=-0.14, 95% CI [-0.25, -0.02]). Regarding the likelihood of completing the survey, higher age was weakly associated with a decrease in the likelihood of completing the survey indicating that one year increase in age would decrease the odds of completion by 3.74% (*OR*=-0.04, 95% CI [-0.05, -0.03]), while gender was moderately associated the likelihood of starting the study so that women had 48.48% higher odds of completing the survey (*OR*=0.40, 95% CI [0.02, 0.77]).

Regarding final completion rate, it can be concluded that 289 invited panel members completed the survey (response rate: 24.53%) and out of those, only 80 participants actually donated their social media data (donation rate: 6.79% among all invited panel members and 27.68% among survey participants).

Figure 2: Overall sample size per research stage.

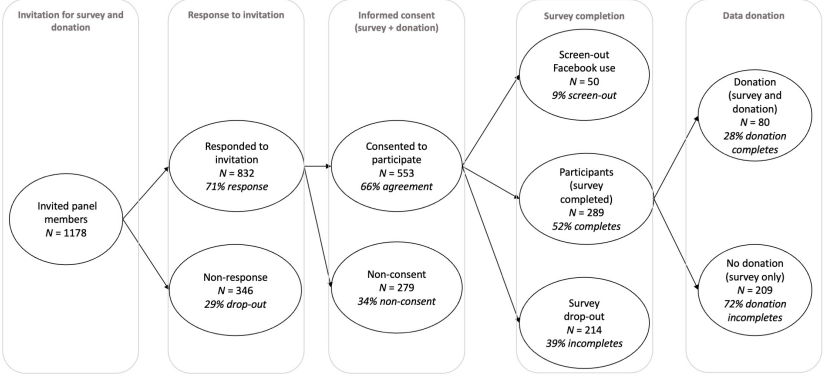


Table 1: Age, gender and education levels distribution among participants who dropped out at different stages (N = 1178)

Stage	Group (N)	*M Age (SD)*	% female	Education		
				Low	Medium	High
Response to invitation	Non-response (346)	68(12)[a]	50%[a]	29%	41%	30%
Informed consent	No consent (279)	67(13)[a]	53%[a]	33%	35%	32%
	Screen-out Facebook use (50)	71(11)[a]	20%[b]	18%	30%	52%
Survey completion	Survey drop-out (214)	65(14)[a]	35%[b]	30%	39%	31%
	Survey only (209)	60(15)[b]	47%[a]	26%	38%	36%
Data donation	Survey and donation (80)	48(13)[c]	49%[a]	27%	38%	35%

Note: numbers in the same column with different superscript differ significantly at least at p <.05

Table 2: Results for logistic regression analyses for participation in different stages of the study

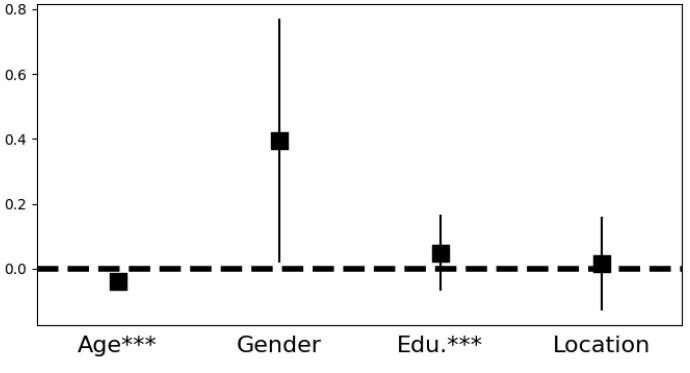
Dependent variable (sample) and predictor	Starting study (all invited, N = 1178)	Consent (all who started study, N = 832)	Survey completion (all consented, N = 553)
Age	-0.03***	-0.03***	-0.04***
Gender	-0.36	-0.72***	0.40*
Level of education	0.01	0.05	0.05
Location	-0.05	-0.13*	0.02
	LL(4, 1160) = -685.88, p<.001	*LL(4, 820) = -498.47, p <.001*	*LL(4, 543) = -350.88 , p <.001*
Nagelkerke R2	.02	.05	.07

Note. The table presents odds ratios. *$p < .05$. **$p < .01$. ***$p < .001$

Predictors of donation behavior

To answer RQ2, we investigated the relation between the donation behavior for all respondents (N = 832) and socio-demographic characteristics. The logistic regression model was statistically significant, *Log-likelihood* (4,820) = -209.83, $p < .001$. The model explained 19.43% (*Nagelkerke* R^2) of the variance in donation behavior. The location (e.g., city, rural, etc.) and gender were not associated with donating data, but higher age was weakly associated with a decrease in the likelihood of donating indicating that one year increase in age would decrease the odds of donating by 5.87% (*OR*=-0.06, 95% CI [-0.08, -0.04]), while higher level of education was moderately associated with an increase in the likelihood of donating indicating that increase in education level would result in 57% increase in the odds of donating (*OR*=0.45, 95% CI [0.25, 0.65], see Figure 3a).

Figure 3: Relation between demographic variables and donation behavior for all respondents (*N* = 832). Odds ratios.



(a) Note: *$p < .05$. **$p < .01$. ***$p < .001$.

RQ3 focuses on dispositional variables. Survey respondents who donated data (compared to respondents who completed the survey, but did not donate their data), have on average more knowledge about algorithmic processes and higher digital literacy skills. The two groups do not differ significantly when it comes to their concerns (Table 3).

Furthermore, we investigated to what extent survey respondents who donated data differed from respondents who did do so when it comes to their trust. No significant differences were concluded for the two trust variables (Table 4).

Table 3: Skills and concerns distribution among survey participants (*N* = 286)

Group	No donation	Donation
M Income (8 categories) (SD)	4.05(1.50)^{A}	4.76(1.42)^{B}
M Self-reported knowledge about algorithms (SD)	2.93(1.64)^{A}	3.53(1.53)^{B}
M Algorithmic awareness in social media (SD)	4.17(1.70)^{A}	5.22(1.23)^{B}
M Internet self-efficacy (SD)	5.33(1.25)^{A}	6.08(0.96)^{B}
M Digital self-efficacy (SD)	3.94(1.52)^{A}	4.83(1.48)^{B}
M Privacy concern (SD)	4.69(1.56)	4.55(1.18)

Note: numbers in the same row with different superscript differ significantly at least at $p < .05$

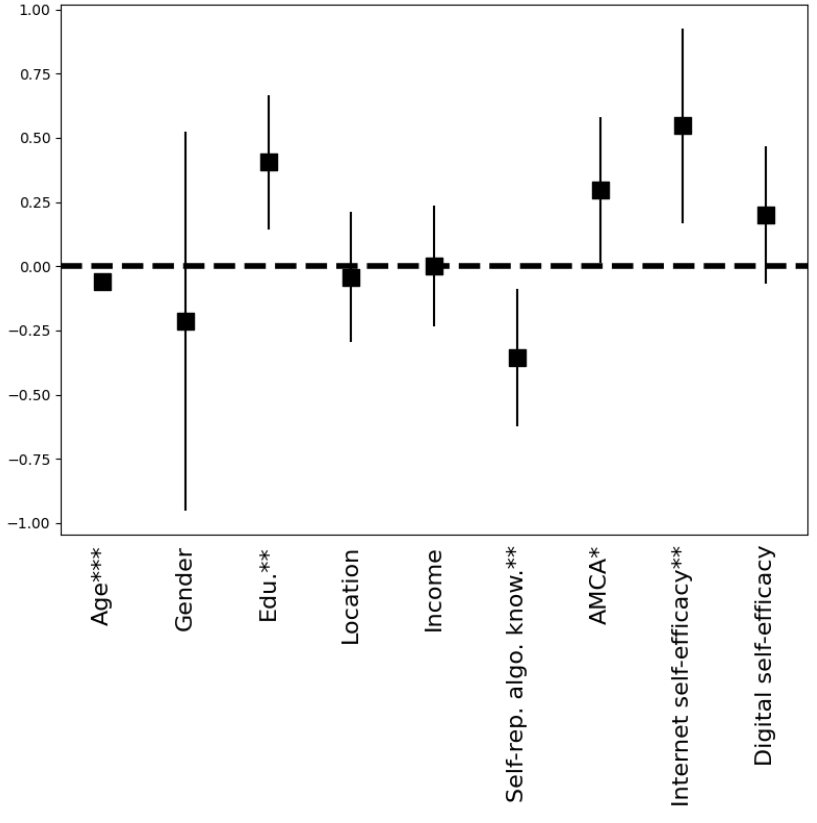
Table 4: Trust distribution among survey participants (*N* = 286)

Group	No donation	Donation
M Trust in online companies (SD)	3.76(1.20)	3.89(0.98)
M Trust in the Web (SD)	3.31(1.29)	3.48(1.08)

Finally, to investigate the relation between the donation behavior for survey participants (N = 286) and dispositional variables, a series of hierarchical logistic regressions was conducted. In the analysis steps, first, only demographic variables were included in the regression. In the second step, skills-related variables were added. In the final step, privacy- and trust-related variables were added. Table 5 shows the results of the hierarchical logistic regressions. Regarding demographics, the first logistic regression model was statistically significant, *Log-likelihood* (5,280) = -140.42, $p < .001$. The model explained 16.7% (*Nagelkerke* R^2) of the variance in donation behavior. Location and gender were not associated with donating data, but higher age score was associated with a decrease in the likelihood of donating (*OR*=-0.04, $p<.001$, 95%CI [-0.06, -0.02]), while higher level of education was associated with an increase in the likelihood of donating (*OR*=0.46, $p<.001$, 95%CI [0.22, 0.71]). Regarding knowledge-related variables, including them in the analysis significantly improved the regression model ($p<.001$, Table 5, Figure 4a). The model including knowledge-related variables explained 25.9% (Nagelkerke R2) of the variance in donation behavior. Self-reported knowledge about algorithms (*OR*=-0.39, $p=.003$, 95%CI [-0.65, -0.13]) was negatively related to donation behavior, while algorithmic awareness in social media (*OR*=0.32, $p=.024$, 95%CI [0.04, 0.60]) and digital literacy skills (*OR*=0.72, $p<.001$, 95%CI [0.37, 1.07]) were positively related to it. Finally, adding other dispositional variables did not significantly improve the model

(p= .279) as privacy- and trust-related variables were not related to the donation behavior.

Figure 4: Relation between demographic and skill-related variables (Model 2) and donation behavior for survey participants. Odds ratios.



(a) Note: *p < .05. **p < .01. ***p < .001.

Reasons for non-donation

To explore reasons for lack of donation we coded answers to the open question asked to respondents who did not donate their data. Overall, 51 participants provided answers. Out of them, 29% (15 responses) failed to provide a valid reason. Technical issues were the most common reason (40%, 20 responses) which can be divided into general issues (14%, 7 responses), is-

sues with the platform (e.g., errors on Facebook, 12%, 6 responses), user on the user-side (e.g., internet speed, 12%, 6 responses) and issues with the data donation tool (error in the donation tool, 2%, 1 response). Around 14% mentioned lack of knowledge (7 responses) and 14% lack of interest (7 responses). Roughly 3% addressed how much time it takes to complete donation procedure as a reason to stop (2 responses).

Table 5: Results for hierarchical regression analysis for data donation behavior among participants (*N* = 289)

Step and predictor		Step 1	Step 2	Step 3
1.	Age	-0.05***	-0.06***	-0.06***
	Gender	0.19	-0.11	-0.14
	Level of education	0.46***	0.40**	0.41**
	Location	-0.03	-0.04	-0.04
	Income	0.12	0.01	-0.02
2.	Self-reported knowledge about algorithms		-0.39**	-0.34*
	Algorithmic awareness in social media		0.32*	0.37*
	Internet self-efficacy		0.55**	0.55**
	Digital self-efficacy		0.20	0.25
3.	Privacy concerns			-0.20
	Trust in online companies			0.07
	Trust in the web			-0.30
		LL(5, 280) = -140.42, p<.001	*LL(9, 276) = -124., p <.001*	*LL(12, 273) = -122.40 , p <.001*
Nagelkerke R2		.17	.26	.27
∆ Nagelkerke R2			.09***	.01

Note. The table presents odds ratios. $^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$.

Discussion

The aim of this study was to examine the representativeness and validity of data collected through donations, focusing on potential selection bias that accumulates with different types of nonresponses. The study utilized a probability sample to investigate who donated their social media data and who dropped out at different stages of data collection. Building on past research, demographic characteristics, digital self-efficacy, and privacy- and trust-related variables were investigated as potential drivers of donation non-response.

Overall, results from this data donation study showed lower donation rate than the intentions to donate reported in past research on collection of digital traces (e.g., Keusch et al., 2019; Kmetty et al., 2024). In fact, we observed high drop-out at all stages of the research in which participants received information on the donation. The most comprehensive information on donation was provided in the informed consent (which for example included a list of the specific files individuals will be asked to request and upload). At this stage, 33.53% of respondents did not consent to participate in the study. This unusually large non-consent rate could potentially be attributed to the information on data donation provided in the informed consent (especially when comparing to other studies that utilized the same panel for survey research and observed minimal drop out at consent). Interestingly, drop-out at this stage is also significantly different from respondents stopping at later stages and similar to the non-response group in terms of demographics. Kmetty et al. (2024) suggested that detailed information about the different types of data included in the DDPs may decrease willingness to participate in donation studies, which could explain high number of non-consents. However, we would argue that transparency needs to be at the core of data donation studies (Araujo et al., 2022). As past research suggests that respondents need to trust that their data will be treated safely and not shared further (Keusch et al., 2019), instead of not providing the information transparently, underlining security measures could possibly lower the drop-out at this stage.

Regarding the type of individuals not included in the data donation sample, the results suggest that higher age and lacking skills are the main drivers of non-donation behavior. The results suggest that data donation behavior decreases with age – participants who donate are even younger than those who decide to participate in the study and follow the remaining procedure. This indicates that when collecting data through donations, a part of the population is missed. Older individuals lacking digital self-efficacy has been

identified as vulnerable in previous studies (Bol et al., 2018; Zarouali, Helberger, & De Vreese, 2021). The fact that those vulnerable groups are left out when collecting digital traces through donations is particularly problematic for studies that focus on digital divides and impact of algorithmic communication on vulnerable groups in the digital society. At the same time, dispositional characteristics related to privacy and trust were found not to be related to the donation behavior. In fact, adding them to the model did not improve it showing their lack of predictive power. This contradicts past research on collection of digital traces where privacy and security concerns were generally seen as one of the main drivers of non-willingness to participate (Keusch et al., 2019; Struminskaya et al., 2020) and actual donation behavior (Ohme et al., 2021). This can be potentially explained by the set-up of the donation website in the current study. Ohme et al. (2021) suggested to counter the effect of privacy concerns by giving respondents more agency when donating data through "selective donations" (p. 307), i.e., giving them the choice not only weather to donate data, but also which parts of the data to donate. This was implemented in the current study and might have offset the effect of privacy concerns and hence is recommended for future data donation studies. Similarly, trust played a smaller role than in other studies on self-disclosure and sharing digital traces. While trust drives data sharing with commercial organizations (Bol et al., 2018; McKnight et al., 2002; Metzger, 2006), it is not related to donating to academic research in the current study. In the invitation, the informed consent, and the donation instructions the respondents were informed about the academic character of the research. Past studies have shown that trust that the research organization will not share the data further (Keusch et al., 2019) and that it is generally a trustworthy research organization increased the intention to participate in studies involving digital traces. As the current research was conducted by a large public university, this might explain the negligent relation with trust. At the same time, we can only conclude the lack of relation between dispositional factors and donation behavior only for participants who completed the survey as they filled in questions on privacy concerns and trust. Participants who dropped out earlier are not included in this analysis. It is hence possible that individuals with high concerns and low trust dropped out at earlier stages (e.g., not providing their consent).

Interestingly, higher self-reported knowledge about algorithms lowered the odds of donating social media data. This effect is opposite to both Internet self-efficacy and algorithmic awareness. It could potentially be explained by individual imaginaries (Duffy & Chan, 2019). When individuals

try to make sense of working of algorithms on e.g., social media, they might use folk theories. Folk theories are ideas that people use to explain the effects of technological systems (DeVito et al., 2017), such as making sense of surveillance by media technologies (Zhang et al., 2024) or algorithmic profiling (Büchi et al., 2023). Folk theories are not necessarily based on facts but are formed based on people's personal experiences and beliefs (Gelman & Legare, 2011). While digital literacy skills and algorithmic awareness are specific and measure skills and knowledge, the self-reported knowledge might rather reflect one's confidence in own ideas. Further research is needed to explore the role of understanding algorithmic processes and where this understanding comes from in participating in data donation studies.

Something we did not measure in our study but, as the answers to open question about reasons not to donate data suggest, might impact data donation behavior as well as research in which data donation is applied, is people's mobile data access and plans. Even though the internet penetration rate in the Netherlands is very high (about 98 percent of the Dutch households has internet access) (Netherlands, 2019), downloading one's data and uploading it to the research environment can require much of the participants' mobile data. Especially for people with lower income levels and restricted mobile internet access (e.g., certain gigabyte limit per month) this is something that researchers should be aware of as a potential barrier to data donation behavior, as well as something that participants should be informed about before taking part in a data donation study. Exploration of the open answers given when participants reported problems with uploading or downloading their data confirmed that five participants explicitly mentioned that uploading or downloading the files took too long, or that downloading "remain stuck" (e.g. "I am already waiting for 15 minutes and Facebook keeps notifying 'in progress' "or "internet connection is too slow due to bad WIFI-connection"). This suggests that a potentially vulnerable group of interest in communication science research that is able to participate in online surveys will be left out from donation samples. This could be countered by collecting digital traces at university labs instead of asking respondents to download and upload DDPs at home.

The current study carries further implications for the design of collection of data donations. Exploration of open answers regarding reasons for not donating data indicated several technical difficulties (12 x e.g. "there was an error message", "password was not accepted") and lack of understanding among respondents (9 times, "e.g., "I don't know how it worked despite the explanation" or "I usually have help with this from my children", "I

found it too difficult, too many steps", "I am not sure how to do this, so rather not do it"). These insights can be used to target underrepresented groups with specific information especially if lack of participation in data donation studies is driven by low Internet self-efficacy as suggested by the results. Providing help when donating data (e.g., through a phone help desk) can improve the representativeness of donation samples and subsequently validity of conclusions drawn from data donation studies. Furthermore, we see an opportunity for improvement in formulating instructions. While the current study provided visual step-by-step instructions on downloading and donating DDPs, future research could experiment with the information provided and engage in co-creation with underrepresented respondent groups to make sure they receive the necessary information. Alternatively, on-location data collection suggested before would allow for immediate help with technical issues.

While offering important insights into representativeness of data donation samples, the current study has multiple limitations. While it presents a comprehensive overview of drop-out from data donation studies at different stages that allows for future improvements to data donation studies, it only focuses on one platform, namely Facebook. Although Facebook is widely used and only 50 respondents were excluded from the study due to non-use, the donation behavior might be different when other or more platforms are involved. Kmetty et al. (2024) showed that in Hungary, the more platforms were included in the request the less likely respondents would be to participate in the study. At the same time, as Internet self-efficacy and technical difficulties were the main driver of non-donation, other platforms might present less of different difficulties to respondents when downloading DDPs. It hence requires further research to conclude to what extent the current findings generalize to different platforms. Furthermore, the current study focuses on characteristics of (non-)respondents and not on characteristics of study design. In order to conclude how data donation studies can deal with the impact of lack of digital self-efficacy and demographic characteristics, future research should focus on testing most effective donation procedures through experimentation. As suggested by past research, adjusting incentives (Skatova & Goulding, 2019) might help to include underrepresented groups. Finally, we cannot rule out that the set-up of the study itself might have had an impact on the donation behavior. In the survey, the respondents were first asked questions about, among others, their privacy concerns and trust and later, they were asked to donate. While these dispositional variables did not lower the likelihood of donating, these questions might

have driven high survey drop-out as participants already knew they would be asked to donate their data. Alternative designs ad question order effects for data donation studies hence deserve further exploration.

To conclude, the current research compliments past studies on willingness to participate in collection of digital traces by focusing on who donates their social media data in academic research and empirically investigating the profile of participants and systematically comparing them to non-participants. The results carry both bad and good news. Groups identified as vulnerable in past research are also missing in donation samples as age and skills are the main drivers of not donating. At the same time, it is due to skills and not due to concerns or lack of trust that individuals decide not to donate to academic research. This offers possibilities for improvement in study design so that all willing individuals are equally able to participate.

References

Acquisti, A., Brandimarte, L., & Loewenstein, G. (2015). Privacy and human behavior in the age of information. *Science*, *347*(6221), 509–514.

Anderson, M., & Kumar, M. (2019). Digital divide persists even as lower-income americans make gains in tech adoption.

Araujo, T., Ausloos, J., van Atteveldt, W., Loecherbach, F., Moeller, J., Ohme, J., Trilling, D., van de Velde, B., De Vreese, C., & Welbers, K. (2022). Osd2f: An open-source data donation framework. *Computational Communication Research*, *4*(2), 372–387.

Baek, T. H., & Morimoto, M. (2012). Stay away from me. *Journal of advertising*, *41*(1), 59–76.

Barnes, S. B. (2006). A privacy paradox: Social networking in the united states. *First Monday*.

Baruh, L., Secinti, E., & Cemalcilar, Z. (2017). Online privacy concerns and privacy management: A meta-analytical review. *Journal of Communication*, *67*(1), 26–53.

Beer, D. (2009). Power through the algorithm? participatory web cultures and the technological unconscious. *New media & society*, *11*(6), 985–1002.

Bietz, M., Patrick, K., & Bloss, C. (2019). Data donation as a model for citizen science health research. *Citizen Science: Theory and Practice*, *4*(1).

Boeschoten, L., Ausloos, J., Möller, J. E., Araujo, T., & Oberski, D. L. (2022). A framework for privacy preserving digital trace data collection through data donation. *Computational Communication Research*, *4*(2), 388–423.

Bol, N., Dienlin, T., Kruikemeier, S., Sax, M., Boerman, S. C., Strycharz, J., Helberger, N., & De Vreese, C. H. (2018). Understanding the effects of personalization as a privacy calculus: Analyzing self-disclosure across health, news, and commerce contexts. *Journal of Computer-Mediated Communication*, *23*(6), 370–388.

Bol, N., Strycharz, J., Helberger, N., van de Velde, B., & de Vreese, C. H. (2020). Vulnerability in a tracked society: Combining tracking and survey data to understand who gets targeted with what content. *New Media & Society, 22*(11), 1996–2017.

Büchi, M., Fosch-Villaronga, E., Lutz, C., Tamò-Larrieux, A., & Velidi, S. (2023). Making sense of algorithmic profiling: User perceptions on facebook. *Information, Communication & Society, 26*(4), 809–825.

CBS. (2019a). Opleidingsniveau. https://www.cbs.nl/nl-nl/nieuws/2019/33/verschil-levensverwachting-hoog-en-laagopgeleid-groeit/opleidingsniveau

CBS. (2019b). Stedelijkheid. https://www.cbs.nl/nl-nl/nieuws/2019/44/meeste-afval-per-inwoner-in-minst-stedelijke-gemeenten/stedelijkheidu

Conner, M., & Norman, P. (2022). Understanding the intention-behavior gap: The role of intention strength. *Frontiers in Psychology, 13*, 923464.

de Vries, D., Piotrowski, J., de Vreese, C., et al. (2022). Resultaten onderzoek digitale competenties (digcom).

DeVito, M. A., Birnholtz, J., & Hancock, J. T. (2017). Platforms, people, and perception: Using affordances to understand self-presentation on social media. *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*, 740–754.

Duffy, B. E., & Chan, N. K. (2019). "you never really know who's looking": Imagined surveillance across social media platforms. *New Media & Society, 21*(1), 119–138.

Eastin, M. S., & LaRose, R. (2000). Internet self-efficacy and the psychology of the digital divide. *Journal of computer-mediated communication, 6*(1), JCMC611.

Eurostat. (2022). How many citizens had basic digital skills in 2021? https://ec.europa.eu/eurostat/web/products-eurostat-news/-/ddn-20220330-1

Gelman, S. A., & Legare, C. H. (2011). Concepts and folk theories. *Annual review of anthropology, 40*, 379–398.

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.

Jürgens, P., Stark, B., & Magin, M. (2020). Two half-truths make a whole? on bias in self-reports and tracking data. *Social Science Computer Review, 38*(5), 600–615.

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.

Kmetty, Z., Stefkovics, Á., Számely, J., Deng, D., Kellner, A., Pauló, E., Omodei, E., & Koltai, J. (2024). Determinants of willingness to donate data from social media platforms. *Information, Communication & Society*, 1–26.

Kruikemeier, S., Boerman, S. C., & Bol, N. (2020). Breaching the contract? using social contract theory to explain individuals' online behavior to safeguard privacy. *Media Psychology, 23*(2), 269–292.

Lazer, D., Hargittai, E., Freelon, D., Gonzalez-Bailon, S., Munger, K., Ognyanova, K., & Radford, J. (2021). Meaningful measures of human society in the twenty-first century. *Nature*, *595*(7866), 189–196.

Malhotra, N. K., Kim, S. S., & Agarwal, J. (2004). Internet users' information privacy concerns (iuipc): The construct, the scale, and a causal model. *Information systems research*, *15*(4), 336–355.

McKnight, D. H., Choudhury, V., & Kacmar, C. (2002). Developing and validating trust measures for e-commerce: An integrative typology. *Information systems research*, *13*(3), 334–359.

Metzger, M. J. (2006). Effects of site, vendor, and consumer characteristics on web site trust and disclosure. *communication research*, *33*(3), 155–179.

Netherlands, S. (2019). Internet. https://longreads.cbs.nl/european-scale-2019/internet/

Non, M., Dinkova, M., Dahmen, B., et al. (2021). *Skill up or get left behind?: Digital skills and labor market outcomes in the netherlands* (tech. rep.). CPB Netherlands Bureau for Economic Policy Analysis.

Ohme, J., Araujo, T., de Vreese, C. H., & Piotrowski, J. T. (2021). Mobile data donations: Assessing self-report accuracy and sample biases with the ios screen time function. *Mobile Media & Communication*, *9*(2), 293–313.

Reeves, B., Ram, N., Robinson, T. N., Cummings, J. J., Giles, C. L., Pan, J., Chiatti, A., Cho, M., Roehrick, K., Yang, X., et al. (2021). Screenomics: A framework to capture and analyze personal life experiences and the ways that technology shapes them. *Human–Computer Interaction*, *36*(2), 150–201.

Sala, E., Knies, G., & Burton, J. (2014). Propensity to consent to data linkage: Experimental evidence on the role of three survey design features in a uk longitudinal panel. *International Journal of Social Research Methodology*, *17*(5), 455–473.

Scheerder, A., Van Deursen, A., & Van Dijk, J. (2017). Determinants of internet skills, uses and outcomes. a systematic review of the second-and third-level digital divide. *Telematics and informatics*, *34*(8), 1607–1624.

Seltzer, E., Goldshear, J., Guntuku, S. C., Grande, D., Asch, D. A., Klinger, E. V., & Merchant, R. M. (2019). Patients' willingness to share digital health and non-health data for research: A cross-sectional study. *BMC medical informatics and decision making*, *19*, 1–8.

Silva, D. E., Chen, C., & Zhu, Y. (2024). Facets of algorithmic literacy: Information, experience, and individual factors predict attitudes toward algorithmic systems. *New Media & Society*, *26*(5), 2992–3017.

Skatova, A., & Goulding, J. (2019). Psychology of personal data donation. *PloS one*, *14*(11), e0224240.

Stier, S., Breuer, J., Siegers, P., & Thorson, K. (2020). Integrating survey data and digital trace data: Key issues in developing an emerging field.

Struminskaya, B., Toepoel, V., Lugtig, P., Haan, M., Luiten, A., & Schouten, B. (2020). Understanding willingness to share smartphone-sensor data. *Public Opinion Quarterly*, *84*(3), 725–759.

Strycharz, J., Ausloos, J., & Helberger, N. (2020). Data protection or data frustration? individual perceptions and attitudes towards the gdpr. *Eur. Data Prot. L. Rev.*, *6*, 407.

van Driel, I. I., Giachanou, A., Pouwels, J. L., Boeschoten, L., Beyens, I., & Valkenburg, P. M. (2022). Promises and pitfalls of social media data donations. *Communication Methods and Measures*, *16*(4), 266–282.

Vorderer, P., Hefner, D., Reinecke, L., & Klimmt, C. (2017). *Permanently online, permanently connected*. Routledge, Taylor & Francis Group London.

Walrave, M., Poels, K., Antheunis, M. L., Van den Broeck, E., & van Noort, G. (2018). Like or dislike? adolescents' responses to personalized social network site advertising. *Journal of marketing communications*, *24*(6), 599–616.

Zarouali, B., Boerman, S. C., & de Vreese, C. H. (2021). Is this recommended by an algorithm? the development and validation of the algorithmic media content awareness scale (amca-scale). *Telematics and Informatics*, *62*, 101607.

Zarouali, B., Helberger, N., & De Vreese, C. H. (2021). Investigating algorithmic misconceptions in a media context: Source of a new digital divide? *Media and Communication*, *9*(4), 134–144.

Zarouali, B., Strycharz, J., Helberger, N., & de Vreese, C. (2022). Exploring people's perceptions and support of data-driven technology in times of covid-19: The role of trust, risk, and privacy concerns. *Behaviour & Information Technology*, *41*(10), 2049–2060.

Zhang, D., Boerman, S. C., Hendriks, H., van der Goot, M. J., Araujo, T., & Voorveld, H. (2024). "they know everything": Folk theories, thoughts, and feelings about dataveillance in media technologies. *International Journal of Communication*, *18*, 21.