

gesis

Leibniz-Institut
für Sozialwissenschaften



Combining survey data and digital behavioral data

GESIS Meet the Experts Series

*Best practice methods in Survey Methodology and
Computational Social Science*

Johannes Breuer & Sebastian Stier, 8 July 2021

Speakers



Dr. Johannes Breuer

- Senior Researcher in the team Data Augmentation, Department Survey Data Curation
- Ph.D. in psychology
- Digital behavioral data, data linking, use and effects of digital media
- Contact: johannes.breuer@gesis.org

Speakers



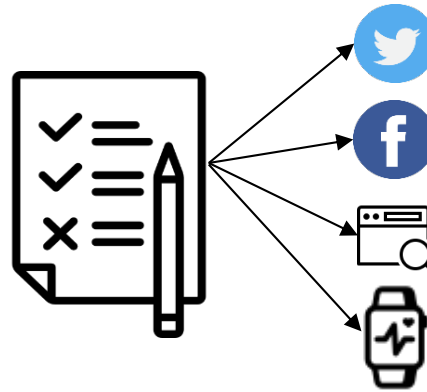
Dr. Sebastian Stier

- Senior Researcher in the team Social Analytics and Services, Department Computational Social Science
- Ph.D. in political science
- Digital behavioral data, political communication, political behavior
- Contact: sebastian.stier@gesis.org

Agenda for today

1. Why?

2. How to?



Combining surveys and
digital behavioral data

3. Challenges

4. Q&A

1. Why combine surveys and digital behavioral data?

Types of digital behavioral data (DBD)



“**Records of activity** (trace data) undertaken through an **online information system** (thus, digital)’ ([Howison et al., 2011](#)) that can be collected from a multitude of technical systems, such as **websites, social media platforms, smartphone apps, or sensors**” ([Stier et al., 2020](#), p. 504)

Digital behavioral data

Strengths

- Direct measures of behavior
- High granularity (temporal resolution)
- High volume and velocity

Limitations

- No information about user characteristics
- No direct measures of attitudes
- No information about offline activities

Survey data

Strengths

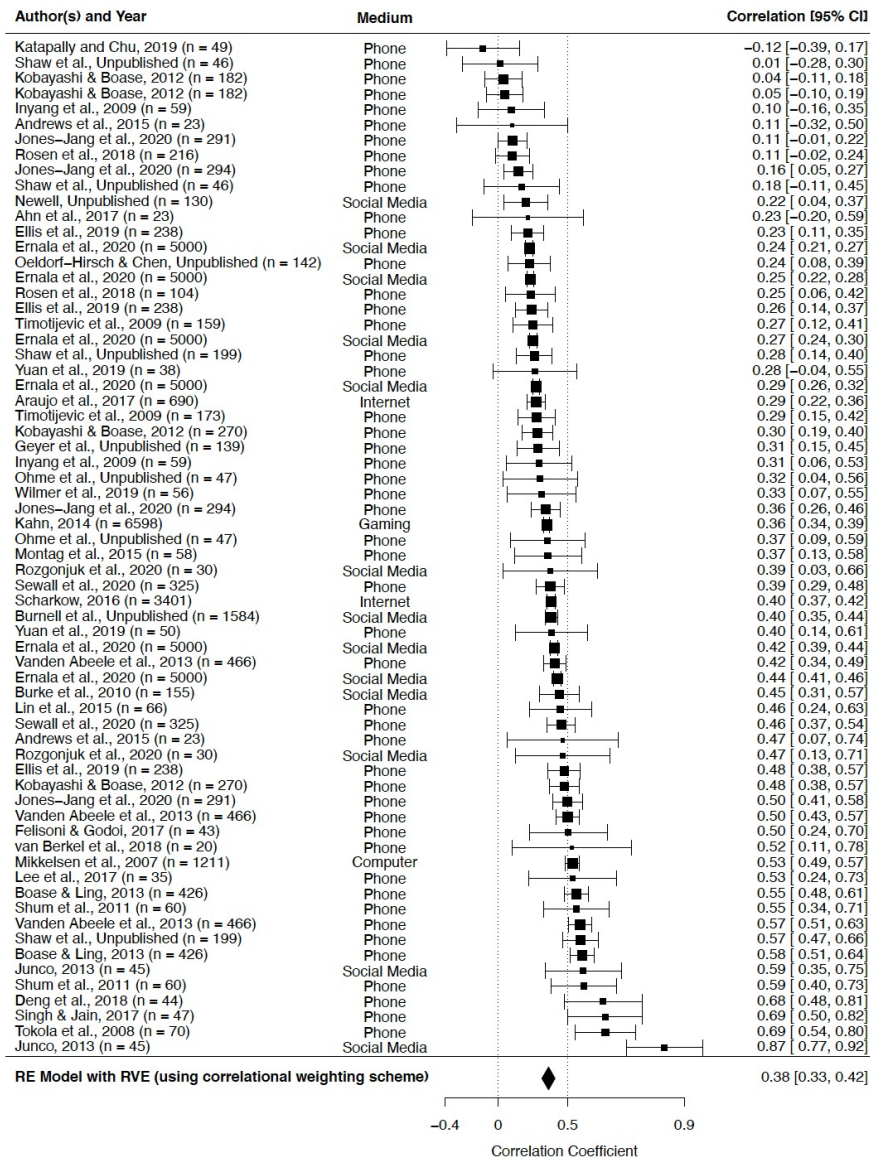
- All types of constructs can be measured, e.g., opinions, attitudes, offline behavior
- Can assess online as well as offline activities
- Probability sampling possible

Limitations

- Self-reports have limited validity
- Self-reports can be biased by social desirability
- Response rates are declining (esp. for telephone surveys)

Example: Do online echo chambers exist?

Inferences from surveys



Article | Published: 17 May 2021

A systematic review and meta-analysis of discrepancies between logged and self-reported digital media use

Douglas A. Parry , Brittany I. Davidson, Craig J. R. Sewall, Jacob T. Fisher, Hannah Mieczkowski & Daniel S. Quintana

Nature Human Behaviour (2021) | [Cite this article](#)

1560 Accesses | 632 Altmetric | [Metrics](#)

„Based on 106 effect sizes, we found that self-reported media use correlates only moderately with logged measurements [...] These findings raise concerns about the validity of findings relying solely on self-reported measures of media use.“ ([Parry et al., 2021](#))

Inferences from DBD

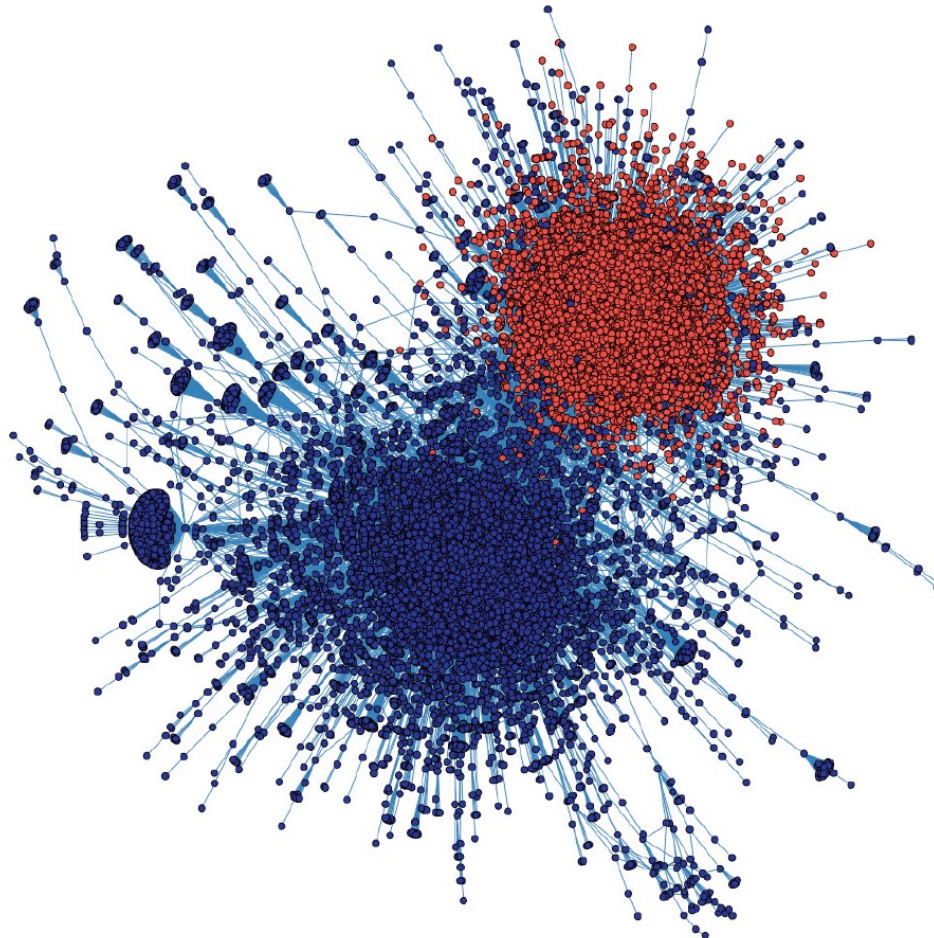
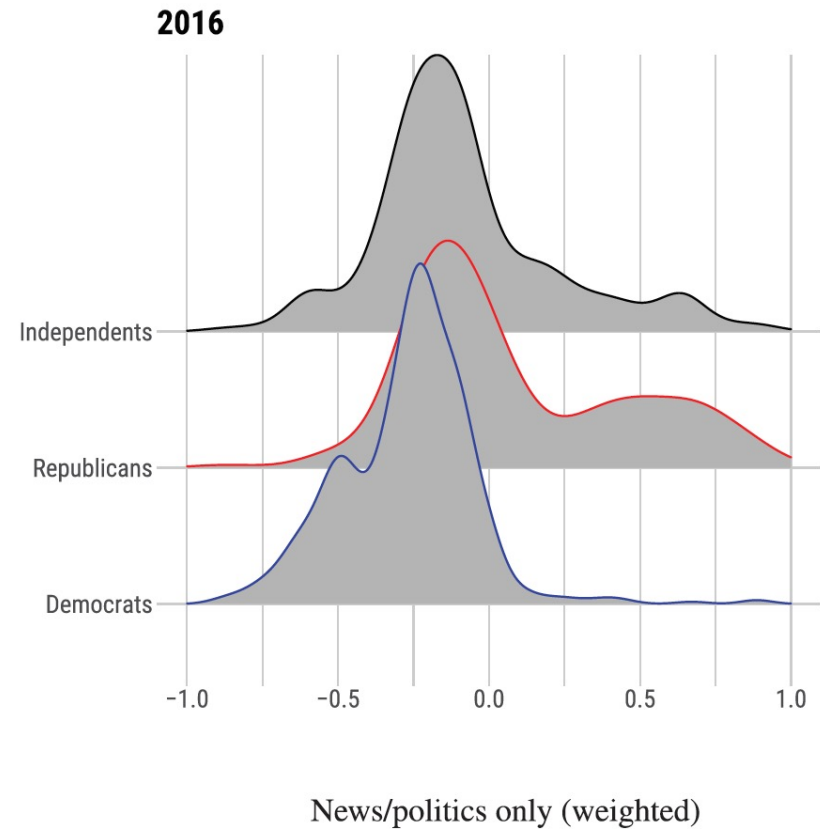
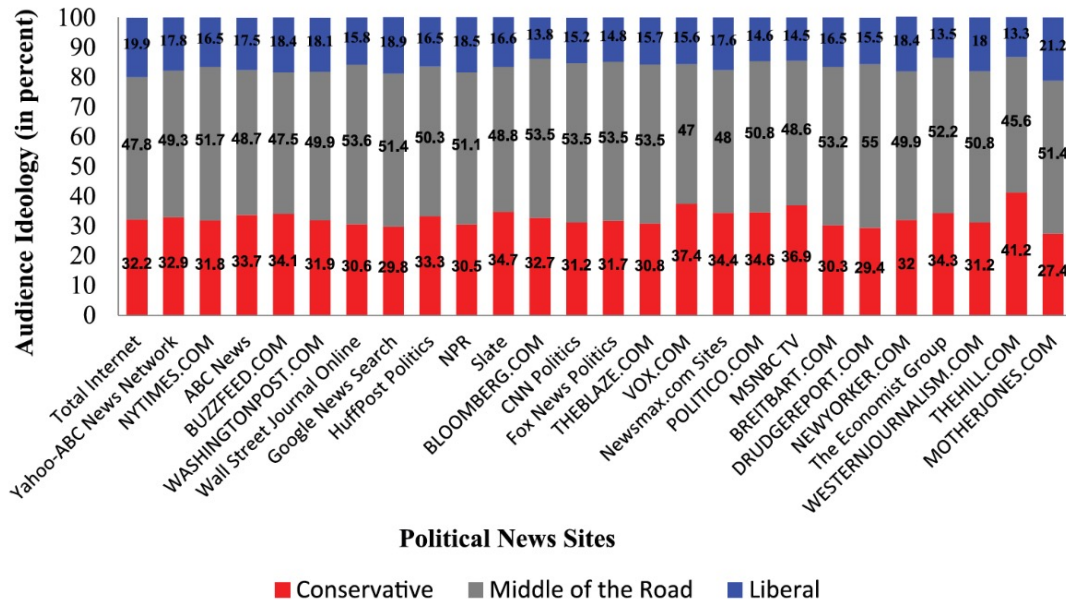


Table 1: Hashtags related to #p2, #tcot, or both. Tweets containing any of these were included in our sample.

Just #p2	#casen #dadt #dc10210 #democrats #dul #fem2 #gotv #kysen #lgf #ofa #onation #p2b #pledge #rebelleft #truthout #vote #vote2010 #whyimvotingdemocrat #youcut
Both	#cspj #dem #dems #desen #gop #hcr #nvsen #obama #ocra #p2 #p21 #phnm #politics #sgp #tcot #teaparty #tlot #topprog #tpp #twisters #votedem
Just #tcot	#912 #ampat #ftrs #glennbeck #hhrs #iamthemob #ma04 #mapoli #palin #palin12 #spwbt #tsot #tweetcongress #ucot #wethepeople

Conover, M., Ratkiewicz, J., Francisco, M., Goncalves, B., Menczer, F., & Flammini, A. (2011). Political Polarization on Twitter. In *Proceedings of the 5th International AAAI Conference on Web and Social Media* (pp. 89–96). AAAI Publications.

When you bring the two together...



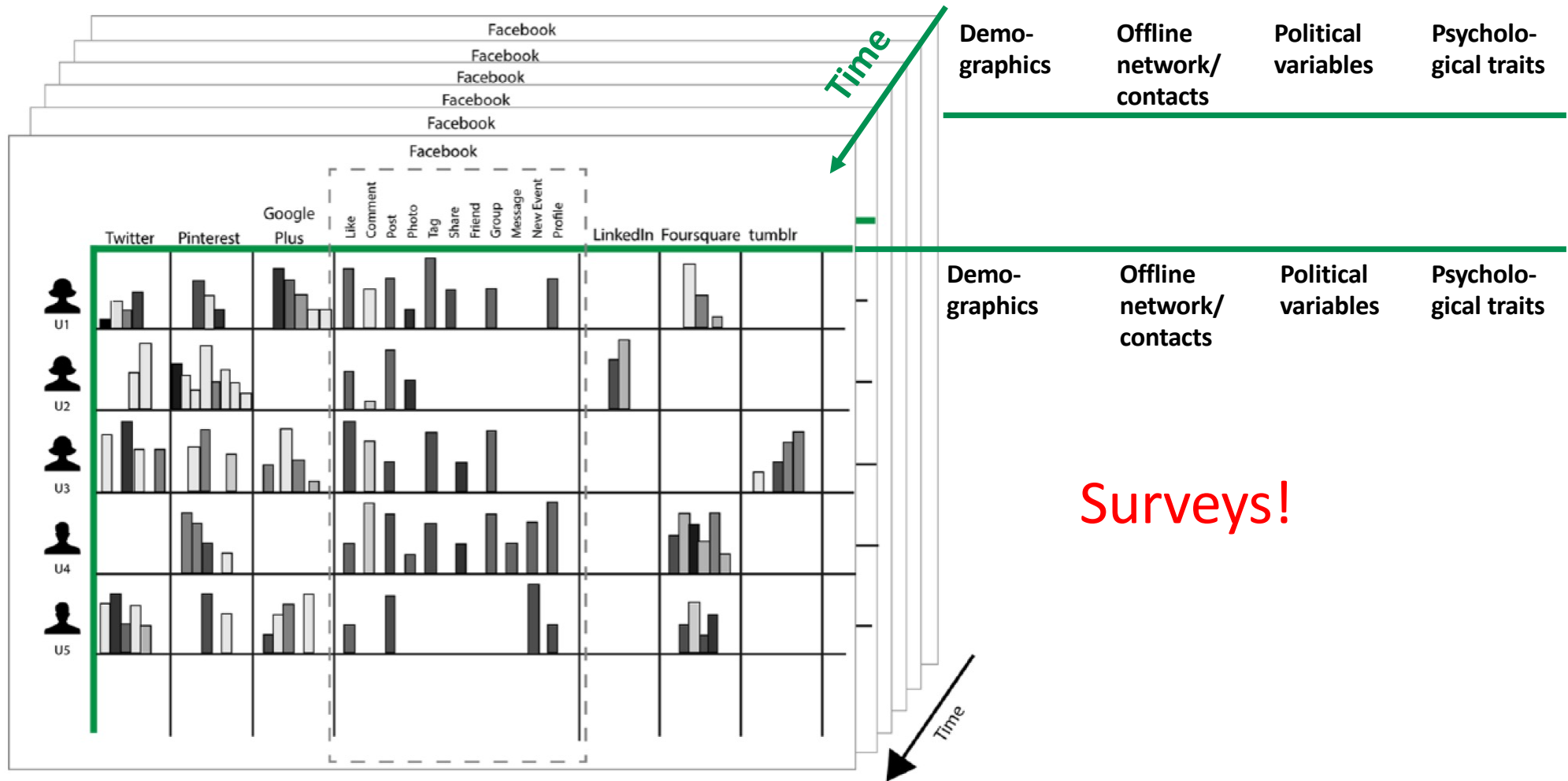
Nelson, J. L., & Webster, J. G. (2017). The Myth of Partisan Selective Exposure: A Portrait of the Online Political News Audience. *Social Media + Society*. <https://doi.org/10.1177/2056305117729314>

Guess, A. M. (2021). (Almost) Everything in Moderation: New Evidence on Americans' Online Media Diets. *American Journal of Political Science*. <https://doi.org/10.1111/ajps.12589>

Linking data from surveys and DBD can be used to combine their unique strengths and to overcome (parts of) their respective limitations.

2. How to combine surveys and digital behavioral data?

Social science in the digital age: the ideal dataset



Resnick, P., Adar, E., & Lampe, C. (2015). What Social Media Data We Are Missing and How to Get It. *The ANNALS of the American Academy of Political and Social Science*, 659(1), 192–206. <https://doi.org/10.1177/0002716215570006>

Ways of linking surveys and DBD

Two possible sequences of data collection



Each option is associated with specific **biases** ([Sen et al., 2021](#))










Ways of linking surveys and DBD

- Two dimensions on which linking approaches can differ:
 1. Overall **temporal sequence** of data collection and linking
 - Data collected specifically for the purpose of linking in the study/project = ex-ante (or direct) linking
 - At least one of the datasets already exists = ex-post linking
 2. **Level of data linking**
 - Individual level (i.e., per user/participant)
 - Aggregated level (e.g., for geographic regions or specific periods of time)

Identifiers

- **Unique identifiers required** for linking surveys with DBD
- User names may be an obvious example (e.g., for social media data)
 - ▶ But risk of failure, e.g., when user names are misreported in surveys
 - Potential solution when starting with/from surveys: Have participants follow or message an account created for the project

Identifiers

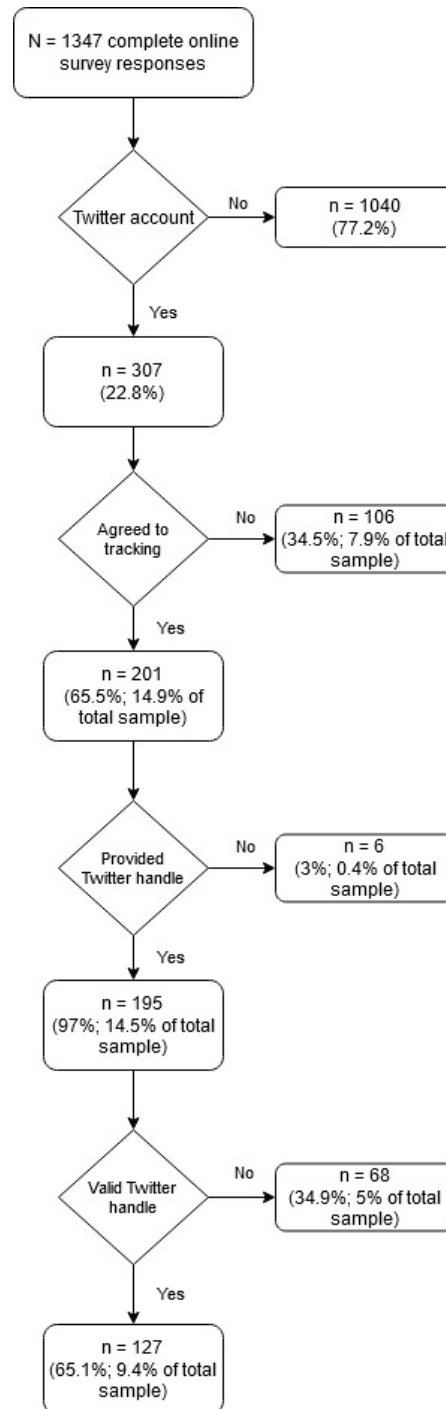
- What identifier(s) to use depends on the type of DBD, how it is collected, and how it is linked with the survey data (see [Breuer et al., 2021](#))
 - ▶ User handle 
 - ▶ Login possible with e-mail address   
 - ▶ User names (typically) real names  
 - ▶ Stable user ID (typically unknown to users)   
- External tools (e.g., browser plugin) may require additional/separate identifiers for linking

3. Challenges in combining surveys and digital behavioral data

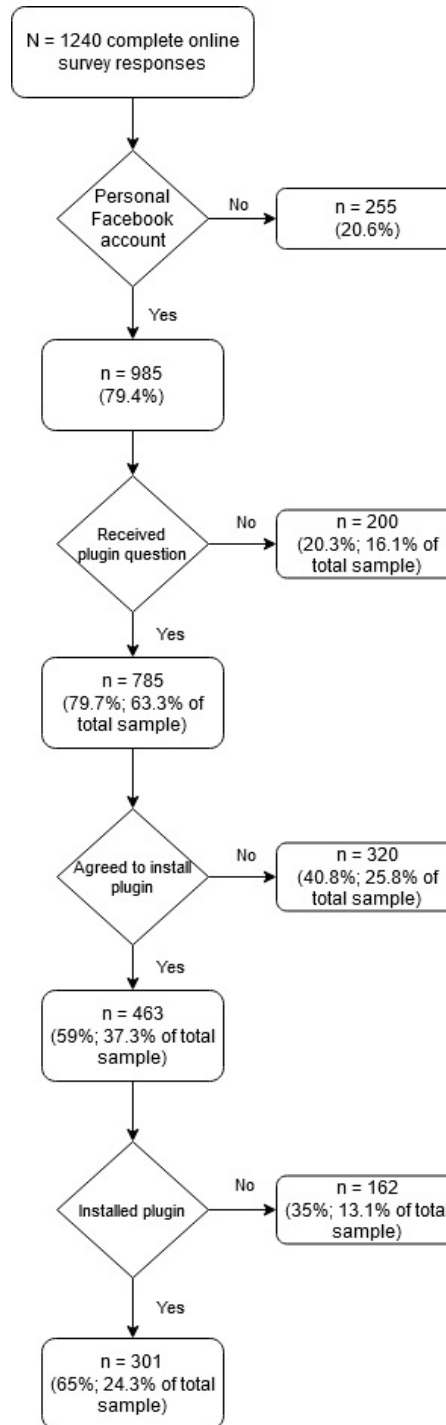
Practical challenges: Willingness to participate

- Web tracking panel with N ~ 2000 participants per month: data from June 2018 to May 2019
- Online surveys including request to link data from:
 - ▶ Twitter (tweets, retweets, etc. via API)
 - ▶ Facebook (public posts in feed via browser plugin for desktop)
 - ▶ Spotify (playlists, plays, etc. via web app)
- Short informed consent in the questionnaire + extended privacy information on GESIS website linked in the short informed consent (see [Breuer et al., 2021](#))

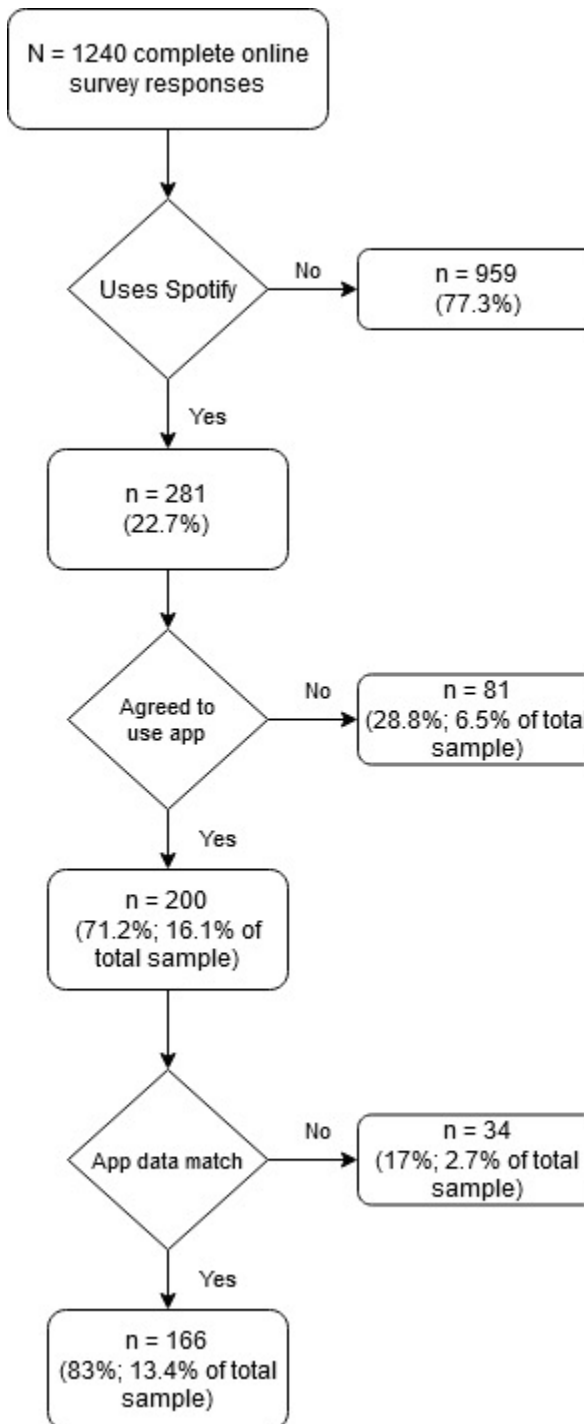
Twitter



Facebook



Spotify



Predictors of willingness to participate

- Age: Younger users more likely to share data
- Use: More frequent users more likely to share
- Respondent burden: For Facebook (browser plugin for desktop) respondents who used desktop for survey more likely to share
- Second study (German non-probability online panel):
 - ▶ Incentive: Higher incentive = more sharing
 - ▶ Respondent burden: Sharing rate much higher for providing screen name vs. data export and upload
 - ▶ Additional factors: Positive survey evaluation & affinity for technology

For details and further results, see [Silber et al. \(2021\)](#)

Legal & ethical issues

- **Contractual agreements:** Do contracts or Terms of Service (ToS) allow linking?
- **Informed consent:** How can it be obtained and what should it look like?
- **Data privacy:** How to deal with the high disclosure risks of linked survey data and DBD?

ToS Example

Twitter Developer Policies

(<https://developer.twitter.com/en/developer-terms/policy>, last accessed 29 June 2021):

“We limit the circumstances under which you may match a person on Twitter to information obtained or stored off-Twitter. Off-Twitter matching involves associating Twitter Content, including a Twitter @handle or user ID, with a person, household, device, browser, or other off-Twitter identifier. You may only do this if you have **express opt-in consent from the person** before making the association, or as described below. In situations in which you don’t have a person’s express, opt-in consent to link their Twitter identity to an off-Twitter identifier, we require that any connection you draw be based only on **information that someone would reasonably expect to be used for that purpose**. In addition, **absent a person’s express opt-in consent** you may only attempt to match your records about someone to a Twitter identity based on:

- ▶ **Information provided directly to you by the person.** Note that records about individuals with whom you have no prior relationship, including data about individuals obtained from third parties, do not meet this standard; and/or
- ▶ **Public data.** “Public data” in this context refers to:
 - Information about a person that you obtained from a public, generally-available resource (such as a directory of members of a professional association)“

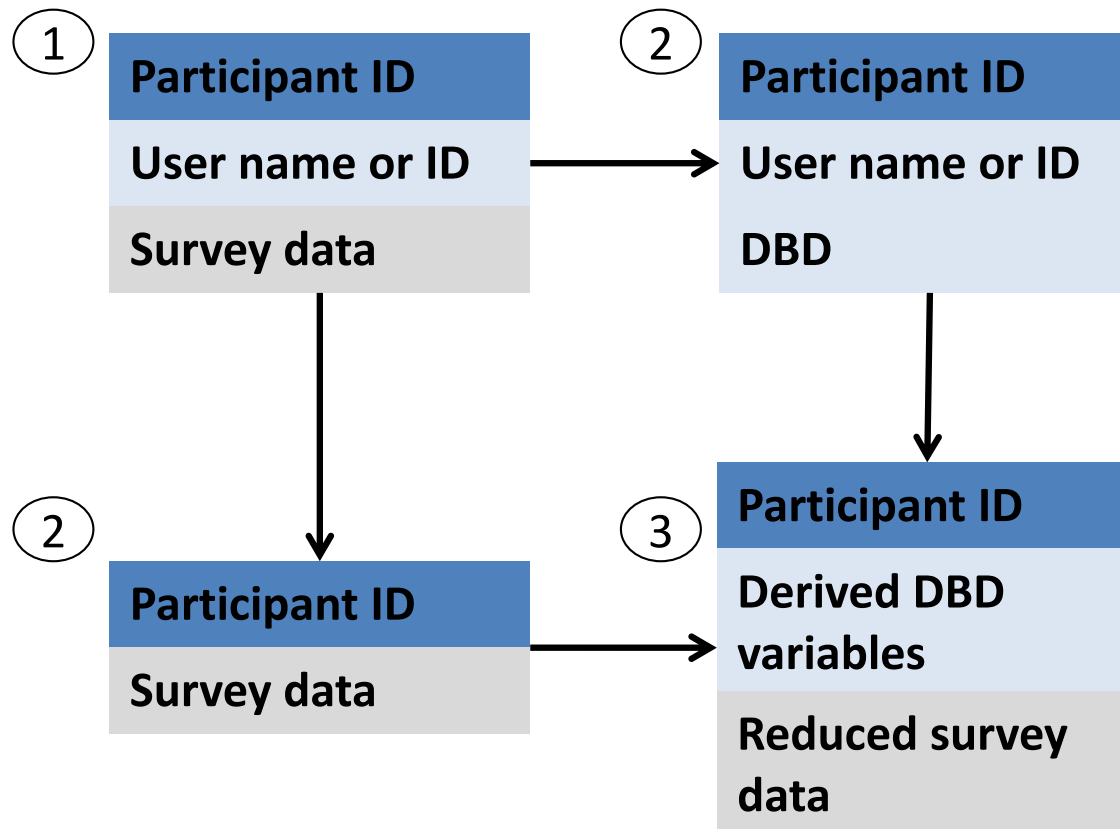
Informed consent

- Informed consent needs to adhere to **legal regulations** (GDPR in Europe) and should satisfy **ethical standards**
- Practical challenge: Properly informing participants without overwhelming them with information and (technical) details
- Substantially easier to obtain when starting with/from survey data
- For some proposed solutions and wording see [Breuer et al. \(2021\)](#) and [Sloan et al. \(2020\)](#)

Data privacy

- Data should be stored and processed in a way that **minimizes disclosure risk**
- In addition to regular data protection measures (passwords, access control, encryption, etc.) the survey data and DBD should be **kept separate as much as possible**

Proposed workflow



Based on [Beuthner et al. \(2021\)](#); originally adapted from [Sloan et al. \(2020\)](#)

Costs and feasibility of linking

- Commercial providers of tracking data are expensive
- Setting up the necessary infrastructure for own tracking tools is technically complex and costly
- Online platforms make surveying their users difficult
- Long-term academic tracking infrastructure is needed
- Plans for a DBD Access Panel at GESIS

Summary

- Linking surveys and DBD can be used to combine their unique strengths
- There are different ways of linking (e.g., ex-ante vs. ex-post)
- To increase willingness to share/link data researchers can use screening procedures, offer incentives and minimize respondent burden
- A set of practical, legal and ethical challenges needs to be addressed ([Stier et al., 2020](#))

Research with Digital Behavioral Data – more to come

New *Meet the Experts* series with talks about CSS methods and data coming soon: September – December 2021

Other options to learn about CSS at GESIS

GESIS Training offers a wide range of seminars, workshops, and other courses, including:

- Sep 13 - Oct. 1, 2021: Fall Seminar in Computational Social Science
- Nov 2 - 5, 2021: Workshop Introduction to Social Media Research Data: Potentials and Pitfalls (Katrin Weller and Indira Sen)

References

- Beuthner, C., Breuer, J., & Jünger, S. (2021). Data Linking—Linking survey data with geospatial, social media, and sensor data. *GESIS Survey Guidelines*. https://doi.org/10.15465/GESIS-SG_EN_039
- Breuer, J., Al Baghal, T., Sloan, L., Bishop, L., Kondyli, D., & Linardis, A. (2021). Informed consent for linking survey and social media data—Differences between platforms and data types. *IASSIST Quarterly*, 45(1), 1–27. <https://doi.org/10.29173/iq988>
- Conover, M., Ratkiewicz, J., Francisco, M., Goncalves, B., Menczer, F., & Flammini, A. (2011). Political Polarization on Twitter. In *Proceedings of the 5th International AAAI Conference on Web and Social Media* (pp. 89–96). AAAI Publications.
- Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., & Lazer, D. (2019). Fake news on Twitter during the 2016 U.S. presidential election. *Science*, 363(6425), 374–378. <https://doi.org/10.1126/science.aau2706>
- Guess, A. M. (2021). (Almost) Everything in Moderation: New Evidence on Americans’ Online Media Diets. *American Journal of Political Science*. <https://doi.org/10.1111/ajps.12589>
- Howison, J., Wiggins, A., & Crowston, K. (2011). Validity Issues in the Use of Social Network Analysis with Digital Trace Data. *Journal of the Association for Information Systems*, 12(12), 767–797. <https://doi.org/10.17705/1jais.00282>
- Nelson, J. L., & Webster, J. G. (2017). The Myth of Partisan Selective Exposure: A Portrait of the Online Political News Audience. *Social Media + Society*, 3(3). <https://doi.org/10.1177/2056305117729314>
- Parry, D. A., Davidson, B. I., Sewall, C. J. R., Fisher, J. T., Mieczkowski, H., & Quintana, D. S. (2021). A systematic review and meta-analysis of discrepancies between logged and self-reported digital media use. *Nature Human Behaviour*. <https://doi.org/10.1038/s41562-021-01117-5>
- Resnick, P., Adar, E., & Lampe, C. (2015). What Social Media Data We Are Missing and How to Get It. *The ANNALS of the American Academy of Political and Social Science*, 659(1), 192–206. <https://doi.org/10.1177/0002716215570006>
- Sen, I., Flöck, F., Weller, K., Weiß, B., & Wagner, C. (2021). TED-On: A Total Error Framework for Digital Traces of Human Behavior on Online Platforms. *ArXiv: 1907.08228 [Cs.CY]*.
- Silber, H., Breuer, J., Beuthner, C., Gummer, T., Keusch, F., Siegers, P., ... Weiß, B. (2021). Linking surveys and digital trace data: Insights from two studies on determinants of data sharing behavior. *OSF*. <https://doi.org/10.31235/osf.io/dz93u>
- 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. <https://doi.org/10.1177/1556264619853447>
- 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. <https://doi.org/10.1177/0894439319843669>
- Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, 359(6380), 1146–1151. <https://doi.org/10.1126/science.aap9559>

Thank you !

gesis

Leibniz-Institut
für Sozialwissenschaften

Leibniz
Leibniz
Gemeinschaft

GESIS Consulting

GESIS offers individual consulting in a number of areas – including survey design & methodology, data archiving, digital behavioral data & computational social science – and across the research data cycle. Please visit our website www.gesis.org for more detailed information.

GESIS consulting is *free of charge* for researchers who conduct

- scientific projects – financed institutionally or by third-party-funds – at universities or publicly funded research institutions, or
- scientific projects at institutions of the Federal Government or the *Länder* or other publicly funded institutions.

For other projects consulting is *subject to a charge* and to available resources.



Expert contact:

johannes.breuer@gesis.org

sebastian.stier@gesis.org

Please find on the GESIS website consulting contacts for:

[Planning Studies](#), [Accessing Data](#), [Analyzing Data](#), [Archiving Data](#)

More Services from GESIS

- [GESIS Survey Guidelines](#) provide short and hands-on explanations to frequent challenges in survey design and methodology.
- Use GESIS data services for [finding data](#) for secondary analysis and [sharing your own data](#).
- Get materials for [capacity building](#) in computational social science and take advantage of our expanding expertise and resources in [digital behavioral data](#).
- Check out the [GESIS blog](#) "Growing Knowledge in the Social Sciences" for topics, methods and discussions from the GESIS cosmos – and beyond.
- Keep up with GESIS activities and subscribe to our monthly [newsletter](#).
-  for publications, tools & services.