UNIVERSITY OF MANNHEIM
School of Social Sciences

Collecting Digital Trace Data via Data Donation

Florian Keusch
University of Mannheim

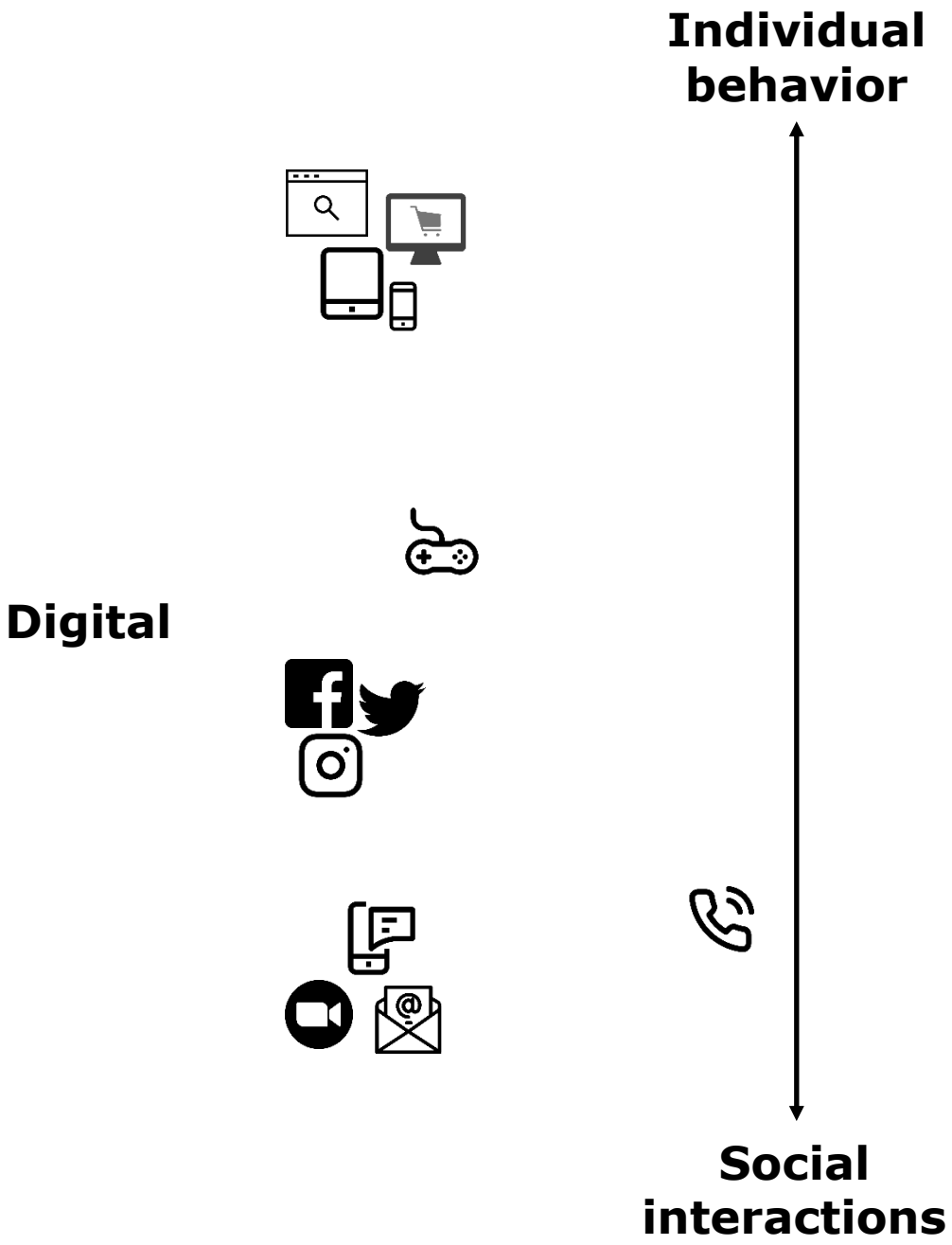
f.keusch@uni-mannheim.de
floriankeusch.weebly.com

Future Directions for Social and Behavioral Science Methodologies in the Next Decade: A Workshop
September 25–26, 2024
Washington, DC

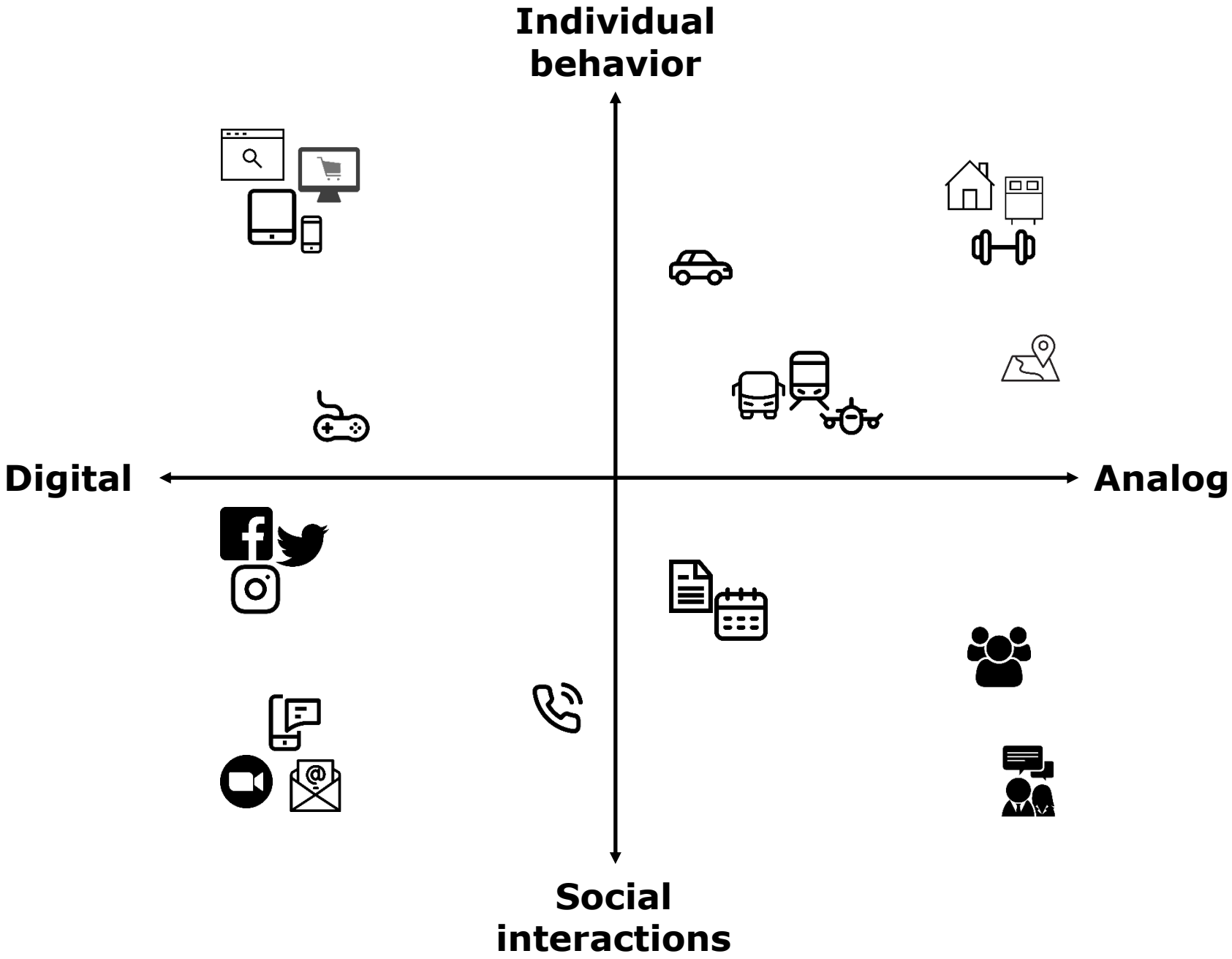
What Are Digital Traces?

"Records of activity (trace data) undertaken through an online information system (thus, digital)"
(Howison et al. 2011:769)

"Behavioral residue [individuals leave] when they interact online"
(Hinds & Joinson 2018:2)



Keusch, F. & Kreuter, F. (2022)



Keusch, F. & Kreuter, F. (2022)

Benefits of Digital Trace Data

- Allow measurement of behavior (in-the-moment) at high frequency
  - Evaluation of moment-to-moment changes
  - Without increasing burden on participants
  - Scalable

Benefits of Digital Trace Data

- Allow measurement of behavior (in-the-moment) at high frequency
  - Evaluation of moment-to-moment changes
  - Without increasing burden on participants
  - Scalable

- Measurement is nonreactive, i.e., without direct solicitation of subject studied
  - Digital trace data should be unaffected by measurement itself
  - Reduced measurement error when measuring…
    - smartphone use (Kobayashi & Boase 2012; Boase & Ling 2013; Andrews, et al. 2015; De Reuver & Bouwman 2015; Revilla et al. 2017; Jones-Jang et al. 2020)
    - online media consumption (Araujo et al. 2017; Haenschen 2020; Revilla et al. 2017; Scharkow 2016)
    - mobility (Stopher et al. 2007; Scherpenzeel 2017)

How to Collect Individual-Level Digital Traces

- Use APIs

How to Collect Individual-Level Digital Traces

- Use APIs

Meta

An Update on Our Plans to Restrict Data Access on Facebook

April 4, 2018

Chesnot / Getty Images

How to Collect Individual-Level Digital Traces

- Use APIs

Meta

An Update on Our Plans to Restrict Data Access on Facebook

April 4, 2018

Chesnot / Getty Images

SCIENCE / TWITTER / TECH

Twitter just closed the book on academic research / Twitter was once an indispensable resource for academic research. That's changed under Elon Musk.

By Justine Calma, a science reporter covering the environment, climate, and energy with a decade of experience. She is also the host of the Hell or High Water podcast.

May 31, 2023, 3:19 PM GMT+2 | 23 Comments / 23 New

https://www.theatlantic.com/technology/archive/2023/02/elon-musk-twitter-ethics-algorithm-biases/673110/

How to Collect Individual-Level Digital Traces

- Use APIs

Meta

An Update on Our Plans to Restrict Data Access on Facebook

April 4, 2018

Chesnot / Getty Images

SCIENCE / TWITTER / TECH

Twitter just closed the book on academic research / Twitter was once an indispensable resource for academic research. That's changed under Elon Musk.

By Justine Calma, a science reporter covering the environment, climate, and energy with a decade of experience. She is also the host of the Hell or High Water podcast.

May 31, 2023, 3:19 PM GMT+2 | 23 Comments / 23 New

https://www.theatlantic.com/technology/archive/2023/02/elon-musk-twitter-ethics-algorithm-biases/673110/

FEATURE

Reddit pricing: API charge explained

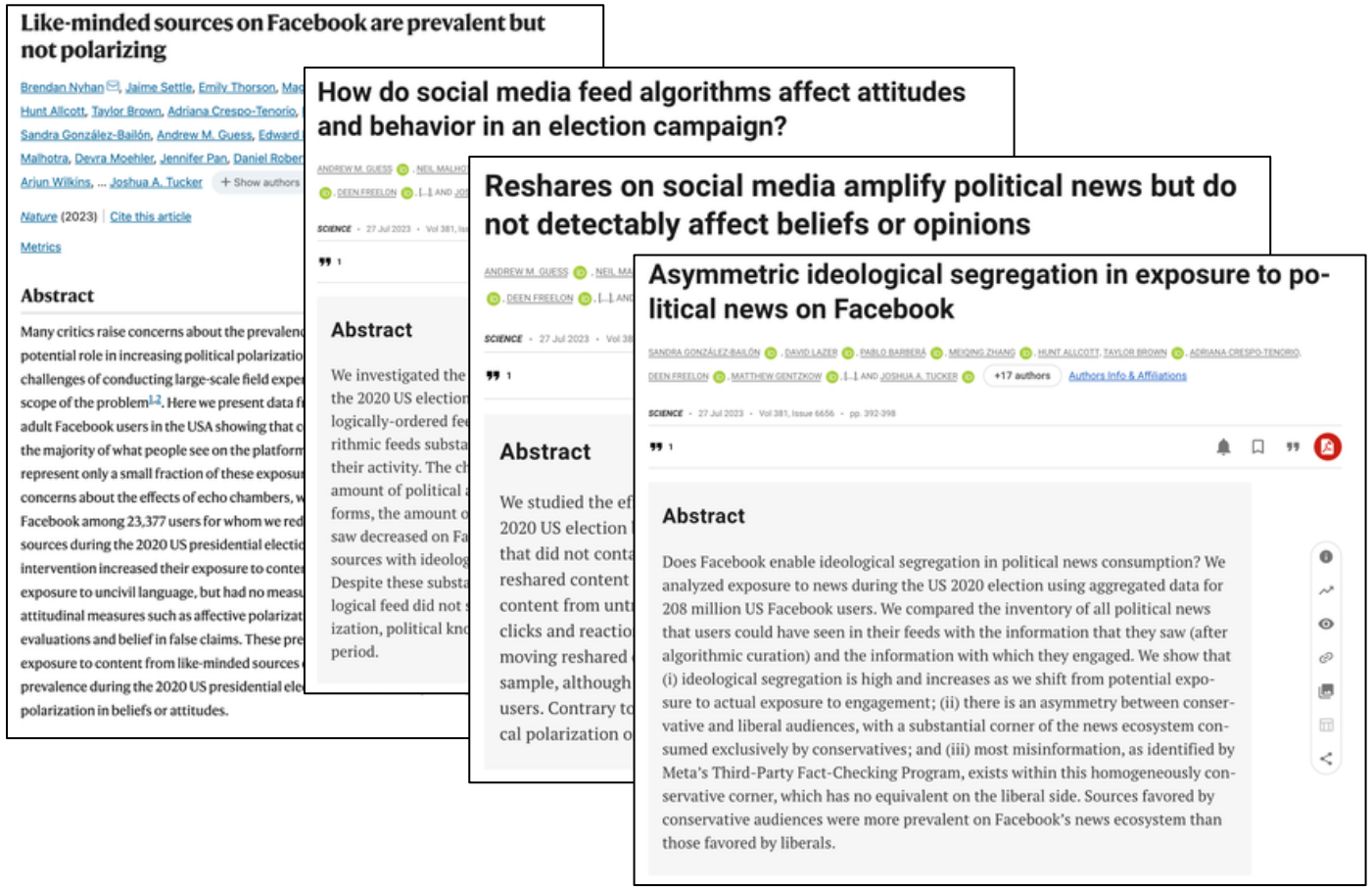
Providing a free API was becoming costly for Reddit. Enterprise-scale developers now have to pay for access to Reddit's data.

By Ben Lutkevich, Technical Features Writer Published: 11 Jul 2023

Shutterstock

How to Collect Individual-Level Digital Traces

- Use APIs
- Collaborate with industry



How to Collect Individual-Level Digital Traces

- Use APIs
- Collaborate with industry

How to Collect Individual-Level Digital Traces

- Use APIs
- Collaborate with industry
- Have users install meters and apps that continuously…
  - …track information on web browsing (URLs, HTML code, screen scraping)
  - …logs usage behavior, (native) mobile browsing, and sensor readings

How to Collect Individual-Level Digital Traces

- Use APIs
- Collaborate with industry
- Have users install meters and apps that continuously…
  - …track information on web browsing (URLs, HTML code, screen scraping)
  - …logs usage behavior, (native) mobile browsing, and sensor readings

Advantages

+ Access to fine-grained, prospective data
+ Relatively easy to combine with self-reports (Keusch & Conrad 2023)
  - e.g., EMAs

How to Collect Individual-Level Digital Traces

- Use APIs
- Collaborate with industry
- Have users install meters and apps that continuously…
  - …track information on web browsing (URLs, HTML code, screen scraping)
  - …logs usage behavior, (native) mobile browsing, and sensor readings

Advantages

+ Access to fine-grained, prospective data
+ Relatively easy to combine with self-reports (Keusch & Conrad 2023)
  - e.g., EMAs

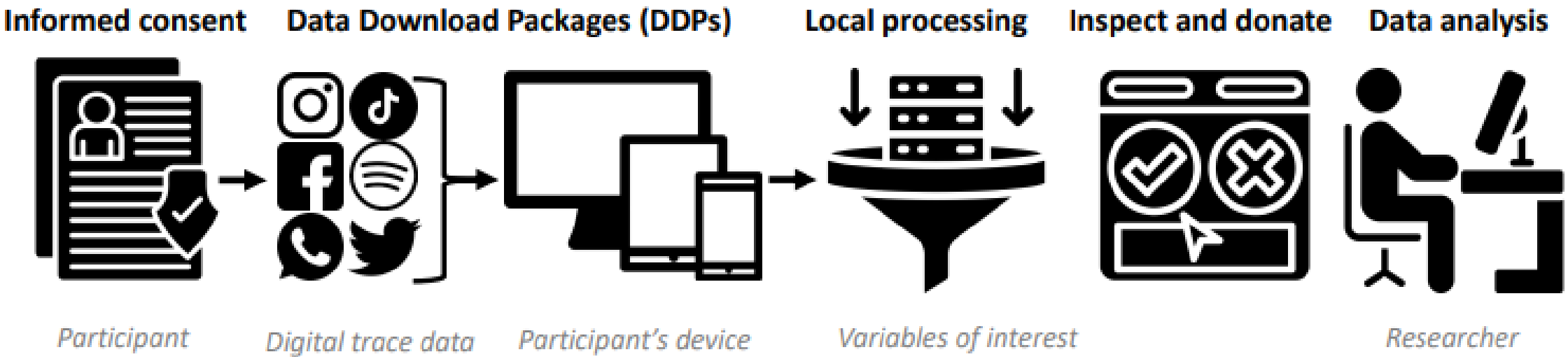
Challenges

- Coverage & nonparticipation (Bosch et al. 2024; Keusch et al. 2022, 2023)
- Ethics & privacy concerns (Keusch et al. 2019; Revilla et al. 2019; Struminskaya et al. 2020, 2021; Wenz et al. 2019; Wenz & Keusch 2023)
- Missing data (Bähr et al. 2022)
- Analytical challenges

How to Collect Individual-Level Digital Traces

- Use APIs
- Collaborate with industry
- Have users install meters and apps
- Ask users to "donate" data
  - Takes advantage of legislation such as GDPR Articles 15 (*Right of access by the data subject*) and 20 (*Right to data portability*)
    - See also California Consumer Privacy Act (CCPA)

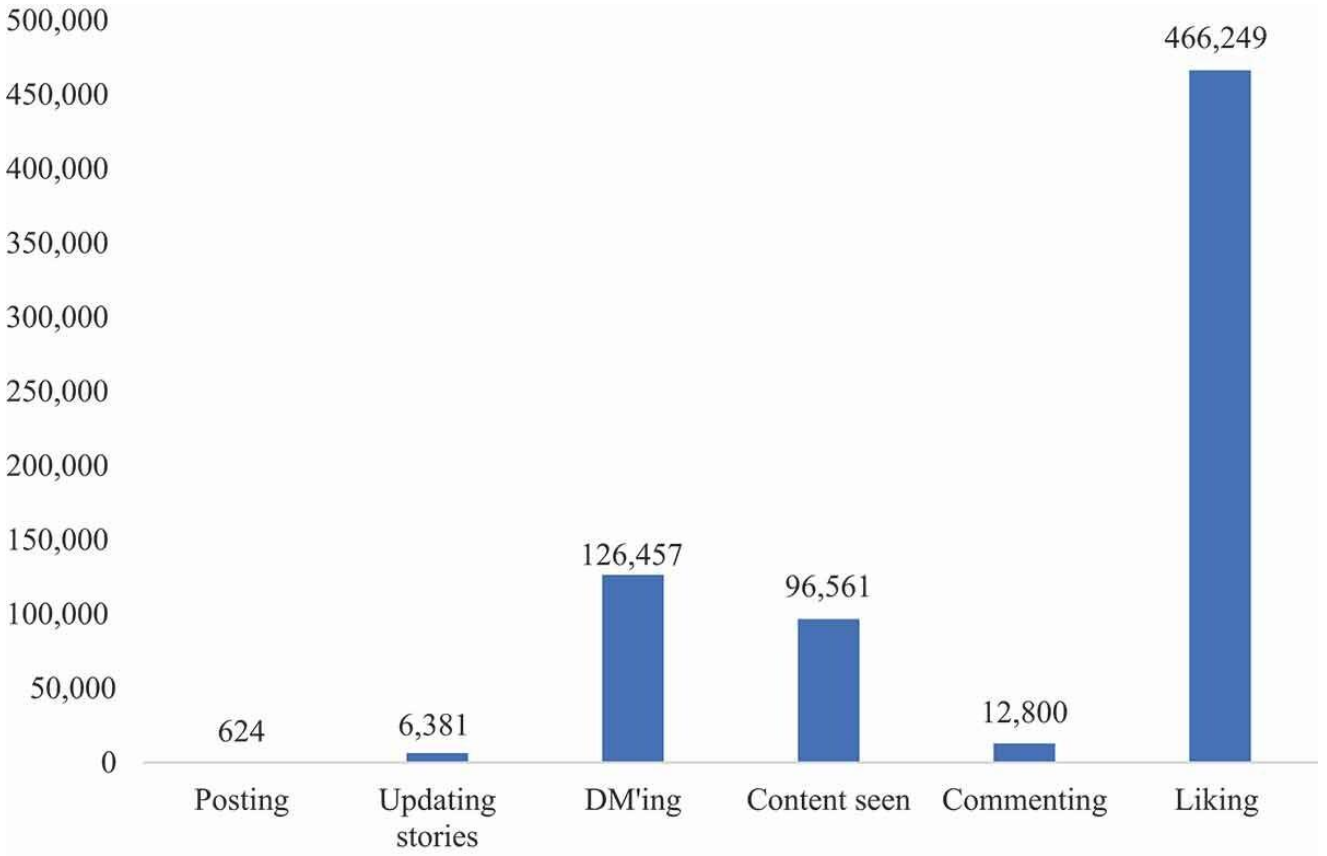
Privacy-preserving Data Donation Platforms



Boeschoten et al. (2023)

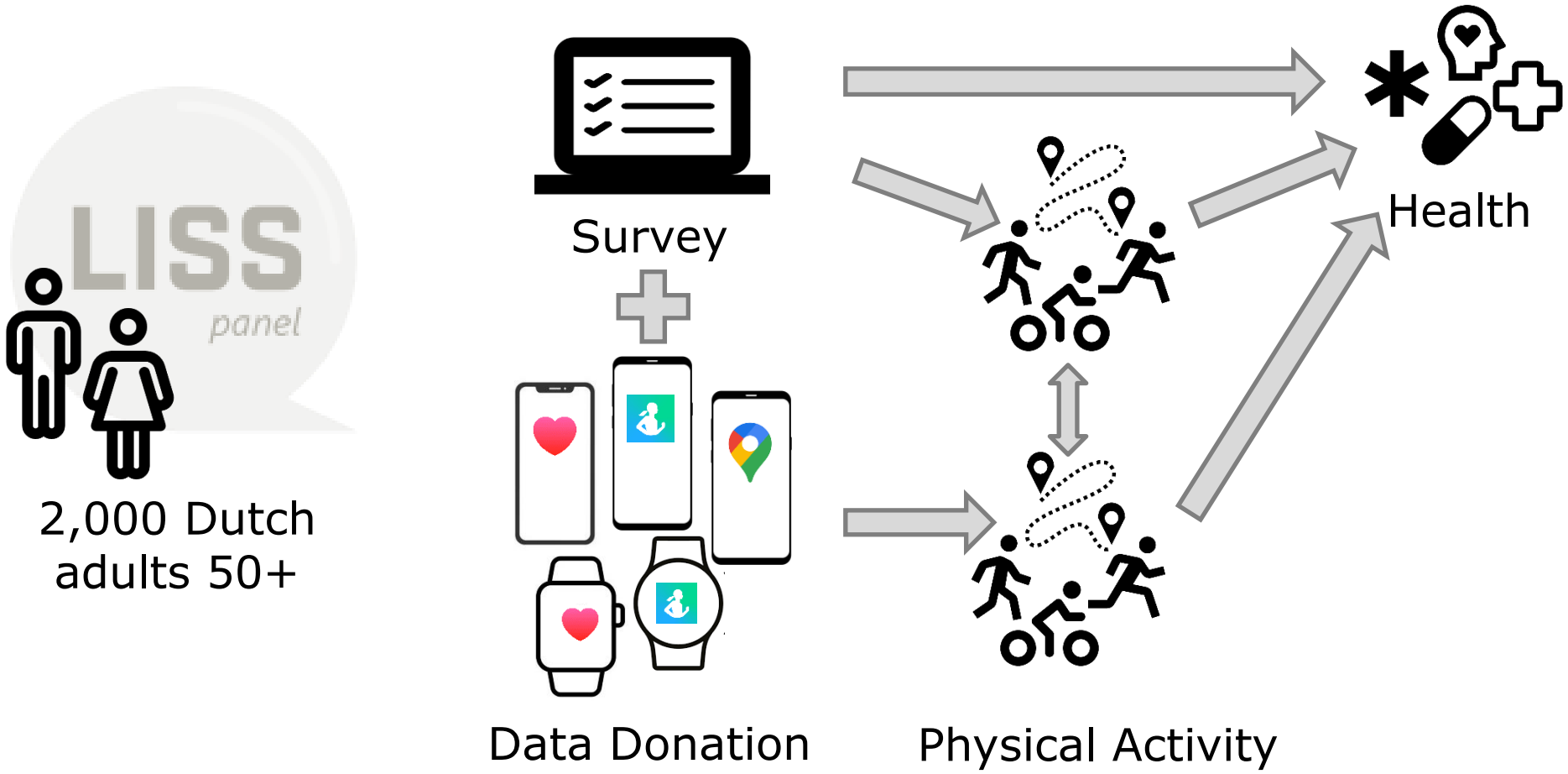
Example Studies

- Instagram activities among 8th and 9th graders (van Driel et al. 2022)



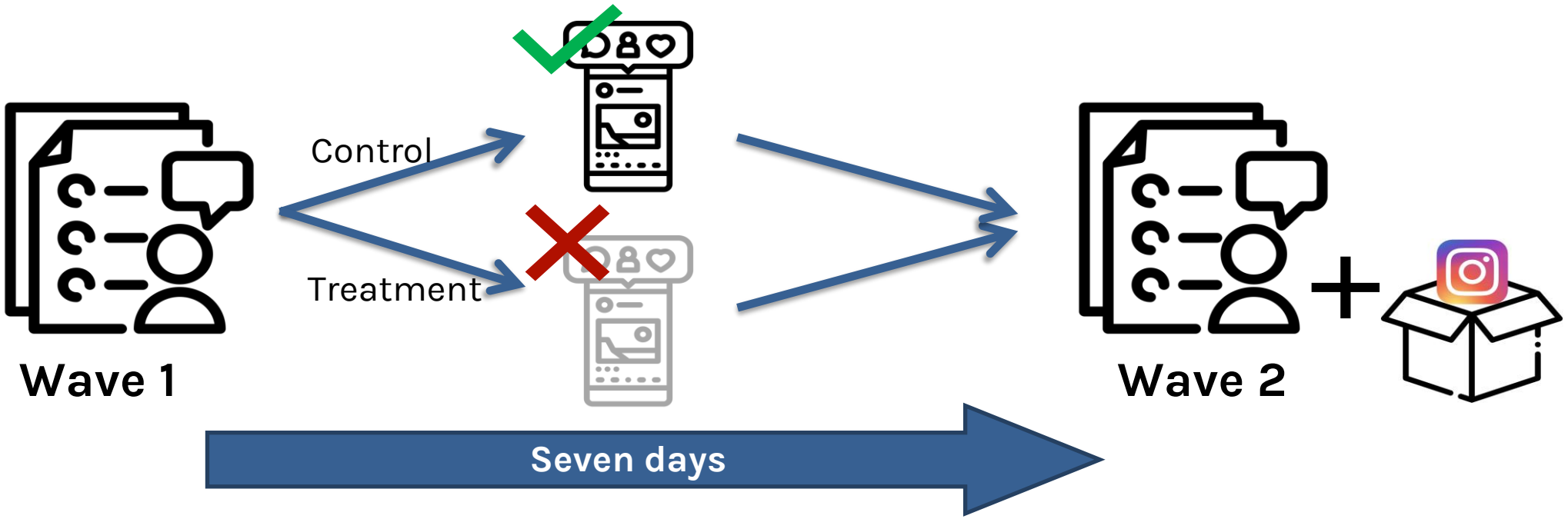
Example Studies

- Instagram activities among 8th and 9th graders (Driel et al. 2022)
- Physical activity of older adults (Keusch & Struminskaya in preparation)



Example Studies

- Instagram activities among 8th and 9th graders (Driel et al. 2022)
- Physical activity of older adults (Keusch & Struminskaya in preparation)
- Effect of Instagram abstinence on body image (Szafran et al. in preparation)



Data Donation

Advantages

+ Allows access to digital trace data that cannot be collected otherwise
  - e.g., https://www.facebook.com/dyi

+ Works for large range of platforms
  - WhatsApp, Facebook, Instagram, Google, LinkedIn, YouTube, Netflix, Apple Health, Fitbit,…

+ User retains full control over what data are donated

Data Donation

Advantages

+ Allows access to digital trace data that cannot be collected otherwise
  - e.g., https://www.facebook.com/dyi

+ Works for large range of platforms
  - WhatsApp, Facebook, Instagram, Google, LinkedIn, YouTube, Netflix, Apple Health, Fitbit,…

+ User retains full control over what data are donated

Challenges

- Data donation process rather cumbersome for users (consent, data download, data upload)

- Technical skills needed to participate

What Do We Know So Far?

- Users report relativly high willingness to donate data (30-75%)
  (Hase & Haim 2024; Keusch et al. 2024; Ohme et al. 2021; Pfiffner and Friemel 2023; Silber et al. 2022; van Driel et al. 2022)
  - Main reason for not being willing: privacy concerns (Keusch et al. 2024)

- Large share (up to 50%) drop out during data donation process
  (Hase & Haim 2024; Keusch et al. 2024; Ohme et al. 2021; Silber et al. 2022; van Driel et al. 2022)
  - Cumbersome process that requires some technical skills

- So far, few studies have assessed nonparticipation bias
  - Some evidence for higher social media use (Hase & Haim 2024; but not Keusch et al. 2024), political interest (Hase & Haim 2024), and "healthy donor" bias (Keusch et al. 2024)

Potential Developments for Methodological Research on Data Donation

- Communicate idea of data donation to participants
  - e.g., stressing data privacy advantages for participants
- Effective integration of data donation into survey infrastructure
  - At what point to ask for donation, how to frame, how to incentivize
- Platforms/data controllers keep changing structure of and access to DDPs (Hase et al. 2024)
  - Constant monitoring (by researchers) and/or standardization (by legislators)
- Unpacking full potential of digital trace data
  - In combination with self-reports, e.g., measurement models (Cernat et al. 2024)

Thank You!

Florian Keusch
University of Mannheim
School of Social Sciences
Social Data Science & Methodology
f.keusch@uni-mannheim.de
http://floriankeusch.weebly.com/
@floriankeusch

References

Andrews, Sally, David A. Ellis, Heather Shaw, and Lukasz Piwek. 2015. "Beyond self-report: Tools to compare estimated and real-world smartphone use." *PLOS ONE* 10(10):e0139004. https://doi.org/10.1371/journal.pone.0139004

Araujo, Theo, Anke Wonneberger, Peter Neijens, and Claes de Vreese. 2017. "How much time do you spend online? Understanding and improving the accuracy of self-reported measures of internet use." *Communication Methods and Measures*, 11:173-90. https://doi.org/10.1080/19312458.2017.1317337

Bähr, Sebastian, Georg-Christoph Haas, Florian Keusch, Frauke Kreuter, and Mark Trappmann. 2022. "Missing data and other measurement quality issues in mobile geolocation sensor data." *Social Science Computer Review* 40:212-35. https://doi.org/10.1177/0894439320944118

Boase, Jeffrey, and Rich Ling. 2013. "Measuring mobile phone use: Self-report versus log data." *Journal of Computer-Mediated Communication* 18:508-19. https://doi.org/10.1111/jcc4.12021

Boeschoten, Laura, Niek C. de Schipper, Adriënne M. Mendrik, Emiel van der Veen, Bella Struminskaya, Heleen Janssen, and Theo Araujo. "Port: A software tool for digital data donation." *Journal of Open Source Software* 8(90):5596. https://doi.org/10.21105/joss.05596

Bosch, Oriol J., Patrick Sturgis, Jouni Kuha, and Melanie Revilla. 2024. "Uncovering digital trace data biases: Tracking undercoverage in web tracking data." *Communication Methods and Measures*. https://doi.org/10.1080/19312458.2024.2393165

de Reuver, Mark, and Harry Bouwman. 2015. "Dealing with self-report bias in mobile Internet acceptance and usage studies." *Information & Management* 52:287-94. https://doi.org/10.1016/j.im.2014.12.002

Haenschen, Katherine. 2020. "Self-reported versus digitally recorded: Measuring political activity on Facebook." *Social Science Computer Review* 38:567-83. https://doi.org/10.1177/0894439318813586

Hase, Valerie and Mario Haim. 2024. "Can we get rid of bias? Mitigating systematic error in data donation studies through survey design strategies." *Computational Communication Research* 6(2). https://doi.org/10.5117/CCR2024.2.2.HASE

Hase, Valerie, Jef Ausloos, Laura Boeschoten, Nico Pfiffner, Heleen Janssen, Theo Araujo, Thijs Carrière, Claes de Vreese, Jörg Haßler, Felicia Loecherbach, Zoltán Kmetty, Judith Möller, Jakob Ohme, Elisabeth Schmidbauer, Bella Struminskaya, Damian Trilling, Kasper Welbers, and Mario Haim. 2024. "Fulfilling data access obligations: How could (and should) platforms facilitate data donation studies?" *Internet Policy Review* 13(3). https://doi.org/10.14763/2024.3.1793

Hinds, Joanne, and Adam N. Joinson. 2018. "What demographic attributes do our digital footprints reveal? A systematic review." *PLOS ONE* 13(11):e0207112. https://doi.org/10.1371/journal.pone.0207112

Howison, James, Andrea Wiggins, and Kevin Crowston. 2011. "Validity issues in the use of social network analysis with digital trace data." *Journal of the Association for Information Systems* 12:767-97. https://doi.org/10.17705/1jais.00282

Jones-Jang, S Mo, Yu-Jin Heo, Robert McKeever, Jung-Hyun Kim, Leigh Moscowitz, and David Moscowitz. 2020. "Good news! Communication findings may be underestimated: Comparing effect sizes with self-reported and logged smartphone use data." *Journal of Computer-Mediated Communication* 25:346-63. https://doi.org/10.1093/jcmc/zmaa009

Keusch, Florian, Sebastian Bähr, Georg-Christoph Haas, Frauke Kreuter, and Mark Trappmann. 2023. "Coverage error in data collection combining mobile surveys with passive measurement using apps: Data from a German national survey." *Sociological Methods & Research* 52:841-78. https://doi.org/10.1177/0049124120914924

Keusch, Florian, Sebastian Bähr, Georg-Christoph Haas, Frauke Kreuter, Mark Trappmann, and Stephanie Eckman. 2022. "Non-participation in smartphone data collection using research apps." *Journal of the Royal Statistical Society. Series A* 185:S225-45. https://doi.org/10.1111/rssa.12827

Keusch, Florian, and Frederick G Conrad. 2022. "Using smartphones to capture and combine self-reports and passively measured behavior in social research." *Journal of Survey Statistics and Methodology* 10:863-85. https://doi.org/10.1093/jssam/smab035

Keusch, Florian, and Frauke Kreuter. 2022. "Digital trace data. Modes of data collection, applications, and errors at a glance." In *Handbook of Computational Social Science. Volume 1: Theory, Case Studies and Ethics*, edited by Uwe Engel et al., 100-118. Milton Park: Routledge. https://doi.org/10.4324/9781003024583-8

Keusch, Florian, Paulina Pankowska, Alexandru Cernat, and Ruben Luksa Bach. 2024. "Do you have two minutes to talk about your data? Willingness to participate and nonparticipation bias in Facebook data donation. *Field Methods*. https://doi.org/10.1177/1525822X231225907

Keusch, Florian, Bella Struminskaya, Christopher Antoun, Mick P. Couper, and Frauke Kreuter. 2019. "Willingness to participate in passive mobile data collection." *Public Opinion Quarterly* 83:210-35. https://doi.org/10.1093/poq/nfz007

Kobayashi, Tetsuro, Jeffrey Boase, Tsutomu Suzuki, and Takahisa Suzuki. 2015. "Emerging from the cocoon? Revisiting the tele-cocooning hypothesis in the smartphone era." *Journal of Computer-Mediated Communication* 20:330-45. https://doi.org/10.1111/jcc4.12116

Ohme, Jakob, Theo Araujo, Claas de Vreese, and Jessica T. Piotrowski. 2021. "Mobile data donations: Assessing self-report accuracy and sample biases with the iOS Screen Time function." *Mobile Media & Communication*, 9:293-313. https://doi.org/10.1177/2050157920959106

Pfiffner, Nico, and Thomas N. Friemel. 2023. "Leveraging data donations for communication research: Exploring drivers behind the willingness to donate." *Communication Methods and Measures* 17:227-49. https://doi.org/10.1080/19312458.2023.2176474

Revilla, Melanie, Mick P. Couper, and Carlos Ochoa. 2019. "Willingness of online panelists to perform additional tasks." *methods, data, analyses* 13:223-52. https://doi.org/0.12758/mda.2018.01

Revilla, Melanie, Carlos Ochoa, and Germán Loewe. 2017. "Using passive data from a meter to complement survey data in order to study online behavior." *Social Science Computer Review* 35:521-36. https://doi.org/10.1177/0894439316638457

Scharkow, Michael. 2016. "The accuracy of self-reported internet use—a validation study using client log data." *Communication Methods and Measures* 10:13-27. https://doi.org/10.1080/19312458.2015.1118446

Scherpenzeel, Annette. 2017. "Mixing online panel data collection with innovative methods." In *Methodische Probleme von Mixed-Mode-Ansätzen in Der Umfrageforschung*, edited by Stefanie Eifler and Frank Faulbaum, 27-49. Schriftenreihe Der ASI - Arbeitsgemeinschaft Sozialwissenschaftlicher Institute. Wiesbaden: Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-15834-7_2

Silber, Henning, Johannes Breuer, Christoph Beuthner, Tobias Gummer, Florian Keusch, Pascal Siegers, Sebastian Stier, and Bernd Weiß. 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:S387-S407. https://doi.org/10.1111/rssa.12954

Stopher, Peter, Camden FitzGerald, and Min Xu. 2007. "Assessing the accuracy of the Sydney Household Travel Survey with GPS." *Transportation* 34:723-41. https://doi.org/10.1007/s11116-007-9126-8

Struminskaya, Bella, Peter Lugtig, Vera Toepoel, Barry Schouten, Deirdre Giesen, and Ralph Dolmans. 2021. "Sharing data collected with smartphone sensors: Willingness, participation, and nonparticipation bias." *Public Opinion Quarterly* 85:423-62. https://doi.org/10.1093/poq/nfab025

Struminskaya, Bella, Vera Toepoel, Peter Lugtig, Marieke Haan, Annemieke Luiten, and Barry Schouten. 2020. "Understanding willingness to share smartphone-sensor data." *Public Opinion Quarterly* 84:725-59. https://doi.org/10.1093/poq/nfaa044

Wenz, Alexander, Annette Jäckle, and Mick P. Couper. 2019. "Willingness to use mobile technologies for data collection in a probability household panel." *Survey Research Methods* 13:1-22. https://doi.org/10.18148/srm/2019.v1i1.7298

Wenz, Alexander and Florian Keusch. 2023. "Increasing the acceptance of smartphone-based data collection." *Public Opinion Quarterly* 87:357-88. https://doi.org/10.1093/poq/nfad019

van Driel, Irene I., Anastasia Giachanou, J. Loes Pouwels, Laura Boeschoten, Ine Beyens, and Patti M. Valkenburg. 2022. "Promises and pitfalls of social media data donations." *Communication Methods and Measures* 16:266-82. https://doi.org/10.1080/19312458.2022.2109608