Utrecht University

Summer Course Survey Research: Advanced Survey Design

Bella Struminskaya & Peter Lugtig

Department of Methodology & Statistics, Utrecht University

Slides by Struminskaya

Utrecht University

Data Collection using Apps, Wearables, Sensors Consent & Ethics

Bella Struminskaya & Peter Lugtig

Department of Methodology & Statistics, Utrecht University

Ethics

Premise under which social and market research can function effectively: research participants provide relevant and accurate information without having to fear about the adverse effects of doing so (Joe et al. 2016)

Ethical principles for protecting the rights of respondents are rooted in:

- The Helsinki Declaration (General Assembly of the World Medical Association, 1964)
- The Belmont Report (1974)

Other reports, e.g., the Menlo Report 2011 on ethical principles guiding information and communication technology research issued by the US Department of Homeland Security (Dittrich et al. 2011)

Three basic ethical principles

1. Beneficence — minimizing harm while maximizing benefit for the individual subjects
2. Justice — burdens of research should not be shared unequally among groups of subjects with some bearing the burdens and other reaping the benefits
3. Autonomy — requires obtaining informed consent from subjects for their research participation

(Singer and Couper, 2011, p. 134-135; see also Joe et al. 2016, p. 79)

Three examples (Salganik 2018)

- Facebook's study of emotional contagion: participants did not provide specific consent and the study has not been subject to a third-party ethical review
- A study in which researchers linked Facebook profile data with administrative records of students without prior consent
- A study in which researchers caused people's computers to visit websites potentially blocked by repressive governments

Is this unique for Big Data Age?

Ethics and consent in the age of big data

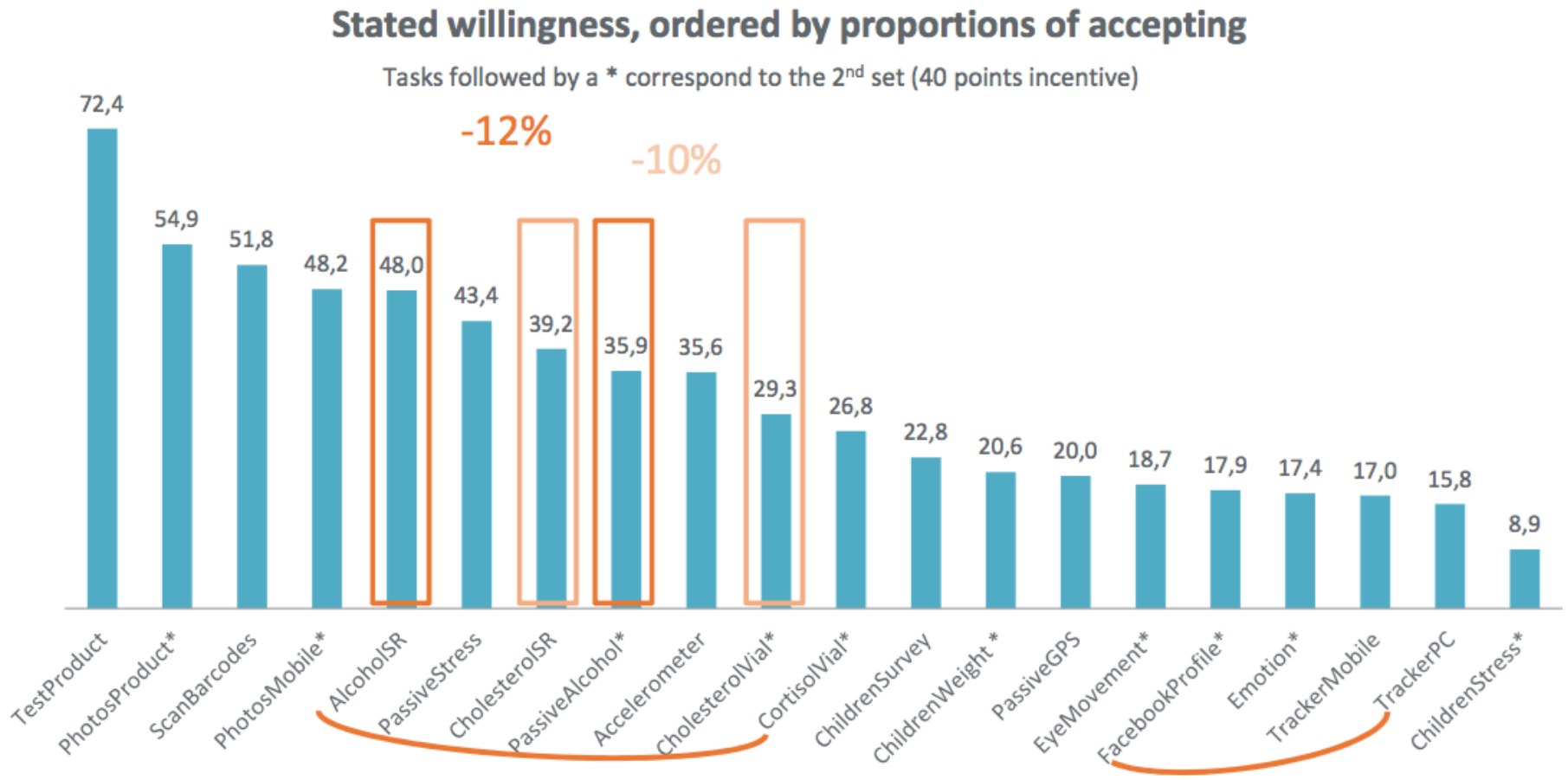
- Argument “no risk involved no need of informed consent” but purpose is to give control to respondents over the information about themselves that they are sharing
- Difficult to anticipate all the purposes for which the information collected on a survey can be used
- Characteristics of big data making it prone to confidentiality breaches: volume, high level of detail, unanticipated secondary use

Paradata: Ethical aspects & privacy concerns

Couper & Singer (2013):

- 3 experiments about giving informed consent to allow for the collection of paradata
- none of the tested versions were able to fully inform respondents about what paradata is
- requiring informed consent reduced survey participation

Willingness to perform additional tasks



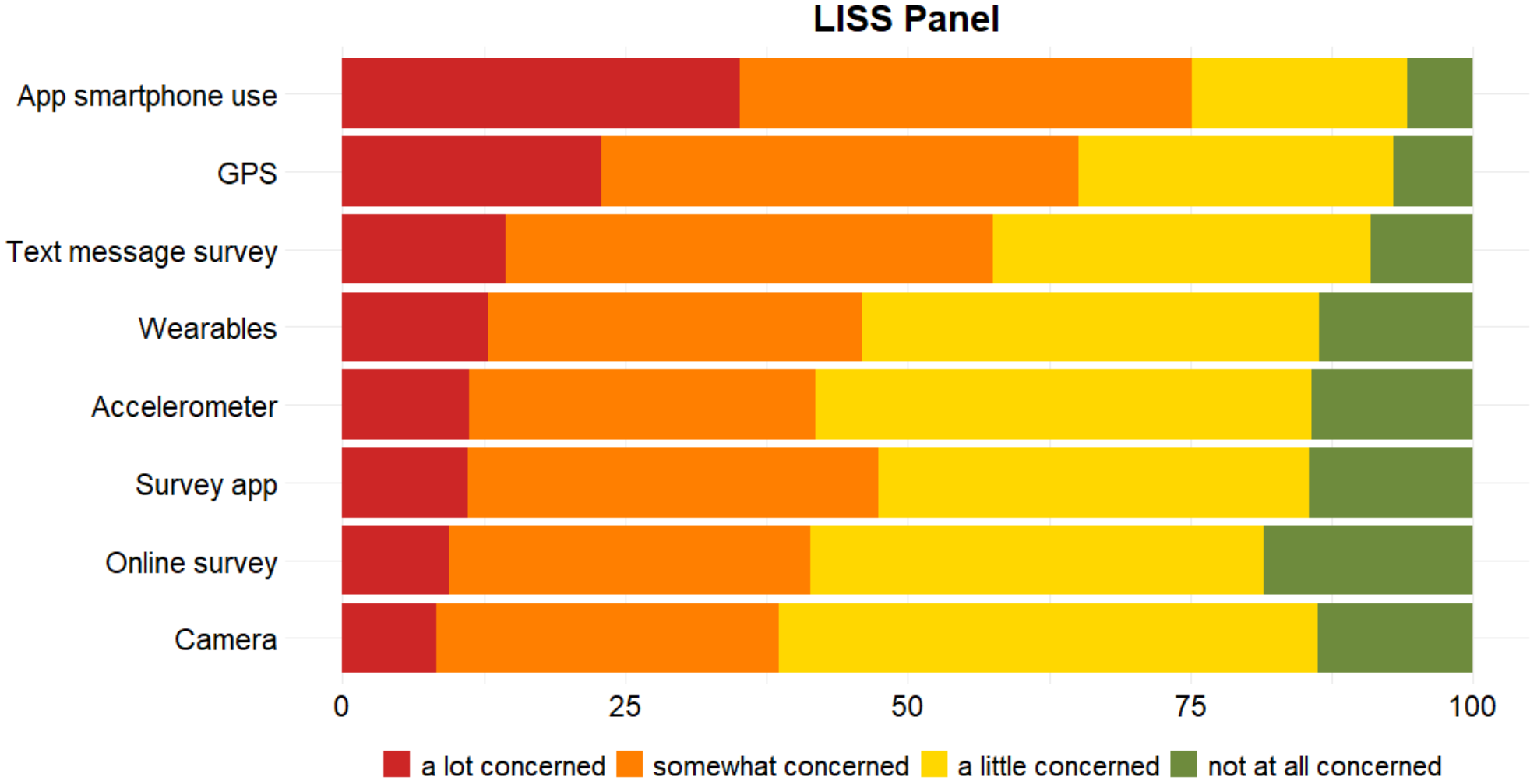
„Willingness is higher for tasks where respondents have control over the reporting of the results, even if this means more effort“

(Revilla, Couper, & Ochoa, 2018)

Privacy Concerns

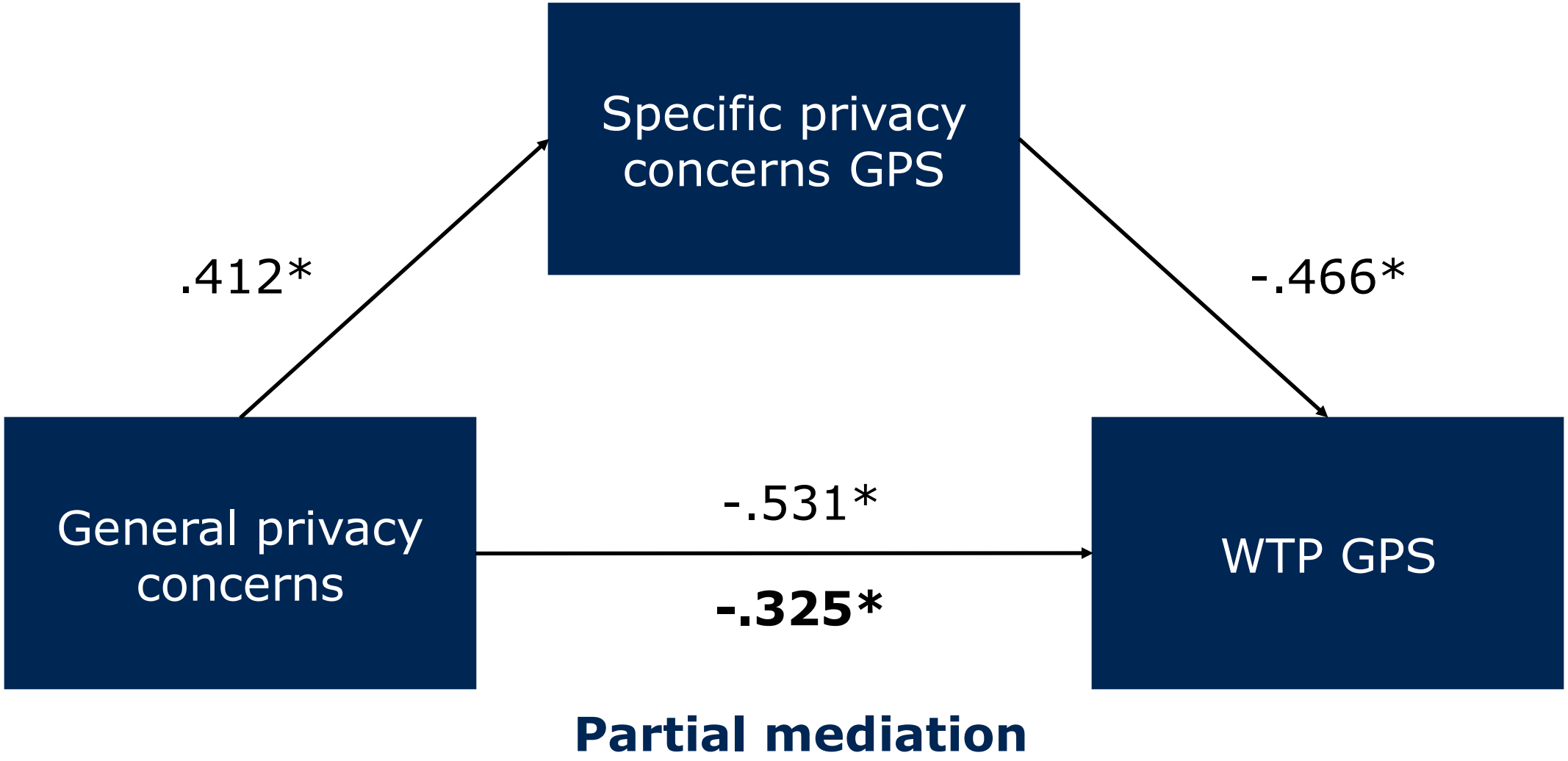
- Participants might have concerns about potential risks related to passive mobile data collection
  - Data streams could be intercepted by unauthorized party
  - Connecting multiple streams of data could re-identify previously anonymous users
  - Information could be used to impact credit, employment, or insurability
  - …
- Main reason against participation (Revilla et al. 2018; Jäckle et al. 2019; Keusch et al. 2019; Struminskaya et al. 2020)
- Higher privacy concerns correlate with lower WTP (Keusch et al. 2019; Revilla et al. 2019; Struminskaya et al. 2020, in press; Wenz et al. 2019)
- The more situations R‘s perceive as violation of privacy (by banks, gov‘t, social media etc.) the lower WTP (Keusch et al. 2019)
- No effect of emphasizing privacy on WTP (Struminskaya et al. 2020, in press)

Concern by Type of Data



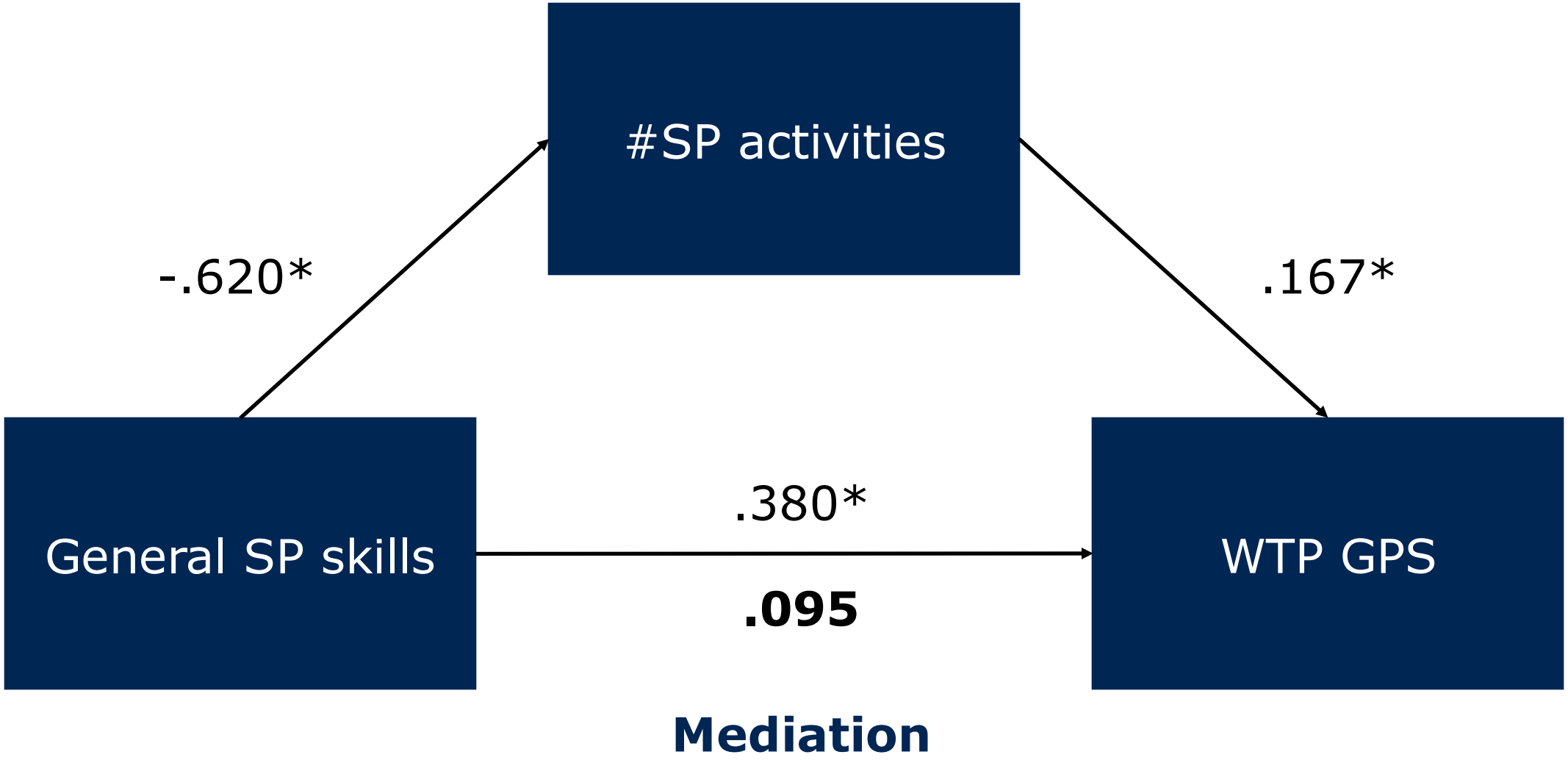
Struminskaya & Keusch (in prep.)

General and Specific Concern



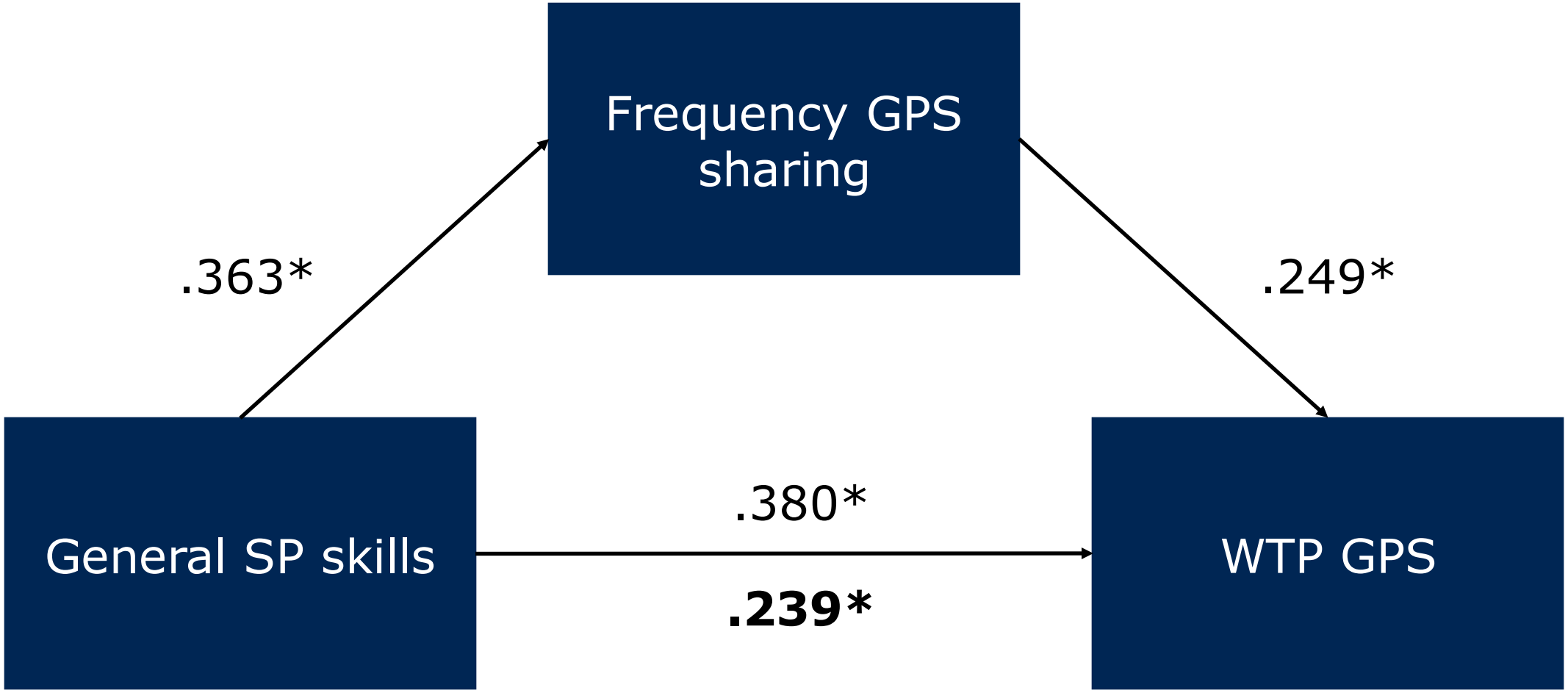
Logistic regression coefficients; *p<.001

General SP Skills and #SP Activities



Logistic regression coefficients; *p<.001

General SP Skills and Frequency GPS Sharing



Partial mediation

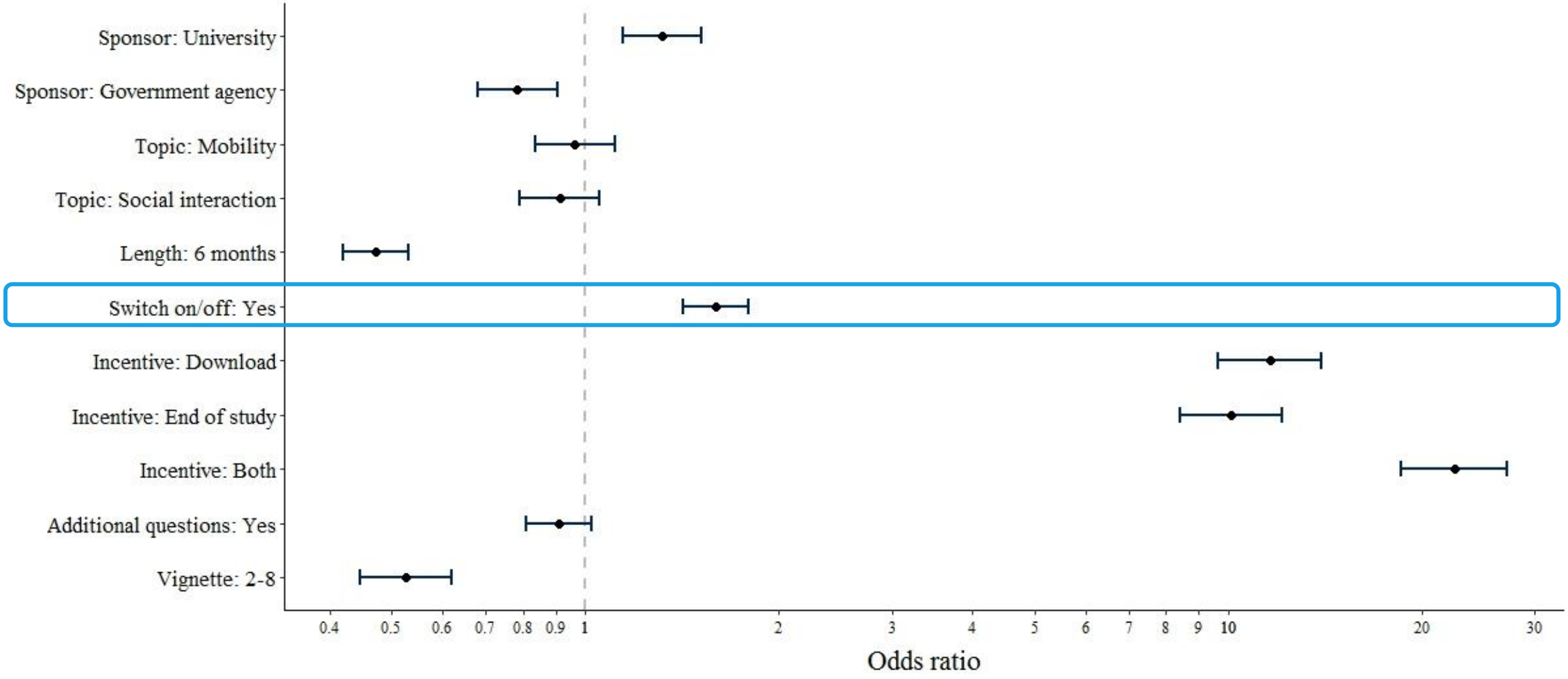
Logistic regression coefficients; *p<.001

Privacy Concerns: Summary

- Level of concern with smartphone data collection differs by type of data collected

Data	Concern	Role of participant	Control	Sensitivity of data
Smartphone usage GPS	High	Passive	Low	High
Activity data	Medium	Passive	Low	Low
Online survey Camera	Low	Active	High	High/Low

Giving autonomy to participants (Keusch et al. 2019)



Odds ratios with 95%-CI from multilevel logistic regression. n=1,947 German smartphone users

Giving autonomy to participants (Struminskaya et al. 2021)

Predictors	Sharing
Order (asked first)	0.02 **
Sponsor University	0.09***
Sponsor Market Research	n.s.
Benefit framing	–0.02*
Autonomy over data collect.	n.s.
Privacy	n.s.

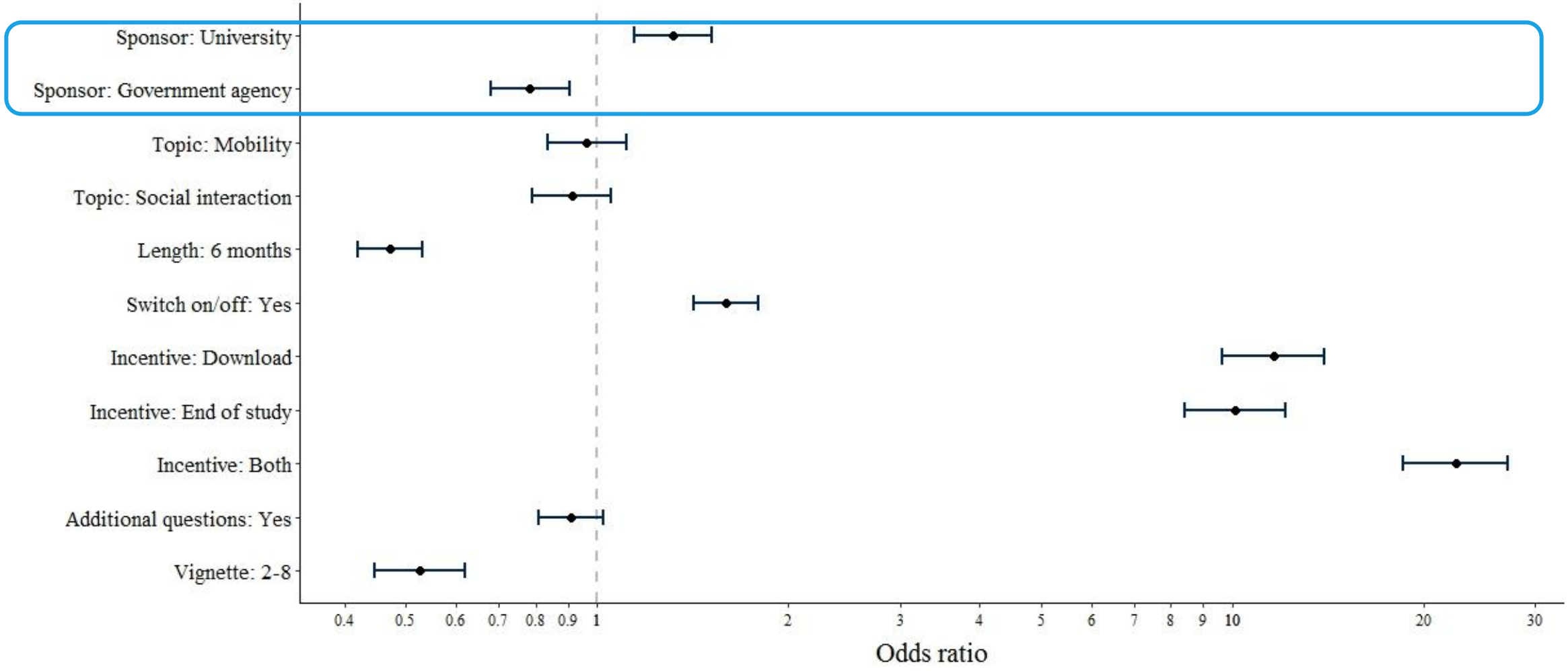
n=2,669; Average marginal effects; covariates not shown

Predictors	WTS GPS	Share GPS	Share video	Share photo house	Share photo receipt	Share photo self
Benefit framing	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.
Autonomy over data collection	.11***	-.06*	n.s.	n.s.	.04*	n.s.
Privacy	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.

n=1,853; Average marginal effects; covariates not shown

In all 3 studies: sig. effects of smartphone use behaviors, mixed findings about the effect of privacy concerns, attitudes toward surveys, prior app download

Sponsor effect (Keusch et al. 2019)



Odds ratios with 95%-CI from multilevel logistic regression. n=1,947 German smartphone users

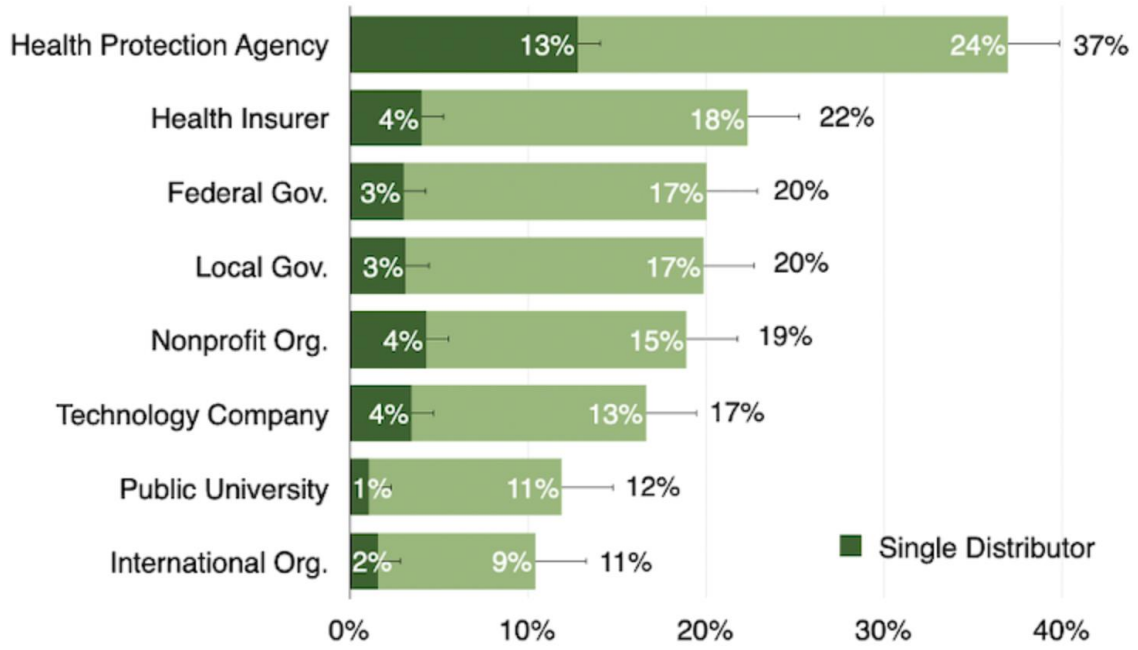
Sponsor effect
(Struminskaya et al. 2020)

Predictors	Sharing
Order (asked first)	0.02 **
Sponsor University	0.09***
Sponsor Market Research	n.s.
Benefit framing	–0.02*
Autonomy over data collect.	n.s.
Privacy	n.s.

n=2,669; Average marginal effects; covariates not shown

https://doi.org/10.1093/poq/nfaa044

Sponsor effect
(Hargittai et al. 2020)

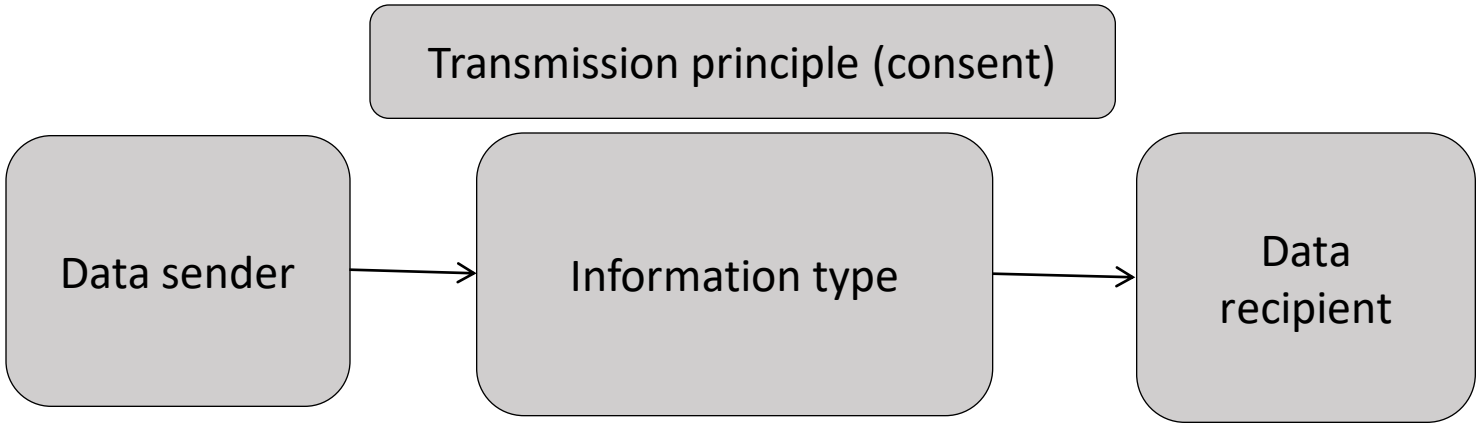


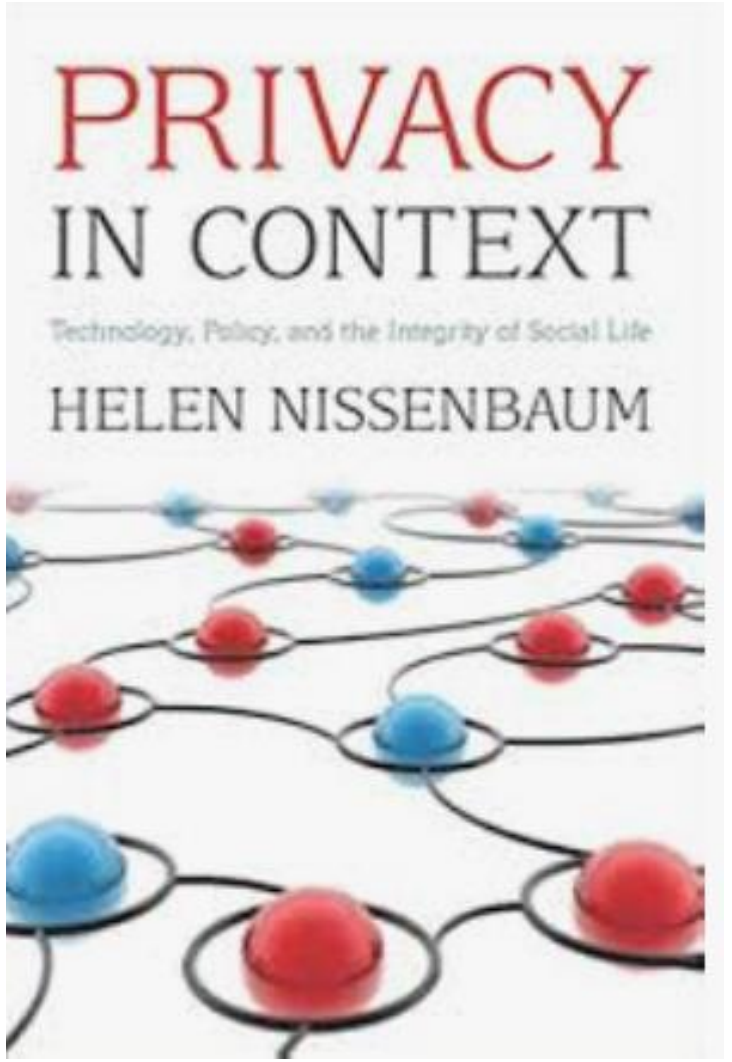
% willing to download contact tracing app if provided by listed distributor;
dark green = responses that only listed that option; light green = responses that also included at least one other option.

https://firstmonday.org/ojs/index.php/fm/article/view/11095/9985

Contextual integrity as a theoretical framework

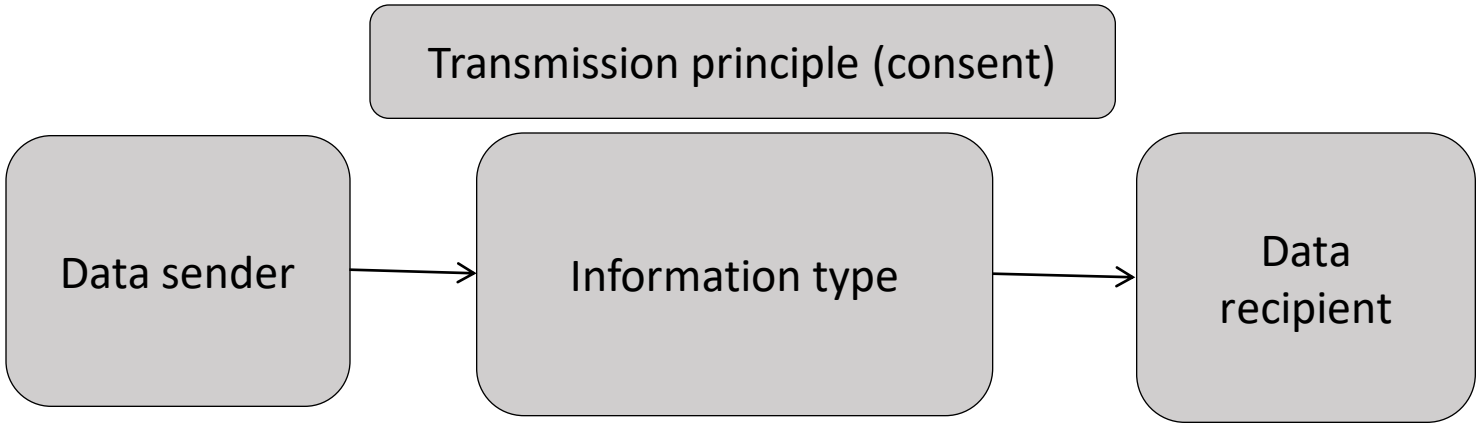
- Assumptions about the ‚audience‘
- Implicit trust/disrtust in the requesting entity
  - Who is asking for data?
  - With whom will it be shared?
  - How/how long will it be stored? Etc.

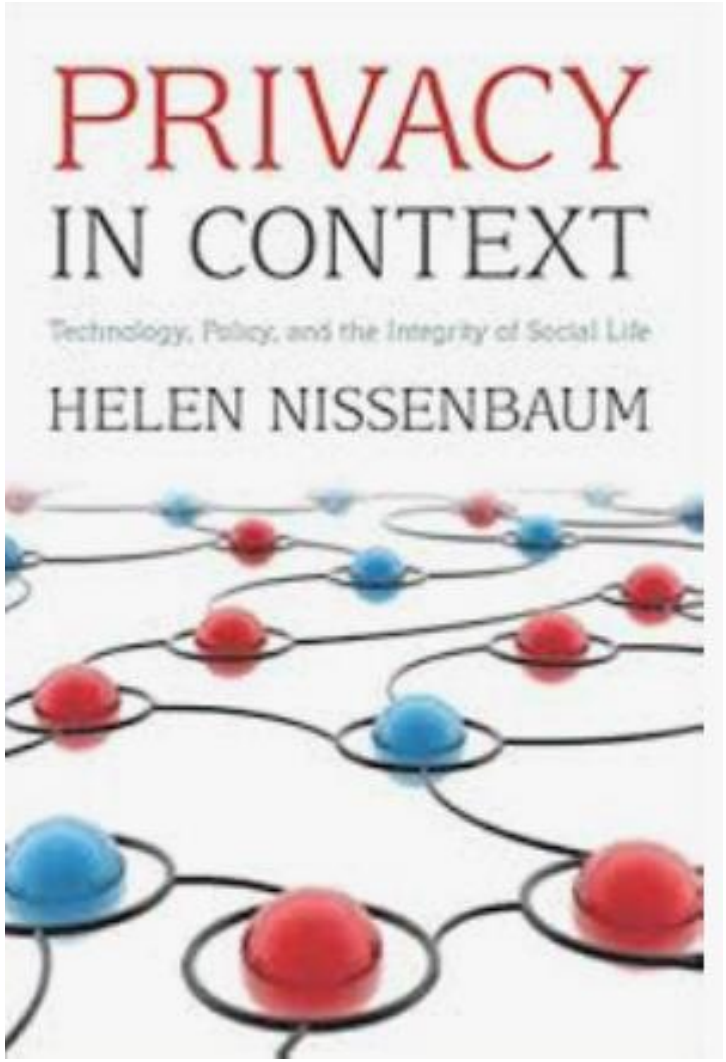




Contextual integrity as a theoretical framework

- Assumptions about the ‚audience‘
- Implicit trust/disrtust in the requesting entity
  - Who is asking for data?
  - With whom will it be shared?
  - How/how long will it be stored? Etc.





Watch Helen Nissenbaum’s Keynote at the 2023 MASS Workshop:
https://www.youtube.com/watch?v=NAXzGZOotiE

Understanding the consent request

When linking your responses to our questionnaires to information that Statistics Netherlands has available about you . . .	Answer (in %)		
	Correct	Incorrect	Don't know
a) your name, gender, and date of birth will be sent to Statistics Netherlands. [TRUE]	34.7	44.2	21.1
b) researchers (from outside Statistics Netherlands) get access to your name, gender, and date of birth. [FALSE]	68.3	11.4	20.3
c) your name, gender, and date of birth will be saved with the linked data. [FALSE]	34.7	39.2	26.1
d) for each project the linked data will always stay at Statistics Netherlands, and will be destroyed after completion of the specific project. [TRUE]	39.9	16.4	43.7
e) results of the study can be traced to you as an individual. [FALSE]	65.5	9.7	24.8
f) every researcher can consult the linked data via the Internet. [FALSE]	66.8	7.5	25.7
g) the Dutch Data Protection Authority supervises the linking and analyses of the data. [TRUE]	65.5	5.6	28.9

(Das & Couper 2014)

NETFLIX PRIZE

In 2006, Netflix offered $1 million to improve their show and movie recommendation system by 10 percent and provided access to a dataset containing over 100 million movie ratings from almost 500,000 Netflix subscribers.[1] Although other researchers developed new predictive algorithms, Narayanan and Shmatikov (2008) reidentified Netflix subscribers by using IMDb data. This breach in privacy (which allowed, for example, the prediction of people's sexuality by their watch history)[2] led to a lawsuit and the cancellation of a follow-up competition.[3]

AOL SEARCHER NO. 4417749

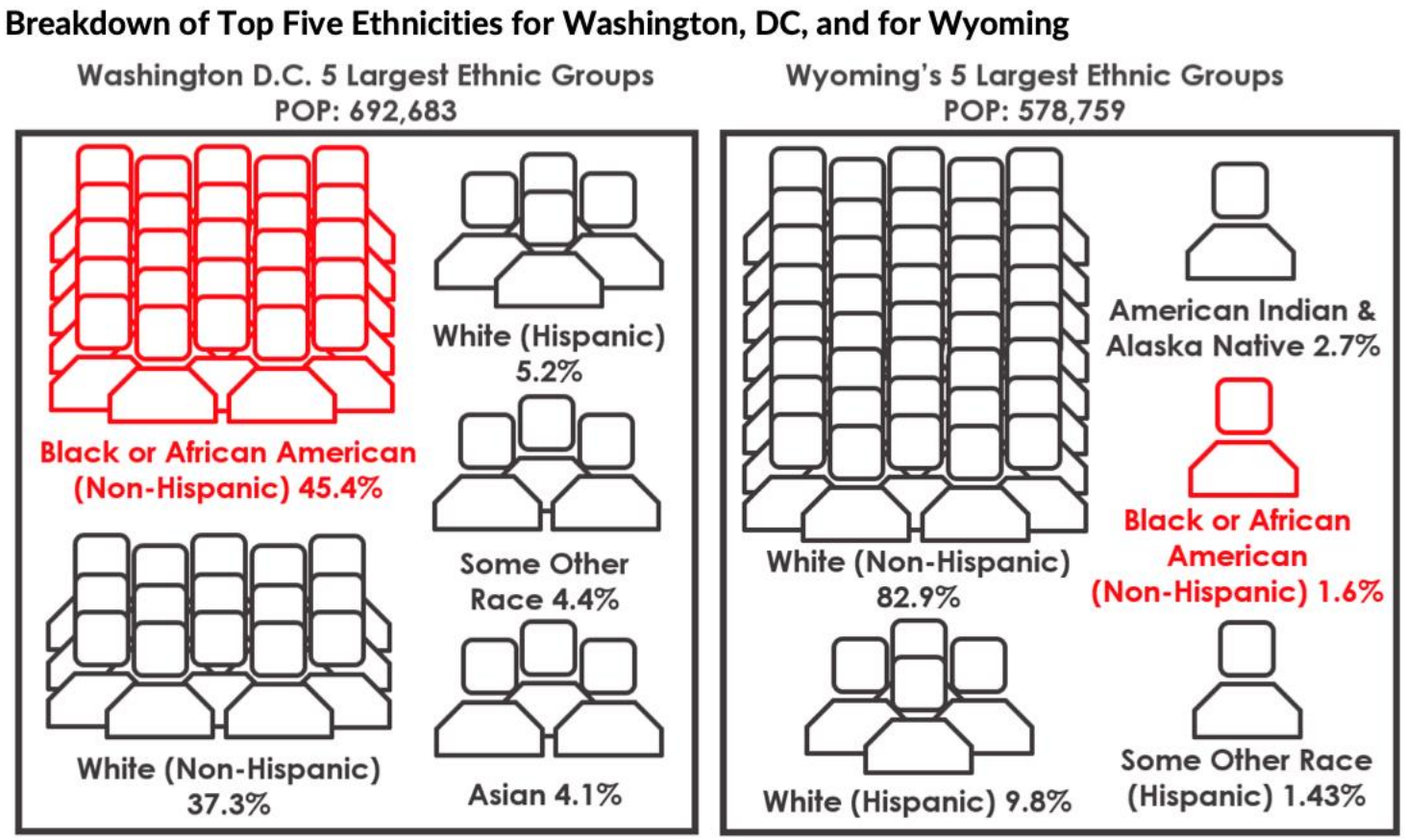
Also in 2006, AOL released an anonymized database of over 20 million web search queries for academic research. Although AOL removed the names and other personally identifiable information, the *New York Times* still identified one of the AOL users based on her searches such as "landscapers in Lilburn, Ga" and "homes sold in shadow lake subdivision gwinnett county georgia." Once identified, her other searches revealed more personal information that included "60 single men" and "dog urinates on everything."[4]

REIDENTIFICATION OF AMERICANS FROM HEALTH DATA

In 2019, computer scientists from the Imperial College London and Université Catholique de Louvain estimated that they correctly reidentified 99.98 percent of Americans from anonymized health data with 15 attributes including zip code, date of birth, gender, and number of children (Rocher, Hendrickx, and De Montjoye 2019).

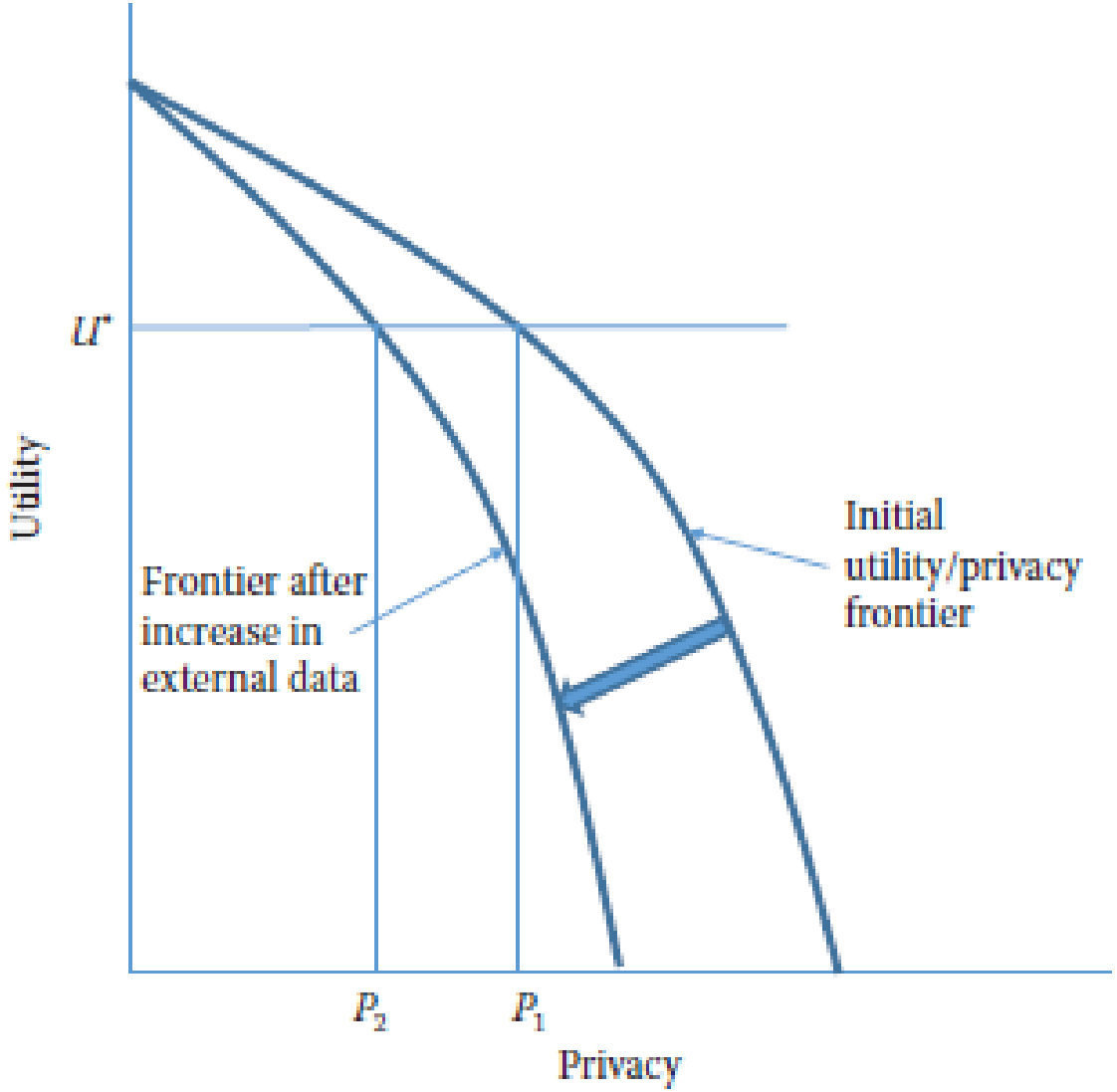
Bowen 2021: https://www.urban.org/sites/default/files/publication/104694/privacy-and-the-public-good_0_0.pdf

“Hiding” a respondent in the dataset



Bowen 2021, based on 2019 US Census data

Privacy-Utility Trade-off

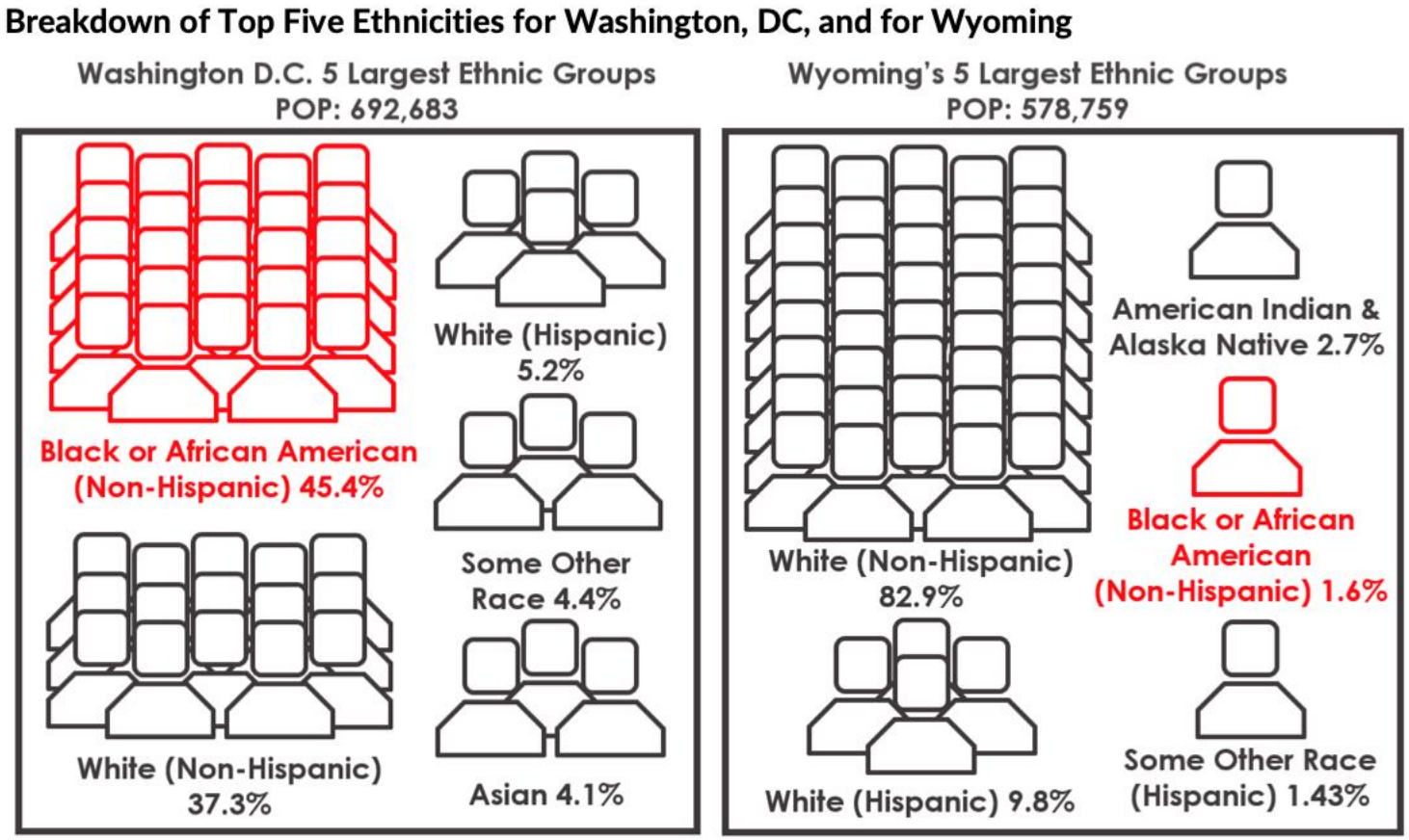


Example: if percentages are reported in aggregated form for a large number of people it is difficult to infer individual values, even if we know that an individual has contributed to the formation of such mean/percentage.
If these means are presented for subgroups or in multivariate tables, the risk for disclosure increases

Greater access → greater risk
Greater detail → greater risk
External data resources → increased risk

Bender et al. 2017, p. 300-301

“Hiding” a respondent in the dataset



Solutions:

1. Change the overall demographic representation significantly

2. Remove the specific population from the data

3. More options: see Bowen 2021

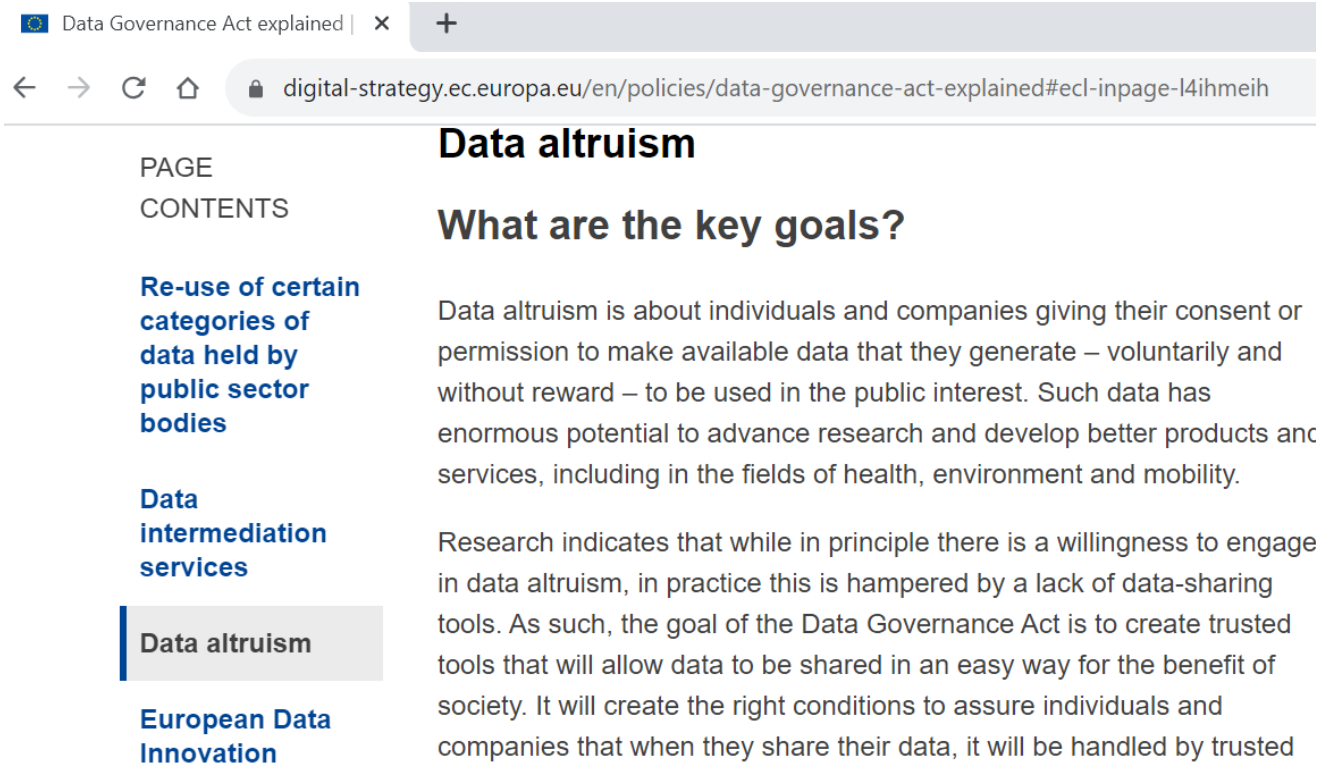
Bowen 2021, based on 2019 US Census data

What (else) to do?

- Need of new access models
  - Redaction / synthetic data & high level of access
  - Rich data but very restricted access → *What about open science?*
- Need to understand privacy & confidentiality concerns to be able to address them
- Further research into norms, wishes, interests of actors (data providers and data recipients)
- Address ethical and legal issues within organizations that protect privacy in a systematic manner
- Still a very new field!

Further steps

- Cross-discipline collaborations e.g., with legal scholars
- Cross-sectoral collaborations


Data Governance Act explained |

digital-strategy.ec.europa.eu/en/policies/data-governance-act-explained#ecl-inpage-l4ihmeih

PAGE CONTENTS

Re-use of certain categories of data held by public sector bodies

Data intermediation services

Data altruism

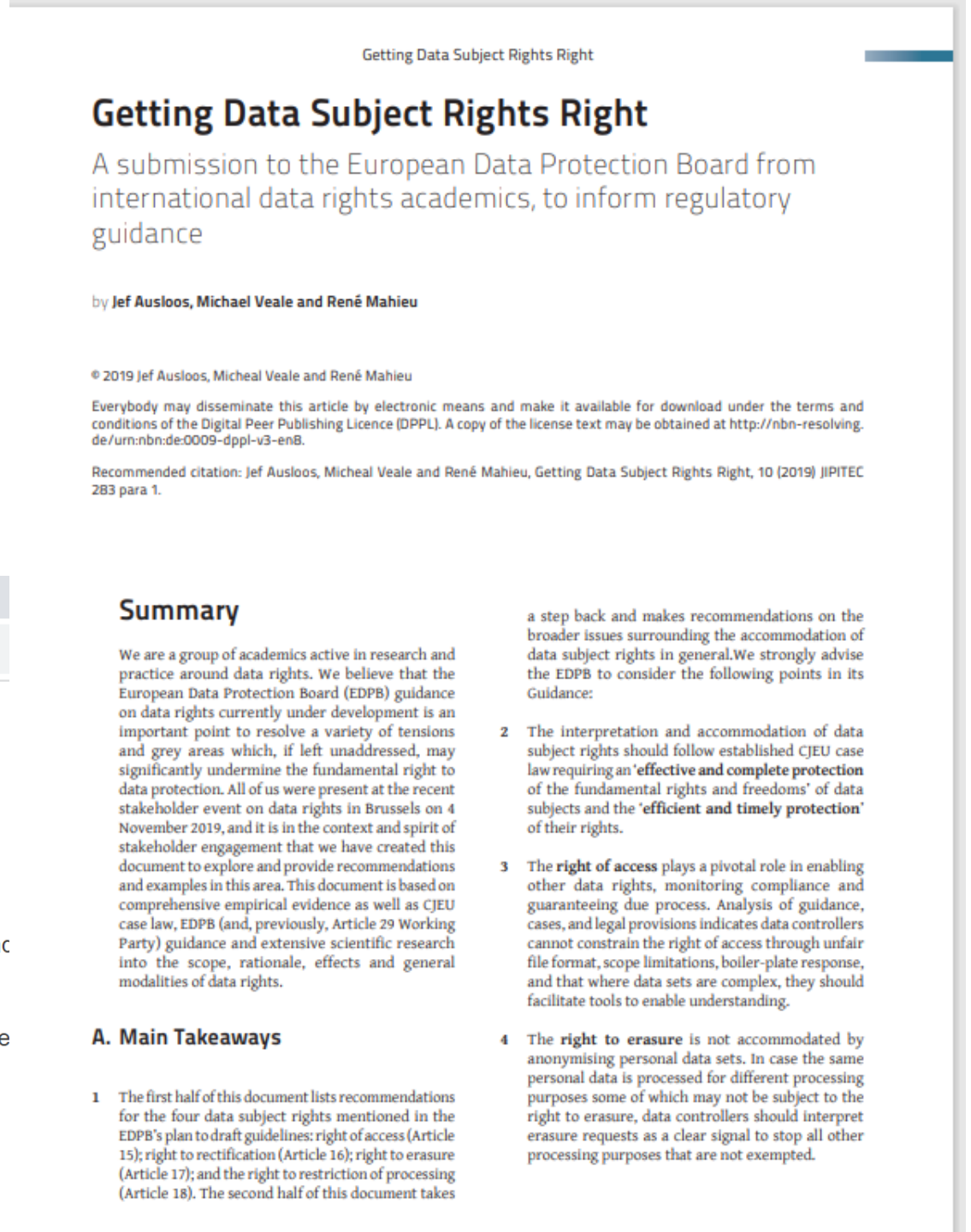
European Data Innovation

Data altruism

What are the key goals?

Data altruism is about individuals and companies giving their consent or permission to make available data that they generate – voluntarily and without reward – to be used in the public interest. Such data has enormous potential to advance research and develop better products and services, including in the fields of health, environment and mobility.

Research indicates that while in principle there is a willingness to engage in data altruism, in practice this is hampered by a lack of data-sharing tools. As such, the goal of the Data Governance Act is to create trusted tools that will allow data to be shared in an easy way for the benefit of society. It will create the right conditions to assure individuals and companies that when they share their data, it will be handled by trusted


Getting Data Subject Rights Right

Getting Data Subject Rights Right

A submission to the European Data Protection Board from international data rights academics, to inform regulatory guidance

by **Jef Ausloos, Michael Veale and René Mahieu**

Recommended citation: Jef Ausloos, Micheal Veale and René Mahieu, Getting Data Subject Rights Right, 10 (2019) JIPITEC 283 para 1.

Summary

We are a group of academics active in research and practice around data rights. We believe that the European Data Protection Board (EDPB) guidance on data rights currently under development is an important point to resolve a variety of tensions and grey areas which, if left unaddressed, may significantly undermine the fundamental right to data protection. All of us were present at the recent stakeholder event on data rights in Brussels on 4 November 2019, and it is in the context and spirit of stakeholder engagement that we have created this document to explore and provide recommendations and examples in this area. This document is based on comprehensive empirical evidence as well as CJEU case law, EDPB (and, previously, Article 29 Working Party) guidance and extensive scientific research into the scope, rationale, effects and general modalities of data rights.

A. Main Takeaways

1 The first half of this document lists recommendations for the four data subject rights mentioned in the EDPB's plan to draft guidelines: right of access (Article 15); right to rectification (Article 16); right to erasure (Article 17); and the right to restriction of processing (Article 18). The second half of this document takes a step back and makes recommendations on the broader issues surrounding the accommodation of data subject rights in general.We strongly advise the EDPB to consider the following points in its Guidance:

2 The interpretation and accommodation of data subject rights should follow established CJEU case law requiring an '**effective and complete protection** of the fundamental rights and freedoms' of data subjects and the '**efficient and timely protection**' of their rights.

3 The **right of access** plays a pivotal role in enabling other data rights, monitoring compliance and guaranteeing due process. Analysis of guidance, cases, and legal provisions indicates data controllers cannot constrain the right of access through unfair file format, scope limitations, boiler-plate response, and that where data sets are complex, they should facilitate tools to enable understanding.

4 The **right to erasure** is not accommodated by anonymising personal data sets. In case the same personal data is processed for different processing purposes some of which may not be subject to the right to erasure, data controllers should interpret erasure requests as a clear signal to stop all other processing purposes that are not exempted.

Codes of ethics and standards

- Code of ethics by World Association of Public Opinion Research (WAPOR) https://wapor.org/about-wapor/code-of-ethics/
- Code of ethics of the American Association of Public Opinion Research (AAPOR) https://www.aapor.org/Standards-Ethics/AAPOR-Code-of-Ethics.aspx
- ESOMAR, and the International Chamber of Commerce (ICC) https://www.esomar.org/what-we-do/code-guidelines
- Ethics Guidelines for Internet-mediated Research by the British Psychological Society https://www.bps.org.uk/news-and-policy/ethics-guidelines-internet-mediated-research-2017
- Laws and regulations developed by governments
  (e.g., the European General Data Protection Regulation, GDPR (EU) 2016/679)

Additional resources (see also Bender et al. 2017; Bowen 2021)

- The American Statistical Association's Privacy and Confidentiality website

http://community.amstat.org/cpc/home

- An overview of federal activities by the Confidentiality and Data Access Committee of the Federal Committee on Statistics and Methodology

http://fcsm.sites.usa.gov/committees/cdac/

- Data dissemination "best practices" by The World Bank and International Household Survey Network

http://www.ihsn.org/home/projects/dissemination

- *Journal of Privacy and Confidentiality* http://repository.cmu.edu/jpc
- Journal *Transactions in Data Privacy* http://www.tdp.cat/
- Workshops and conferences by The United Nations Economic Commission on Europe hosts and produces occasional reports

http://www.unece.org/stats/mos/meth/confidentiality.html

References / further reading

- Bender, S., Jarmin, R., Kreuter, F., & Lane, J. 2017. Privacy and confidentiality. In: Foster, I., Ghani, R., Jarmin, R. S., Kreuter, F., Lane, J. Big data and social science, CRC Press, pp. 299-311.
- Bowen, C. 2021. Personal privacy and the public good: Blancing data privacy and data utility. Urban Institute. https://www.urban.org/sites/default/files/publication/104694/privacy-and-the-public-good_0_0.pdf
- Joe, Kathy, Raben, Finn, & Phillips, Adam. 2016. The ethical issues of survey and market research. In C. Wolf, D. Joye, T. W. Smith, Y. Fu: The SAGE Handbook of Survey Methodology. London: SAGE Publications Ltd. Pp. 77-86. DOI: 10.4135/9781473957893
- Hargittai, E., Redmiles, E. M., Vitak, J. and Zimmer, M. Americans' willingness to adopt a COVID-19 tracking app: The role of app distributor. *First Monday*, Volume 25, No. 11-2 Nov 2020. https://firstmonday.org/ojs/index.php/fm/article/download/11095/9985; https://dx.doi.org/10.5210/fm.v25i11.11095
- 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, https://doi.org/10.1093/poq/nfz007
- Salganik, M. 2018. Bit by bit: Social research in the digital age. Princeton: Princeton University Press
- Singer, Eleanor, and Mick P. Couper. 2011. Ethical considerations in Internet surveys. In: M. Das, P. Ester, & L. Kaczmirek: Social and Behavioral Research and the Internet. London: Routledge Taylor & Francis Group. Pp. 133-162.
- Singer, E., and M. P. Couper. 2011. Ethical considerations in Internet surveys. In: M. Das, P. Ester, & L. Kaczmirek: Social and Behavioral Research and the Internet. London: Routledge Taylor & Francis Group. Pp. 133-162.
- Struminskaya, B., Toepoel, V., Lugtig, P., Haan, M., Luiten, A., Schouten, B. 2020, Understanding Willingness to Share Smartphone-Sensor Data, *Public Opinion Quarterly*, Vol. 84, Issue 3: 725–759, https://doi.org/10.1093/poq/nfaa044