DIGITAL DEMOGRAPHY: ANALYZING WEB AND SOCIAL MEDIA DATA

DAY 2 – BASICS OF DIGITAL DEMOGRAPHY

EDSD NOVEMBER 2022

Tom Theile

Lab of digital and computational demography,

Max-Planck-Institute for Demographic Research, Rostock

AGENDA AND SCHEDULE OF THE WEEK

- 9:30 – 13:00 Monday to Thursday
- Friday: assignment
- Block 1: Internet, Webscraping, APIs
- Block 2: Digital Demography
- Block 3: Social Media Data
- Extra: Digital Datasets

FINAL ASSIGNMENT

- A small number of short tasks
- To be submitted until Friday afternoon 5pm
- Everyone has to submit the answers by email
- Email has to contain your full name!
- More information on Thursday

DIGITAL DEMOGRAPHY.

What is ‘digital’ demography?

DIGITAL DEMOGRAPHY (NOT SERIOUS!)

- Demography is old
- It is your job to come up with novel methods to capture/analyze/draw-conclusions-from demographic data
- →Computers!

DIGITAL DEMOGRAPHY

Digital Data <--> Digital methods

DIGITAL DEMOGRAPHY

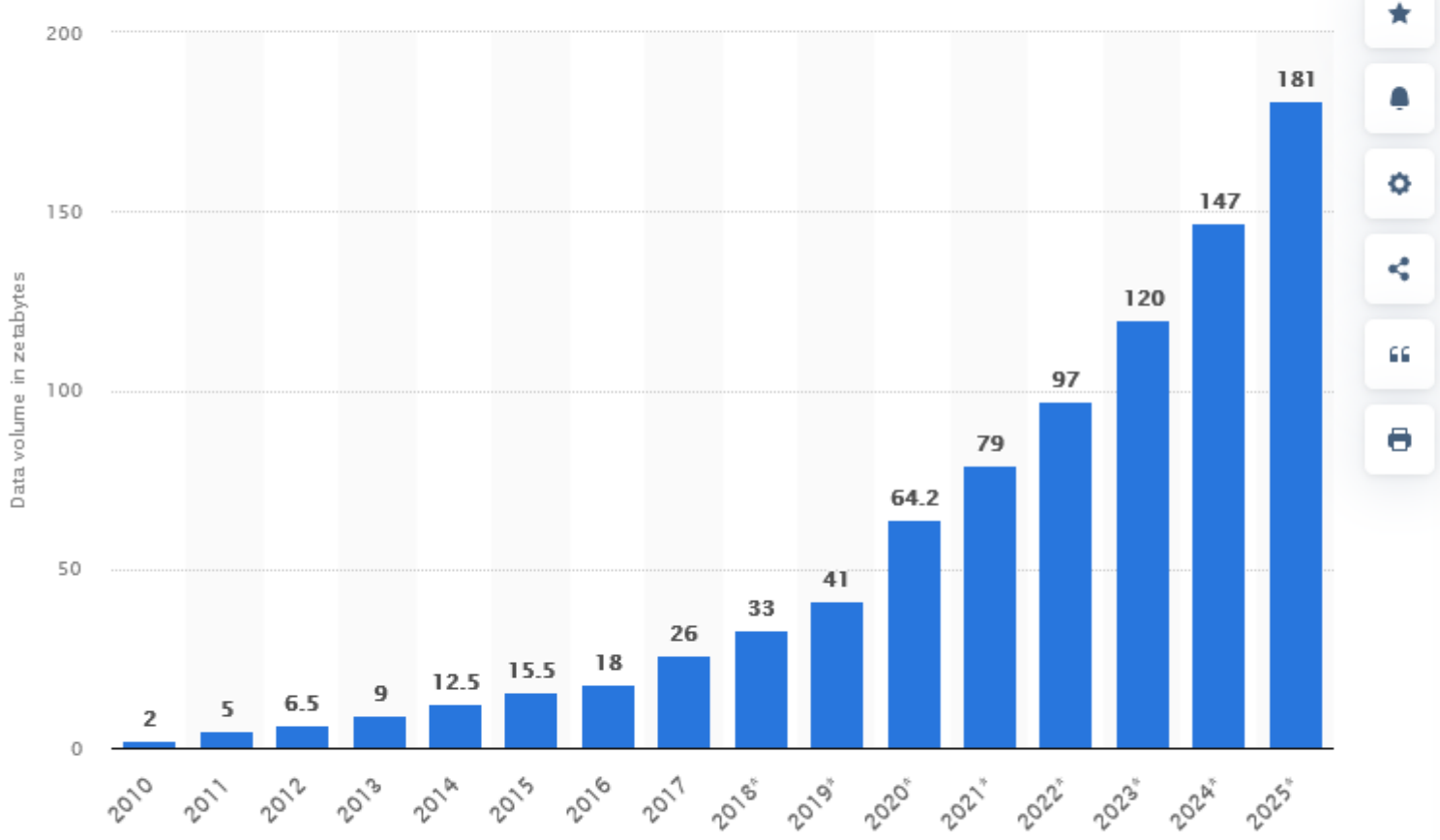
- 2nd next birthday: Did you learn something that was completely new to you?
- 3rd next birthday: What kind of data will you use in your research and In what time did your planned data source first appear?
- 4th next birthday: What makes digital trace data a type of ‘new’ big data for population research that is different from ‘old’ big population data sources?

What did you not understand?

Criticism of the paper?

Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2020, with forecasts from 2021 to 2025

(in zettabytes)



© Statista 2022

Additional Information

Show source

DOWNLOAD

PDF XLS PNG PPT

Sources

- Show sources information
- Show publisher information
- Use Ask Statista Research Service

Release date

June 2021

Region

Worldwide

Survey time period

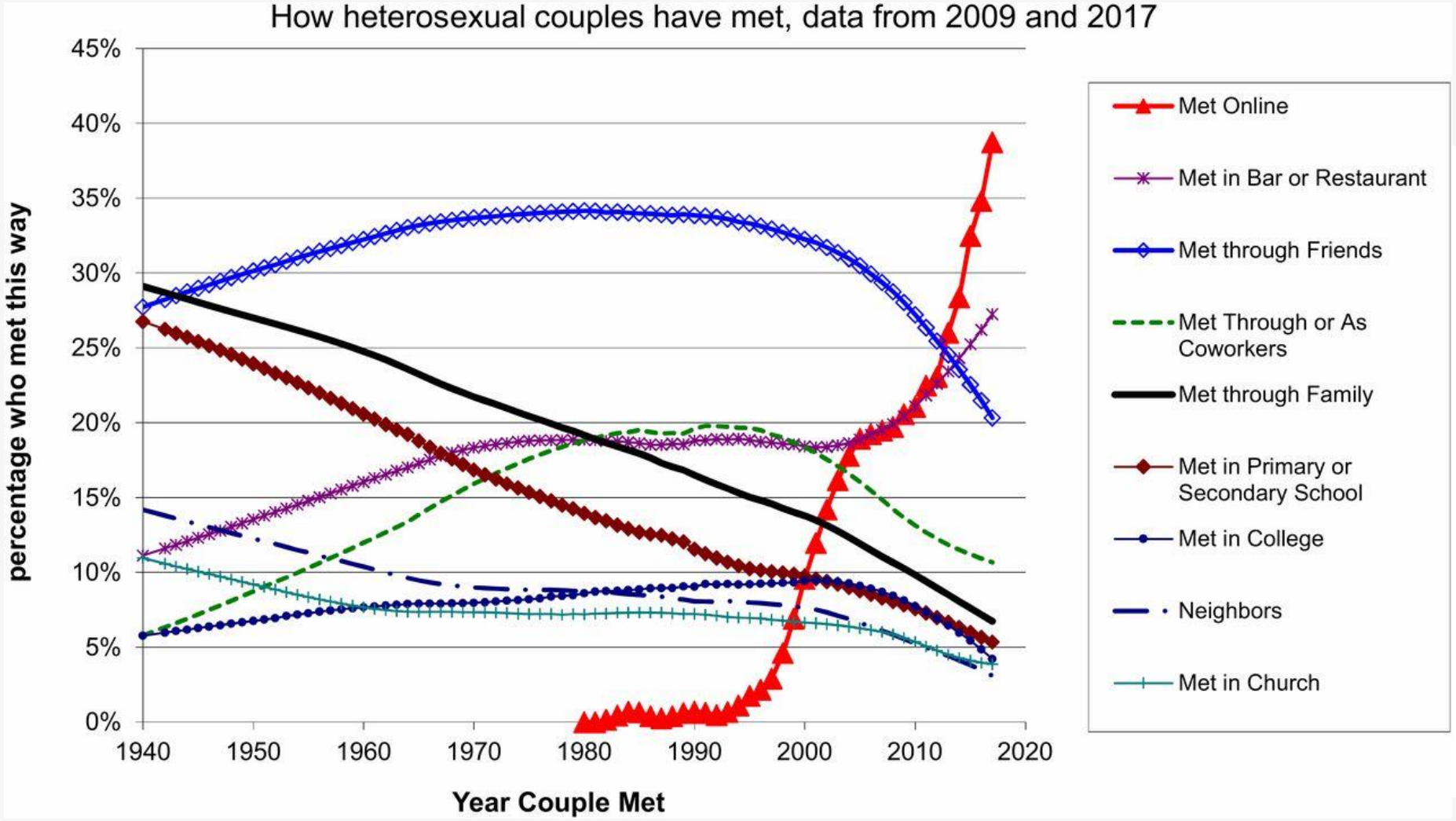
2010 to 2020

Supplementary notes

* The data was taken from various publications released over several years: Forecast for the years 2018 and 2019 as of 2018; Forecast for 2020 as of May 2021; Forecast for 2021 to 2025 as of March

Source: https://www.statista.com/statistics/871513/worldwide-data-created/

DIGITAL TRANSFORMATIONS HAVE CHANGED OUR LIVES



Source: https://www.pnas.org/doi/10.1073/pnas.1908630116 - *Rosenfeld 2019*

DIGITAL DATA SOURCES FOR DEMOGRAPHIC RESEARCH

1. Digital Trace Data (online and offline)
    1. Social media
    2. Mobile phones
    3. Wearable devices, etc...
2. Crowd-sourced online data
    1. Wikipedia and Wikidata
    2. DNA and online genealogies
    3. Petitions, etc...
3. Online Surveys
4. Simulations (made-up data)

DIGITAL DEMOGRAPHY

What:

1. Apply formal demography to digital trace data (1)
    1. "Digital censuses" in Facebook and Linkedin
    2. Life tables to estimate survival within platforms (users are born and die)
2. Advance theory (2)
    1. Demographic (holy trinity: mortality, fertility, migration)
    2. Sociological theory (explain social action)

1 Cesare, N., Lee, H., McCormick, T., Spiro, E., and Zagheni, E. (2018). Promises and pitfalls of using digital traces for demographic Research. Demography 55(5):1979–1999.
2 Edelmann, A., Wolff, T., Montagne, D., and Bail, C.A. (2020). Computational Social Science and Sociology. Annual Review of Sociology 46(1):61–81

DIGITAL DEMOGRAPHY

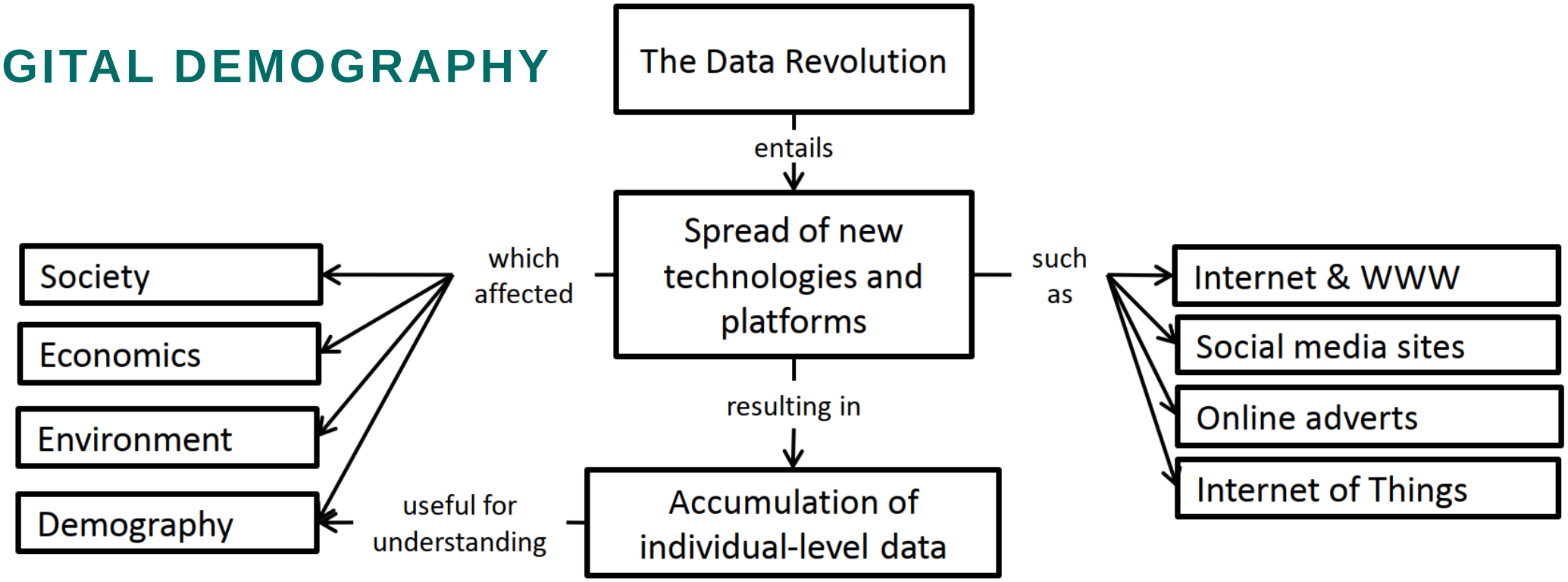


Fig. 1 The Data Revolution and new sources of data for demographic analysis.

- Digital vs analogue
- Online vs offline
- ‘Big’ vs ‘small’ data

IS ‘BIG DATA’ NEW DATA?

DEMOGRAPHER COLLECTING BIG DATA FOR THE 1925 US CENSUS

https://upload.wikimedia.org/wikipedia/commons/6/6f/Volkstelling_1925_Census.jpg

LIMITATIONS OF TRADITIONAL DATA SOURCES

- Costly
- Outdated
- Time consuming
- Inconsistent
- Unavailable
- Lack of data on emigration
- Incomplete answers/misunderstanding of questions etc.
- Immigrants are often underrepresented in traditional data sources.
- limited in hard-to-reach contexts and societies.

WHAT IS BIG DATA?

is “information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation” 6 [1]. It can be described in 3 Vs which are:

- *Volume: As the name suggests, the size of the data is Big, hence the volume of the data.*
- *Velocity: Big data such as Twitter allow us to stream data at real-time.The rate at which we obtain data is faster than the traditional data sources.*
- *Variety: Traditional data are mostly structured data. Big data, on the other hand, come in various forms. It can be videos, photos, texts, and audios. It requires a thorough data processing before extracting information/knowledge from it.*

Jisu Kim 2022; Salganik, M. (n.d.). Bit by Bit: Social Research in the Digital Age. Princeton, NJ: Princeton University Press.

WHAT IS DIGITAL TRACE DATA?

What makes digital trace data a type of 'new' big data for population research that is different from 'old' big population data sources?

•by-product of digital activity
•always collected, not once in a while
•Unlike rectangular dataframes with rows and columns, many digital traces data are unstructured, messy, and come in formats unfamiliar to many demographers
•The variety of formats, units of analyses, and also sizes of these data sets, which may contain
• millions of records, often require computational approaches for data management, retrieval, and analysis that are not yet a part of mainstream demo-graphic training.
•not comparable over longer time periods - platforms, users and algorithms change fast
•these data often come from and are owned by private companies

Ridhi Kashyap: Has demography witnessed a data revolution? Promises and pitfalls of a changing data ecosystem? 2021

DIGITAL TRACES ARE BY-PRODUCTS OF OUR ONLINE PRESENCE

Digital breadcrumbs are unavoidable

- Pre-GDPR: largely unchecked
- Marketing-led
- Not collected for social-scientific research

Diego Alburez-Gutierrez, EDSD 2022; 2

BIG DATA FOR SOCIAL RESEARCH: THE GOOD

Twitter, Facebook, Yahoo, ...

- Big
- Free (sometimes) or cheap (often)
- Granular data
- Large scale data
- Continuously generated, always-on
- Non-reactive
- Information/opinion shared by users from an uncontrolled environment
- Various forms of data: video, image, text, audio etc.

Jisu Kim 2022; Salganik, M. (n.d.). Bit by Bit: Social Research in the Digital Age. Princeton, NJ: Princeton University Press.

BIG DATA FOR SOCIAL RESEARCH: THE BAD

- incomplete
- inaccessible
- nonrepresentative (within- and out-of sample)
- drifting (population, behavioural, system)
- algorithmically confounded (observing a casino)
- inaccessible
- dirty
- sensitive

Salganik, M. (n.d.). Bit by Bit: Social Research in the Digital Age. Princeton, NJ: Princeton University Press.

CURRENT TOPICS IN DIGITAL DEMOGRAPHY

1. Methodological developments
    1. Inference from non-representative samples
    2. Understand and address online bias
    3. Nowcasting demographic processes
2. Understanding internet users and online use
    1. Infer demographics (age, sex, location, SE status, etc) from image and text
    2. Track inequalities in online access
    3. Consequences of platform use for users
3. Migration (internal and external)
    1. Estimate flows and stocks
    2. Mobility by subgroups (e.g. undocumented, highly-skilled)
    3. Cultural assimilation of immigrants
4. Mortality and morbidity
5. Online and offline fertility dynamics
6. Time use and well-being

ISSUES WITH TRADITIONAL SURVEYS

Efficacy: traditional sampling methods are outdated and less feasible

Coverage: decreased response rates, difficult to sample from hard-to-reach populations

Resources: expensive and time consuming

Recency: quickly become outdated, long period between new data collections

Comparability: lack of common definitions across countries

ISSUES WITH TRADITIONAL SURVEYS

Cost-effectiveness: less expensive than traditional surveys

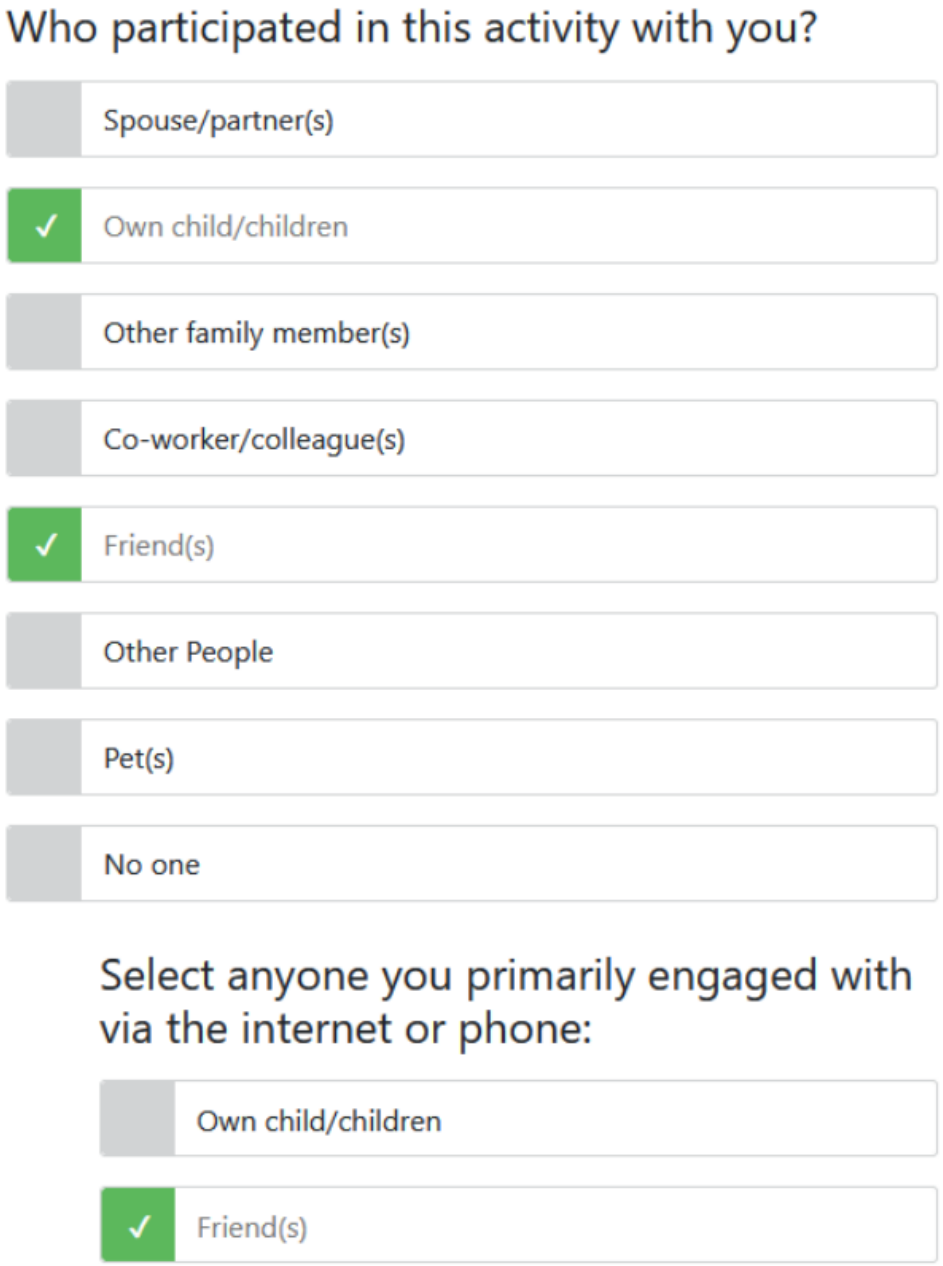
Coverage: targeted sub-populations, hard-to-reach populations

Timeliness: easy and timely implementation, data collection & analysis in near real-time

Flexibility: less burdensome, user-friendly interfaces, easy to manage

Recency: continuous data collection, easy to make edits

Comparability: cross-national surveys, comparative data collections, common definitions



- Rinderknecht, R. G., Doan, L., & Sayer, L. C. 2022. “MyTimeUse: An Online Implementation of the Day-Reconstruction Method” *Journal of Time Use Research.*

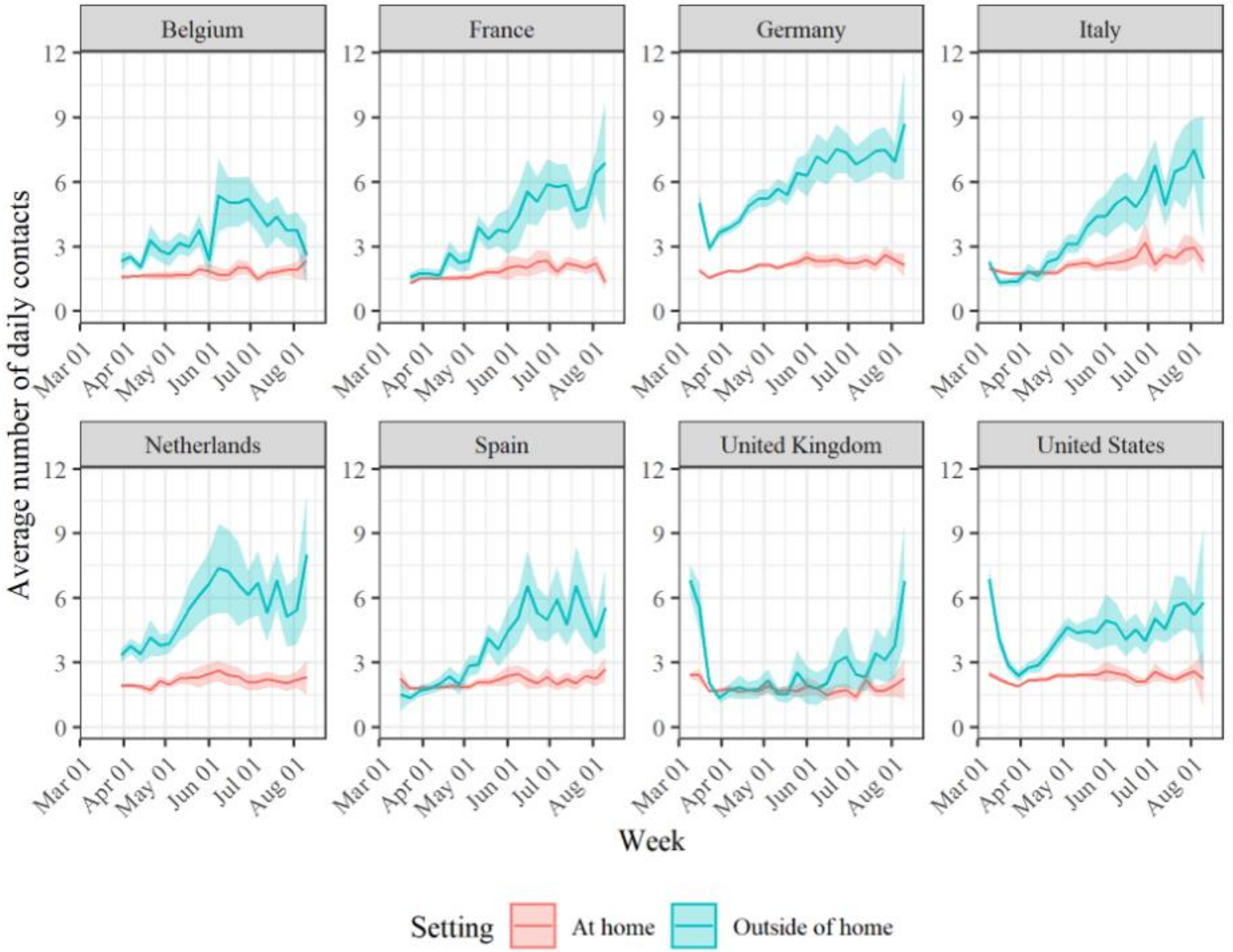
ONLINE SURVEYS ON SOCIAL MEDIA

Relative advantage: scale

Important for:

- Targeting specific subgroups or regions
- Conducting research outside of Western countries

ONLINE SURVEYS ON SOCIAL MEDIA



Grow, A., Perrotta, D., Del Fava, E., Cimentada, J., Rampazzo, F., Gil-Clavel, S., & Zagheni, E. (2020). Addressing public health emergencies via Facebook surveys: advantages, challenges, and practical considerations. Journal of medical Internet research, 22(12), e20653.

ONLINE SURVEYS ON SOCIAL MEDIA

Table 4. Comparison of Sample Sizes at Different Stages of the Sampling and Survey Process.

	Number of Users Belonging to the Target Population (According to Facebook)[a]	Unique Users Reached with Ads[a]		Payed Link Clicks[a]		Completed Questionnaires[b]	
		n	Percentage of Targeted FB Users in this Country	*n*	Percentage of Targeted FB Users in this Country	*n*	Percentage of Targeted FB Users in this Country
Austria	15,000	7,918	52.79	408	2.72	117	0.78
Ireland	54,000	28,107	52.05	1,314	2.43	425	0.79
Switzerland	9,000	3,432	38.13	215	2.39	62	0.69
United Kingdom	410,000	50,979	12.43	1,257	0.31	424	0.10

Note. [a]Source of absolute figures: Facebook (FB) advertisement statistics. Relative values: own calculation. [b]Based on paradata. Only respondents who reached the questionnaire via the FB advertisements.

Budget of €500, and 96% of the 1,028 respondents belonged to target population.

• Pötzschke, S., & Braun, M. (2017). Migrant sampling using Facebook advertisements: A case study of Polish migrants in four European countries. Social Science Computer Review, 35(5), 633-653.

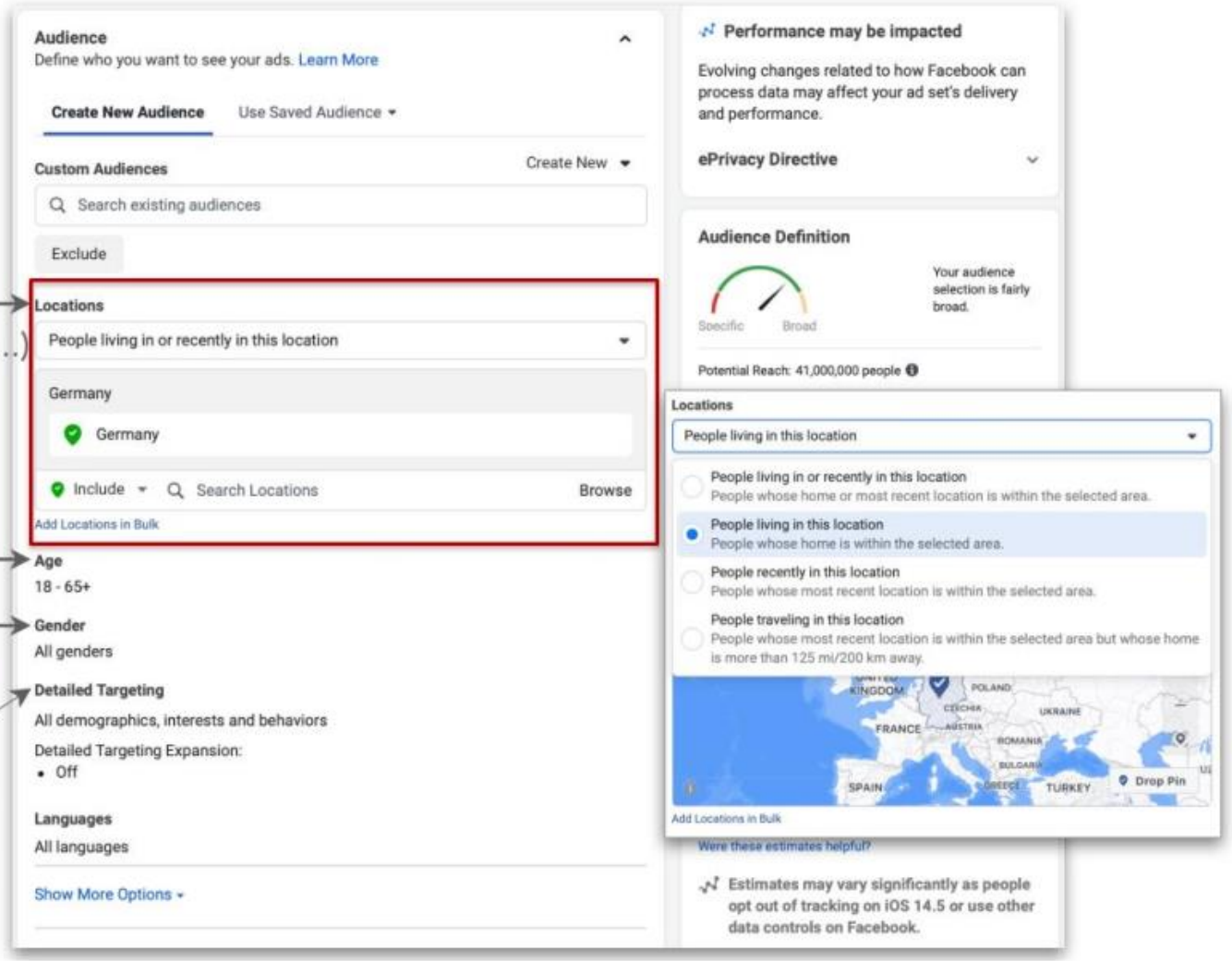
ONLINE SURVEYS ON SOCIAL MEDIA

Location
(country, region, city, ...)

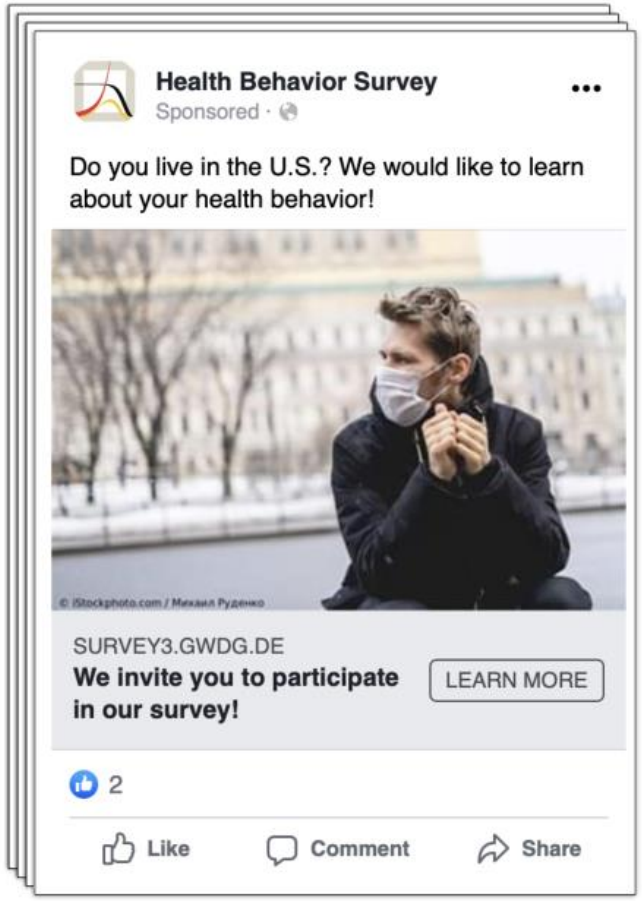
Age (ranges)

Gender
(men, women, both)

Other matching
criteria
(interests, industry,
behaviours, ...)

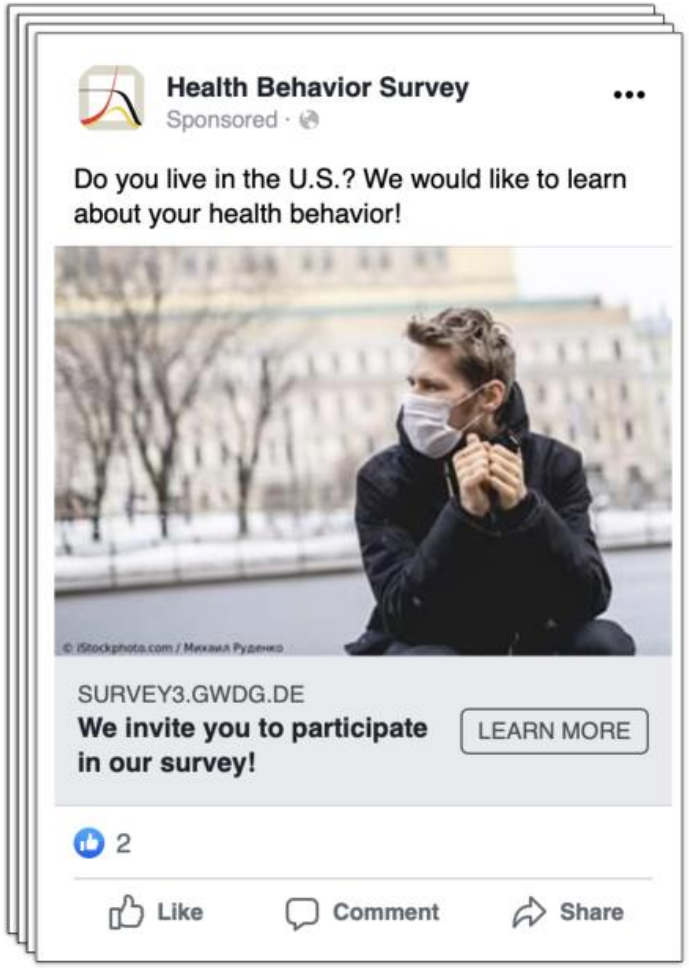


ONLINE SURVEYS



Example of FB ad in the US.

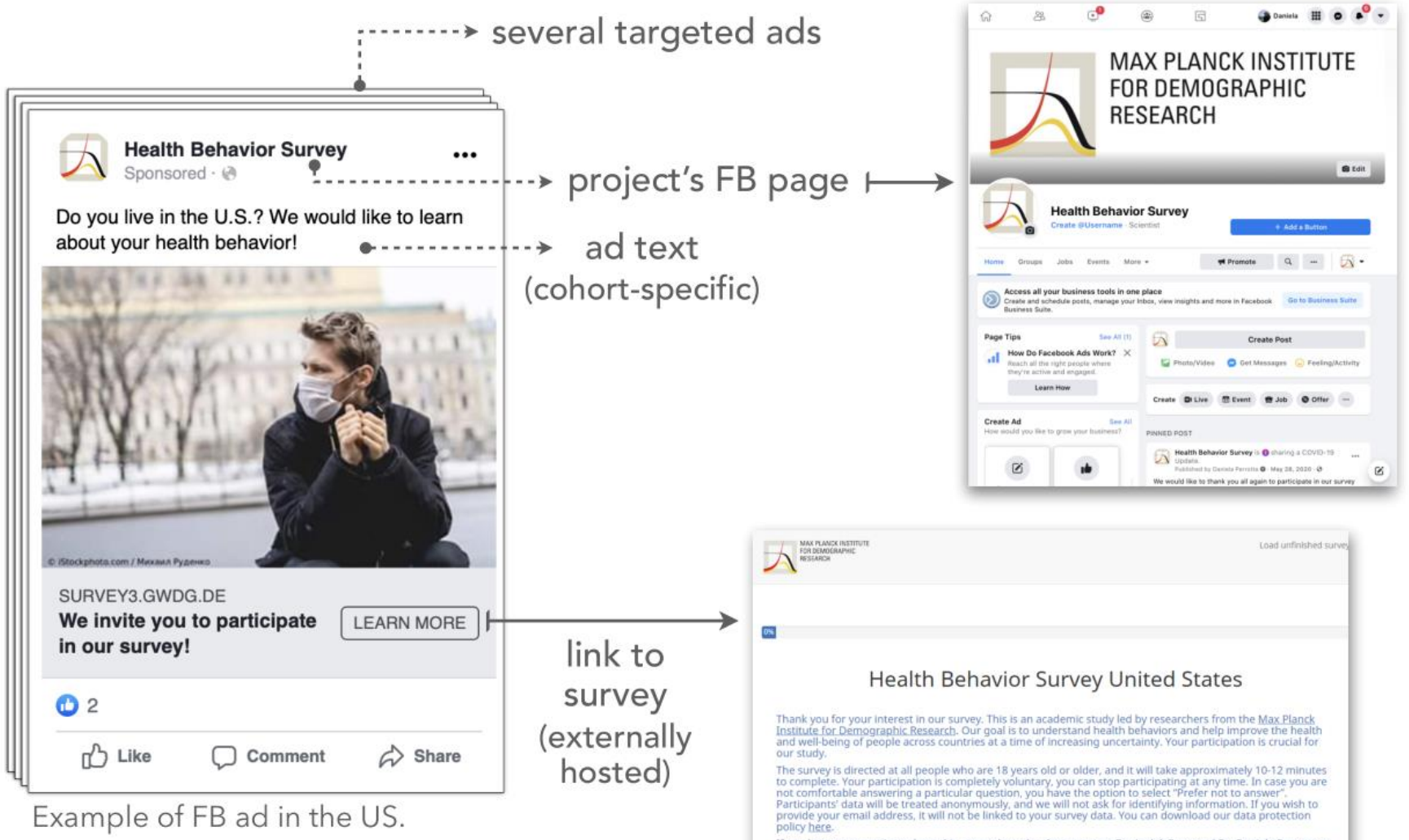
RESPONDENTS RECRUITMENT VIA FACEBOOK



Example of FB ad in the US.

Grow A, Perrotta D, Del Fava E, Cimentada J, Rampazzo F, Gil-Clavel S, Zagheni E. Addressing Public Health Emergencies via Facebook Surveys: Advantages, Challenges, and Practical Considerations. JMIR, 2020

RESPONDENTS RECRUITMENT VIA FACEBOOK



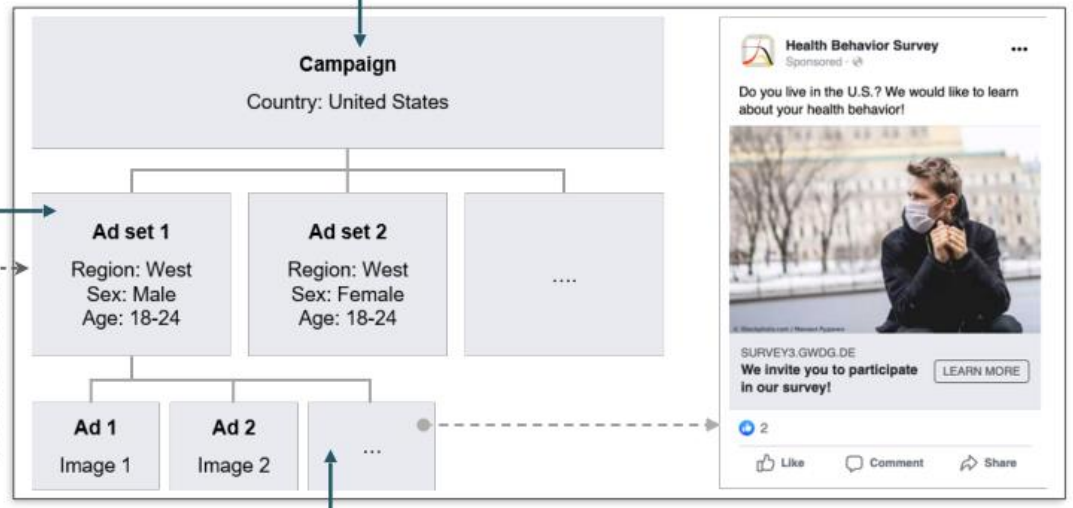
Example of FB ad in the US.

RESPONDENTS RECRUITMENT VIA FACEBOOK

ONE AD CAMPAIGN PER COUNTRY

ONE AD SET PER DEMOGRAPHIC GROUP

- sex (M, F)
- age (18-24, 25-44, 44-64, 65+)
- region of residence (NUTS1/US Census regions)



SIX AD IMAGES WITHIN EACH AD SET


1 – Male athlete
©Adobe Stock/grki


2 – Group of athletes
©Adobe Stock/nd3000


3 – Woman blowing nose
©iStockphoto/Goodboy Picture Company


4 – Couple blowing noses
©iStockphoto/Goodboy Picture Company

5 – Woman wearing mask
©Adobe Stock/shintartanya

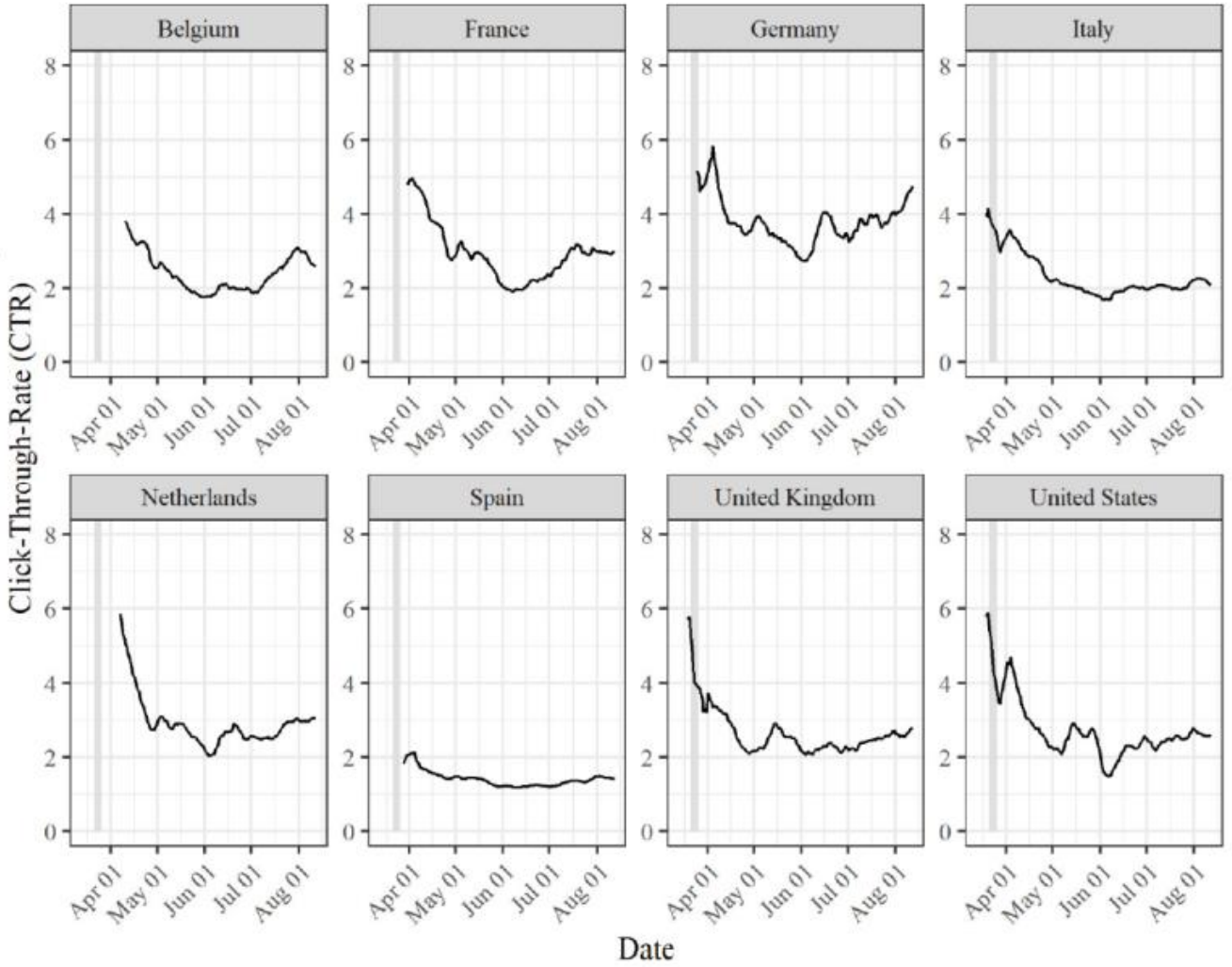

6 – Man wearing mask
©iStockphoto/Михаил Руденко

Grow A, Perrotta D, Del Fava E, Cimentada J, Rampazzo F, Gil-Clavel S, Zagheni E. *Addressing Public Health Emergencies via Facebook Surveys: Advantages, Challenges, and Practical Considerations*. JMIR, 2020

RESPONDENTS RECRUITMENT VIA FACEBOOK

Click-through rate (CTR) =
click-throughs / impressions

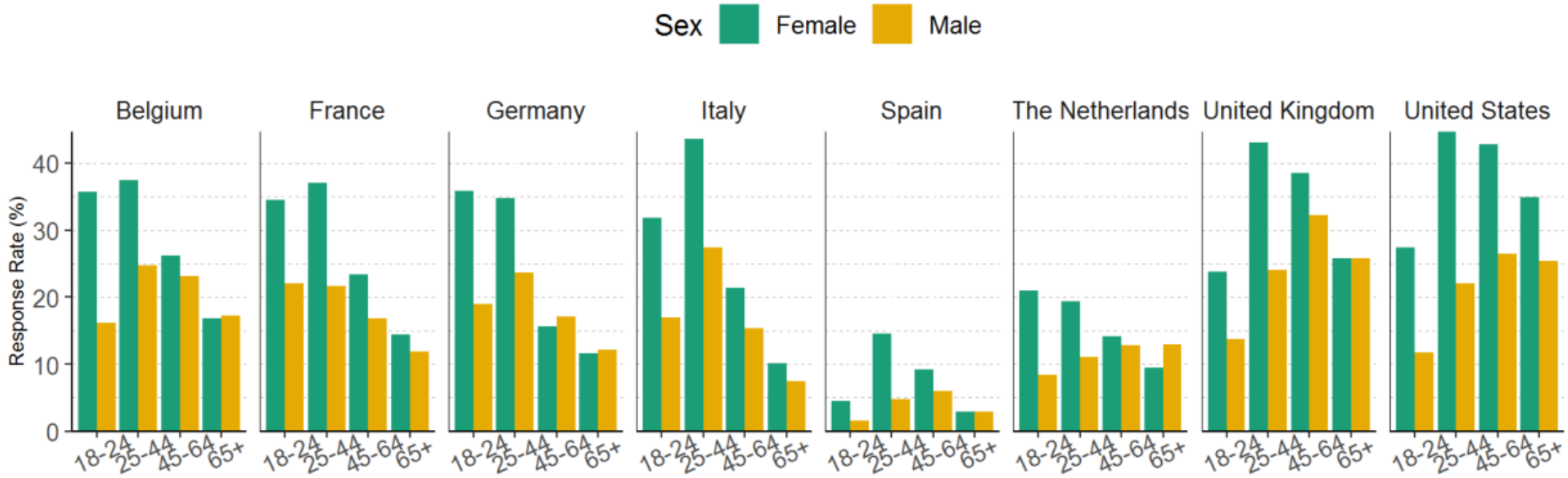
Facebook users were more likely to click on our ads in the early phases of the survey



RESPONDENTS RECRUITMENT VIA FACEBOOK

Response rate (%) = completed questionnaires / click-throughs

- response rate (overall): from 6% in Spain to 31% in UK and US
- response rate higher for women



Daniela Perrotta, PHDS 2021

CROWDSOURCED PLATFORMS

Relative advantage: convenience

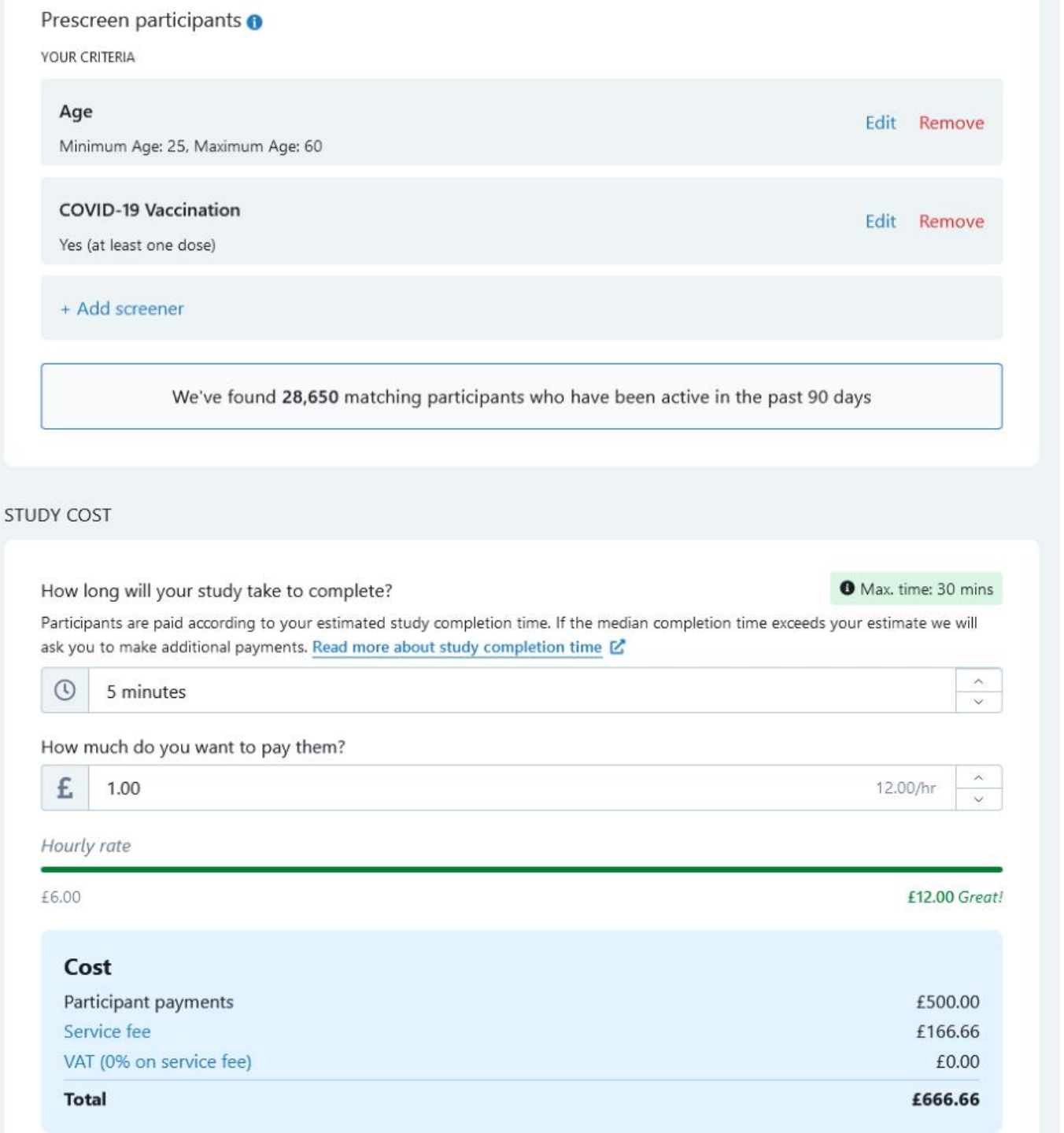
• They facilitate participant payment, re-recruitment, and messaging without requesting personally identifying information. These features allow for complex research designs.

amazon mechanical turk

www.mturk.com

Prolific

www.prolific.co



Aguinis, H., Villamor, I., & Ramani, R. S. (2021). MTurk research: Review and recommendations. Journal of Management, 47(4), 823-837.

CROWDSOURCED PLATFORMS

Relative disadvantages

• Population size is relatively small. While researchers were able to recruit ~1,000 Polish migrants in four countries with Facebook, and Facebook reported ~500,000 such users on their platform, Prolific reports only ~400 such respondents available for recruitment in these countries.

• Participants tend to be concentrated in the West, especially the U.S.

• You cannot use these platforms to recruit voluntary samples

amazon mechanical turk

www.mturk.com

Prolific

www.prolific.co

SOCIAL MEDIA AND CROWDSOURCED PLATFORMS

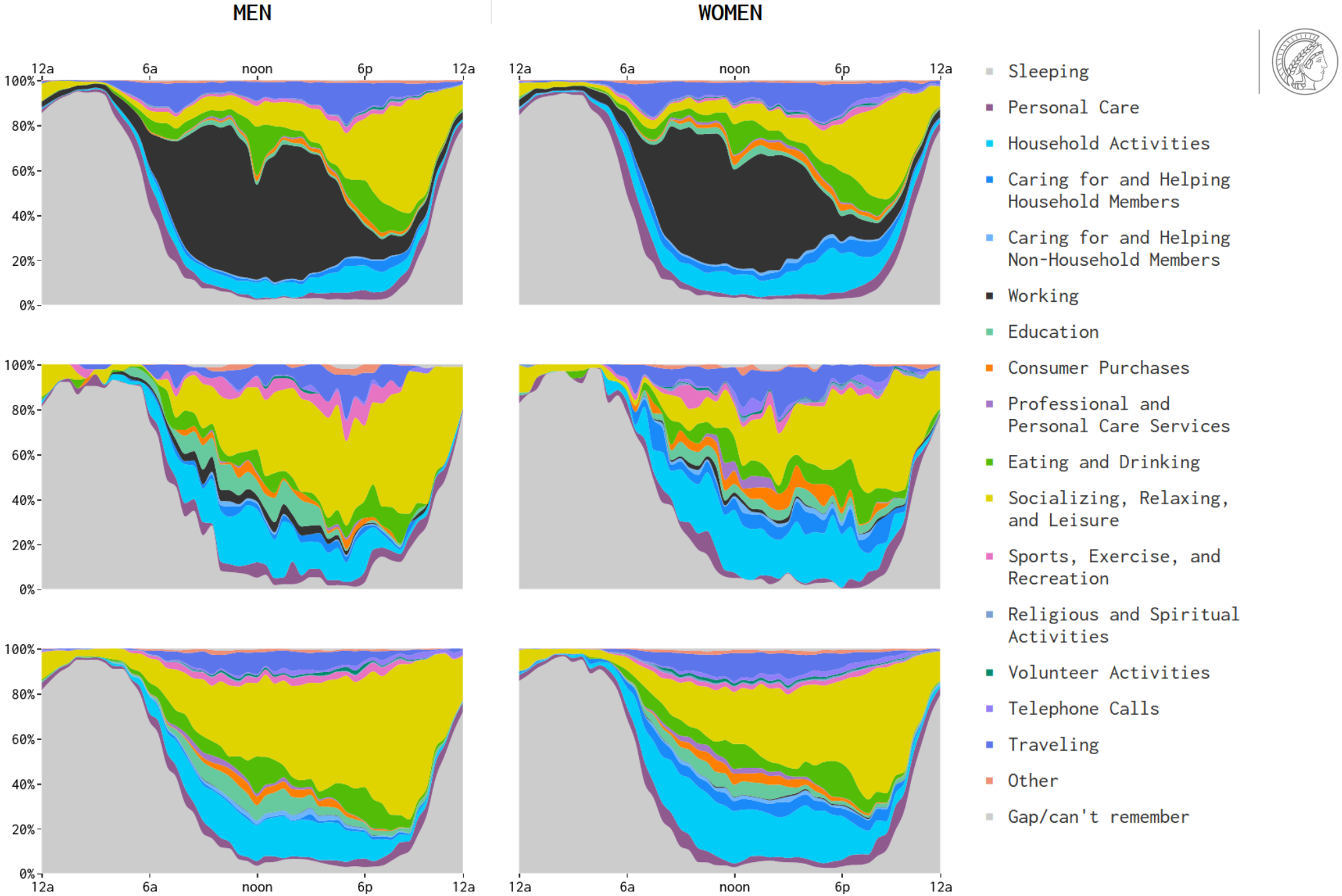
What do you think are potential problems?

ONLINE SURVEYS ON SOCIAL MEDIA

Challenges of recruitment via Facebook:

- Facebook is a “black box”: Facebook will “optimize” your advertisement in ways that skew participant demographics.
  - Also: Self-selection based on interest in the survey topic
  - Partial solution: run lots of advertisements targeting specific demographic groups. (Grow et al. 2020)
- Facebook will sometimes cause delays, either by reviewing your advertisements or claiming problems with your payment method.
  - Delays are unpredictable, but they are more common if your ad is related to social or political topics or if you offer to pay participants.
- Facebook has no built-in methods for paying participants or advertising specifically to people who previously completed your survey.
- Lack of attention of respondents

Rinderknecht, Gordon, PHDS 2022



Source: American Time Use Survey 2020 / By: FlowingData

THANK YOU FOR YOUR ATTENTION!

Tom Theile

Research Software Engineer

theile@demogr.mpg.de