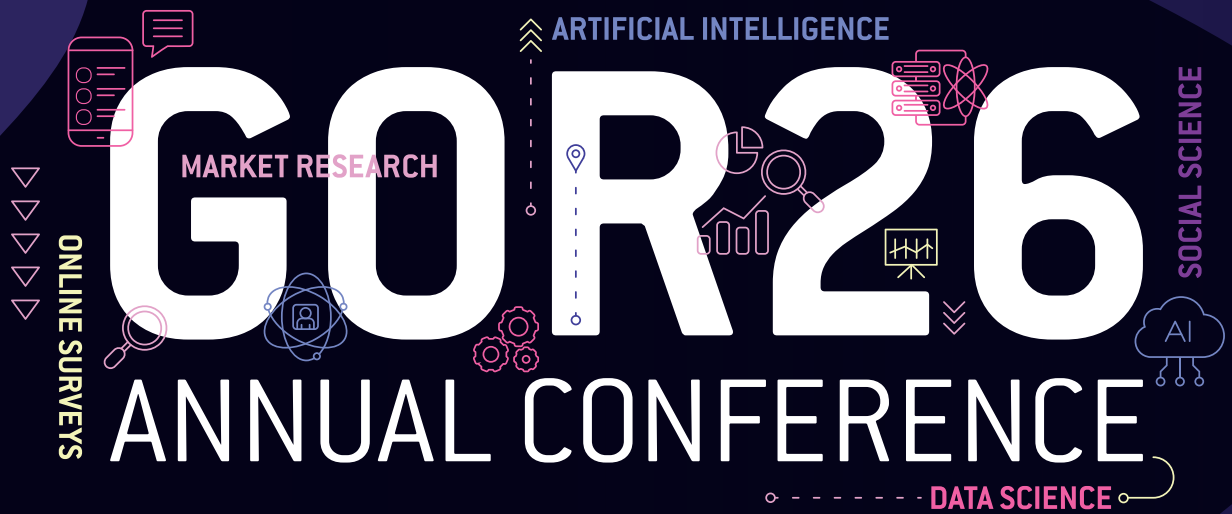
28TH GENERAL ONLINE RESEARCH CONFERENCE

26 - 27 FEBRUARY 2026 // ANNUAL CONFERENCE:
RHEINISCHE HOCHSCHULE COLOGNE, CAMPUS VOGELSANGER STRASSE

25 FEBRUARY 2026 // GOR WORKSHOPS:
GESIS – LEIBNIZ-INSTITUT FÜR SOZIALWISSENSCHAFTEN IN COLOGNE



GOR26
ANNUAL CONFERENCE

Rheinische Hochschule Köln

GOR
organized by DGOF

OTTO HELLWIG, SIMON KÜHNE, BELLA STRUMINSKAYA, CARINA CORNESSE, YANNICK RIEDER, LISA DUST, (EDS.)

28TH GENERAL ONLINE RESEARCH CONFERENCE PROCEEDINGS, COLOGNE 2026

ISBN 978 - 3 – 9822985 - 7 - 3

OTTO HELLWIG, SIMON KÜHNE, BELLA STRUMINSKAYA, CARINA CORNESSE, YANNICK RIEDER, LISA DUST, (EDS.)

GERMAN SOCIETY FOR ONLINE RESEARCH
DEUTSCHE GESELLSCHAFT FÜR ONLINE-FORSCHUNG (DGOF) E.V.
WWW.DGOF.DE

EDITING: Anna Hristova, Anna Conrad **LAYOUT AND TYPESET:** Nicole Propach, Köln

CONTENT

DGOF BOARD

DR. OTTO HELLWIG
Chairman DGOF Board,
Lakmoos AI, Hungary

DR. CARINA CORNESSE
GESIS Mannheim, Germany

LISA DUST
Civey, Germany

PROF. DR. SIMON KÜHNE
Bielefeld University, Germany

YANNICK RIEDER
Janssen-Cilag GmbH, Germany

PROF. DR. BELLA STRUMINSKAYA
Utrecht University, The Netherlands

PROGRAMME COMMITTEE

PROF. DR. SIMON KÜHNE
(Programme Chair), Bielefeld University

PROF. DR. BELLA STRUMINSKAYA
(Vice Programme Chair), DGOF Board & Utrecht University

DR. OTTO HELLWIG
DGOF Board & Lakmoos AI

DR. CARINA CORNESSE
DGOF Board & GESIS Mannheim

LISA DUST
DGOF Board & Civey

DR. FREDERIK FUNKE
datenmethoden.de & LimeSurvey GmbH

PROF. DR. FLORIAN KEUSCH
Mannheim University

YANNICK RIEDER
DGOF Board & The Janssen Pharmaceutical Companies of Johnson & Johnson

DR. OLAF WENZEL
Wenzel Marktforschung

CONFERENCE CHAIR

PROF. DR. SIMON KÜHNE
Programme Chair
DGOF Board & Bielefeld University

PROF. DR. BELLA STRUMINSKAYA
Vice Programme Chair
DGOF Board & Utrecht University

LOCAL PARTNER

RHEINISCHE HOCHSCHULE COLOGNE

DGOF OFFICE

ANNA HRISTOVA, ANNA CONRAD

GOR 26: INTERNATIONAL BOARD

DR. EVA AIZPURUA National Centre for Social Research

ALEXANDRA ASIMOV GESIS Leibniz Institute for the Social Sciences

KATJA BIRKE Produkt + Markt

KAREN BLANKE Destatis

PROF. MARCEL Das Centerdata

PROF. EDITH DE LEEUW Utrecht University

BEAT FISCHER intervista AG

DR. FREDERIK FUNKE Dr. Funke – SPSS & R Trainings

DR. GEORG-CHRISTOPH HAAS Institute for Employment Research

NIKLAS HAUPT Miios GmbH

FRANK HEUBLEIN Frank Heublein M3 Services

JULIA HOFFMANN Heidelberg Materials AG

PROF. ANNETTE HOXTELL Fachhochschule Erfurt

DR. WOJCIECH JABLONSKI Netherlands Interdisciplinary Demographic institute

STEFAN KNAUFF Bielefeld University

DR. SEBASTIAN KOCAR University College London

FABIENNE KRAEMER GESIS Leibniz Institute for the Social Sciences

DR. THOMAS KRÜGER Umfragezentrum Bonn – Prof. Rudinger GmbH (uzbonn GmbH)

DR. TANJA KUNZ GESIS Leibniz Institute for the Social Sciences

ANDRÉ LANG Insius

PROF. BETTINA LANGFELDt Kassel University

DANIIL LEBEDEV GESIS Leibniz Institute for the Social Sciences

PROF. JI-PING LIN Academia Sinica

JAN MACKEBEN Institut für Arbeitsmarkt- und Berufsforschung

DR. OLGA MASLOVSKAYA University of Southampton

EMANUEL MAXL Context-Research

PROF. NATALJA MENOLD Dresden University of Technology

PROF. GUSTAVO S. MESCH University of Haifa

LONG NGUYEN Bielefeld University

MAREIKE OEHRL september Strategie & Forschung GmbH

DANIELLE REMMERSWAAL Utrecht University & CBS

DR. TOBIAS RETTIG University of Mannheim

FRANZISKA RIEDER Omnicom Media Group Germany GmbH

DR. EIKE MARK RINKE University of Leeds

DR. JOSS ROSSMANN GESIS Leibniz Institute for the Social Sciences

CAMILLA SALVATORE Utrecht University

GERO SERFAS Q Agentur für Forschung

DR. KAZUHIKO SHIBUYA Alma Mater Europaea

ROBIN SPICER Bilendi GmbH

DR. MARKUS STEINBRECHER Bundeswehr Center for Military History and Social Sciences

PROF. BELLA STRUMINSKAYA Utrecht University

DR. TABEA TESCH Royal Canin (Mars Inc.)

DORIAN TSOLAK Bielefeld University

CHINTAN TURAKHIA SSRS

DR. ALEXANDER WENZ University of Mannheim

DR. OLAF WENZEL Wenzel Marktforschung

DR. GEORG WITTENBURG Inspirient GmbH

ANNE ZIMMER RTL Deutschland GmbH

ZAZA ZINDEL Bielefeld University

DEAR GUESTS OF THE GOR 26!

WELCOME TO THE 28TH EDITION OF THE GENERAL ONLINE RESEARCH CONFERENCE. WE ARE VERY HAPPY TO HOST THIS YEAR'S GOR IN COOPERATION WITH THE RHEINISCHE HOCHSCHULE KÖLN.

BACK TO OUR ROOTS FOR THE SEVENTH TIME

This year marks the seventh time that the GOR has visited its founding city of Cologne. This year, the Rheinische Fachhochschule is opening its doors to us as a partner of the GOR again at its Campus Centre in Cologne-Ehrenfeld.

Rheinische Hochschule Köln (RH) is a private, state-recognized university of applied sciences located in Cologne, North Rhine-Westphalia, Germany. It was officially established in 1971 and continues the tradition of the Rheinische Ingenieurschule für Bau- und Maschinenbauwesen, which was founded in 1958. The institution offers undergraduate and graduate degree programs in areas such as engineering, business administration, law, social sciences, and health. RH is organized under private sponsorship and is subject to state supervision according to the Higher Education Act of North Rhine-Westphalia. The university maintains several campuses in Cologne and has a student population of approximately 5,700.

As in previous years, we have a great conference programme lined up for you including keynotes, discussions, presentations, awards, posters, workshops and much more. Again we come with a structure across thematic categories into sessions. In doing so, we want to break down the silos and network people even more closely. This is entirely in line with our claim 'Connecting People: Innovating Market Research and Social Data Science'. You can choose between four parallel sessions and one invited session. In addition, there are three award competitions: i) the GOR Impact & Innovation Award 2026 for the most impactful market research case, ii) the GOR Thesis Award 2026 for the best thesis in online research, iii) the GOR Poster Award 2026 for the best poster of the conference. The DGOF Best Paper Award 2026 for the best paper in online research will also be awarded at GOR. In addition, this year's GOR will honour the winner of the 'Best of FAMS Award', an annual competition organised by the industry associations ADM, BVM and DGOF for the best final theses by trainees studying to become market and social research specialists (FAMS).

Our keynotes this year take different looks at the development of our profession. On Thursday, 26th February, Dr. Katharina Schüller, Founder & Owner of STAT-UP Statistical Consulting & Data Science will talk about "More efficiency, new problems? AI in empirical research" and on Friday, 27th February, Prof. Dr. Peter Lugtig from Utrecht University, will welcome us to the age of data integration and discuss 'But how to do it?'.

On the pre-conference day, Wednesday, 25th February, four workshops will take place located at the GESIS Office Cologne. We will conclude Thursday's conference with three events. First, at 5:15 p.m., the DGOF general meeting will take place at the RFH. At 7 p.m., the Early Career Speed Networking Event will begin at the Blue Shell, where we will kick off this year's GOR Party at 8 p.m. in cooperation with the 'Middance Crisis' party series.

We are particularly grateful for the enthusiastic support of and collaboration with our partners at Rheinische Hochschule Köln: Prof. Dr. Christian Bosau and his team. We would also like to thank our sponsors and media partners. And, of course, a big THANKS to you, the conference participants, presenters, and speakers at this event!

HAVE A GREAT TIME AT THE GENERAL ONLINE RESEARCH CONFERENCE 2026!
Dr. Otto Hellwig

DEUTSCHE GESELLSCHAFT FÜR ONLINE-FORSCHUNG E.V.

ONLINE RESEARCH IS A DYNAMIC, INNOVATIVE FIELD, WITH CONSTANTLY EMERGING CHALLENGES AS WELL AS OPPORTUNITIES FOR RESEARCH AND PRACTICE.

The German Society for Online Research (Deutsche Gesellschaft für Online Forschung) (DGOF) is a modern, innovative association, which has focused on the interests of the actors in the field of online research since its establishment in 1998.

It is the association's goal to be the leader in this field. DGOF seeks to bridge different research fields (such as sociology, psychology, political science, economics, market and opinion research, data science) using online research methods and facilitates the transfer between academic research and the industry.

DGOF campaigns for the establishment and the development of online research as well as the interests of online researchers in Germany. Online research ranges from online based data collection methods (e.g., web surveys in online panels); to mobile research with smartphones, tablets, and wearables; to the collection and analysis of social media data, administrative data, data from passive measurements, and other big data sources.

DGOF organizes the General Online Research (GOR) conference and the Research Plus event series which support professional and collegial exchanges between researchers and practitioners across academia and the industry. By bringing together scientific findings, commercial needs, and practical applications for best practices, DGOF provides a sustainable input for further developments in online research.

CHANGE THROUGH INNOVATION IS A KEY CHARACTERISTIC OF OUR RESEARCH FIELD. DGOF IS A FACILITATOR FOR THIS CHANGE: WWW.DGOF.DE

1 DGOF MEANS DEVELOPMENT:
Online research is more than just web surveys. We constantly expand our portfolio and our expertise with the development, encouragement, and establishment of innovative digital methods, passive measurement, and big data methods. In addition, we focus on the relationship between the Internet and society.

2 DGOF CONNECTS:
We are a bridge between different research disciplines and across commercial applications.

3 DGOF IS DIVERSE:
We support our members' interests, for the dissemination of knowledge, for exchange, and for discussion, as well as for the establishment and implementation of scientific standards.

4 DGOF IS INNOVATIVE:
We are a facilitator of new issues such as big data and data science.

5 DGOF IS DISRUPTIVE:
We support change. It is our practice to foster acceptance for new methods in research, and we are always on the lookout for new developments.

DGOF

DEUTSCHE GESELLSCHAFT FÜR ONLINE-FORSCHUNG – DGOF E. V.
GERMAN SOCIETY FOR ONLINE RESEARCH
HUHNSGASSE 34B
50676 COLOGNE (GERMANY)

PHONE: +49 (0)221 272318180
E-MAIL: OFFICE@DGOF.DE
WWW.DGOF.DE // WWW.GOR.DE

DR OTTO HELLWIG served as CEO of respondi AG from its founding in 2005 until the sale of the respondi Group to Bilendi in 2021. Following the acquisition, he took on the role of Corporate Integration Director at Bilendi & respondi.

Since 2025, he has been Manager DACH at Lakmoos AI. Dr. Hellwig has been active in the field of market and social research since the early 1990s. He holds degrees in social sciences, psychology, and media studies, and worked for several years as a research assistant at the Institute for Applied Social Research at the University of Cologne, where he earned his doctorate in 2000. He has served as Chairman of the DGOF Board since March 2013.

PROF. DR. SIMON KÜHNE is Professor of Applied Socia Data Science at the Faculty of Sociology at Bielefeld University.

His methodological research is focused on Survey Methodology and Computational Social Science, esp. social media. He also works on aspects of social inequality in the areas of sexual and gender diversity, racism and discrimination, and health. Simon Kühne has been actively involved in shaping the GOR for several years. After serving as a reviewer and track chair, he has shared responsibility for the program as (Vice) Program Chair since 2022. He has been a member of the DGOF Board since 2023 and he is the Vice Programme Chair of the GOR 2025 Conference.

PROF. DR. BELLA STRUMINSKAYA is associate professor at the department of Methodology and Statistics at Utrecht University, The Netherlands, and a researcher at Statistics Netherlands. Her research focuses on innovations in data collection such as using apps, wearables, and sensors in surveys and official statistics, as well as data donation of digital trace data.

She is a country team lead and member of the Management Board of Survey of Health, Ageing and Retirement in Europe (SHARE), member of Methods Advisory Board of the European Social Survey (ESS) and member of the Scientific Advisory Board of Statistics Sweden.

Bella Struminskaya has been a board member of the German Society for Online Research (DGOF) since 2017 and is the programme chair of the GOR 25 conference.

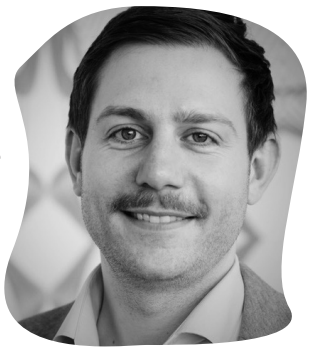
DR. CARINA CORNESSE is Head of The Department Survey Design and Methodology at GESIS Leibniz Institute for the Social Sciences. Previous to her current role, she was a guest professor of Social Stratification and Survey Methodolgoy at Free University (FU) Berlin and held different positions at the German Institute for Economic Research, University of Bremen, and University of Mannheim.

Her research focuses on survey methodology, digital inequality, and online research. She is particularly interested in innovative methods for recruiting and surveying individuals and households as well as achieving inclusiveness and diversity in survey samples given unequal digital opportunities, habits, and preferences. She is a member of the DGOF Board as well as the chair of the DGOF Research Funding committee since 2023 and local host of GOR 2025.

LISA DUST is Senior Analytics Consultant at Civey. Her focus is on translating data into strategic, actionable insights that enable effective decision-making. She has over 20 years of experience across corporate and agency roles in market research, analytics, and consultancy.

Lisa has been a member of the DGOF since 2025 and regularly serves as a juror at scientific market research conferences.

YANNICK RIEDER is Emea Market Research Manager at Johnson & Johnson Innovative Medicine.



His passion is implementing innovative approaches to gain in-depth customer and market insights and translating them into actionable strategies. He has successfully done this for over 10 years on both the agency and corporate sides. Yannick is a DGOF Board Member since 2023.

DEAR PARTICIPANTS

DEAR PARTICIPANTS OF THE GOR 2026,

it is a real pleasure to welcome you to the General Online Research Conference 2026 – and to do so once again at the Rheinische Hochschule Köln - University of Applied Sciences. That the DGOF has chosen our university as host for the second time is something we truly appreciate. We see it as a sign of trust and as confirmation that the GOR and the RH are a very good match.

As someone who has been part of the DGOF community for many years and who has attended numerous GOR conferences as a researcher, it is especially nice for me to welcome you this time again as the local host. Together with the DGOF – German Society for Online Research, we are delighted to create a setting where excellent research, practical relevance, and open exchange come together.

The RH stands for applied research, practice-oriented teaching, and close connections to industry and society. This makes hosting the GOR particularly fitting – after all, online research thrives on exactly this interplay between theory, methods, data, and real-world application. We strongly believe that good ideas emerge where different perspectives meet, and conferences like the GOR provide the perfect space for this.

There is also hardly a better place for a conference on online research than Cologne. The city and the whole Rhine area are one of the central hubs for online research in Germany, home to numerous agencies, institutes, panels, and technology providers. Many of the organizations shaping our field are based here – which makes Cologne not only easy to reach, but also highly relevant from a professional point of view. And of course, Cologne is always worth a visit: open, lively, welcoming, and never boring.

The GOR has always been more than just a conference program – it is a community. Beyond keynotes, talks, and workshops, it is the conversations in the hallways, during coffee breaks, or late in the evening that often leave the strongest impression. And speaking of evenings: the GOR party has long been a tradition and a highlight for many participants. It is a wonderful reminder that networking, collaboration, and good ideas also benefit from relaxed conversations, music, and a shared drink.

I encourage you to make the most of the coming days – engage in discussions, ask questions, reconnect with familiar faces, and meet new colleagues. Take time for exchange, curiosity, and maybe even a little bit of Cologne's famous "rheinische Gelassenheit".

I hope you will experience the GOR 2026 not only as a successful conference, but as an inspiring few days filled with knowledge, encounters, and memorable moments. We are very happy to have you here and wish you a stimulating, enjoyable, and unforgettable stay at the RH and in Cologne.

Warm regards,
PROF. DR. CHRISTIAN BOSAU
Local Host GOR 2026
Rheinische Hochschule Köln

ORGANIZERS

DGOF

Rheinische Hochschule Köln

SPONSORS

Bilendi

Datapods

ESRA
EUROPEAN SURVEY RESEARCH ASSOCIATION

GIM. BETTER INSIGHTS.

GESIS Leibniz Institute for the Social Sciences

horizoom
people first.

infas

inspirient
cognitive analytics

MNFORCE

moweb research

norstat

produkt+markt
marketing research

TESTSET

talk
ONLINE PANEL

verian

xelper

MEDIA PARTNER

marktforschung.de

COOPERATION PARTNERS

succeet
insights | data | analytics
2026
18 – 19 March, 2026 @ RMCC
Wiesbaden/Frankfurt, Germany

WdM
9–17 June, 2026
by succeet meet & succeed

WEDNESDAY, 25/FEB/2026

WEDNESDAY

11.30 AM - 12.30 PM

BEGIN WORKSHOP CHECK-IN GESIS – Leibniz-Institut für Sozialwissenschaften Köln

12.30 PM - 03.00 PM

WORKSHOP 1
GESIS, West I

WORKSHOP 2
GESIS, West II

03.00 PM - 03.30 PM

BREAK GESIS – Leibniz-Institut für Sozialwissenschaften Köln

03.30 PM - 06:00 PM

WORKSHOP 3
GESIS, West I

THURSDAY, 26/FEB/2026

THURSDAY

08:00 AM - 09:00 AM

BEGIN CHECK-IN Rheinische Hochschule, Campus Vogelsanger Straße

09:00 AM - 10:00 AM

1: OPENING AND KEYNOTE 1: DR. KATHARINA SCHÜLLER Location: RH, Auditorium

10:00 AM - 10:15 AM

BREAK RH, Lunch Hall/ Cafeteria

10:15 AM - 11:15 AM

2.1: AI AND SURVEY RESEARCH
RH, Seminar 01

2.2: PARADATA AND METADATA
RH, Seminar 02

2.3: MEDIA STUDIES
RH, Seminar 03

2.4: INNOVATION IN MEASUREMENT INSTRUMENTS
RH, Seminar 04

GOR THESIS AWARD MASTER
RH, Auditorium

11:15 AM - 11:30 AM

BREAK RH, Lunch Hall/ Cafeteria

11:30 AM - 12:30 PM

3.1: APP-BASED DATA COLLECTION
RH, Seminar 01

3.2: VIDEO AND IMAGES IN SURVEY RESEARCH
RH, Seminar 02

3.3: MARKET AND CUSTOMER RESEARCH
RH, Seminar 03

3.4: ONLINE RESEARCH ON YOUTH AND MENTAL HEALTH
RH, Seminar 04

GOR THESIS AWARD PHD
RH, Auditorium

12:30 PM - 01:30 PM

LUNCH BREAK RH, Lunch Hall/ Cafeteria

01:30 PM - 02:30 PM

4.1: POSTER SESSION
RH, Auditorium

4.2: POSTER SESSION
RH, Auditorium

4.3: POSTER SESSION
RH, Auditorium

4.4: POSTER SESSION
RH, Auditorium

4.5: POSTER SESSION
RH, Auditorium

02:00 PM - 03:45 PM

GOR IMPACT AND INNOVATION AWARD
RH, Seminar 03

02:00 PM - 03:30 PM

5.1: DATA QUALITY AND MEASUREMENT ERROR I
RH, Seminar 03

5.2: ONLINE PANELS I
RH, Seminar 02

5.3: AI AND SOCIETY
RH, Seminar 04

03:30 PM - 04:00 PM

BREAK RH, Lunch Hall/ Cafeteria

04:00 PM - 05:00 PM

6.1: DATA QUALITY AND MEASUREMENT ERROR II
RH, Seminar 01

6.2: ONLINE PANELS II
RH, Seminar 02

6.3: MODELLING PEOPLE, INFORMING POLICY: NEW APPROACHES IN THE AI ERA
RH, Seminar 02

05:15 PM - 06:30 PM

DGOF: MEMBER MEETING RH, Seminar 01

07:00 PM - 08:00 PM

EARLY CAREER SPEED NETWORKING EVENT Blue Shell

08:00 PM - 02:00 AM

GOR 26 PARTY Blue Shell

FRIDAY, 27/FEB/2026

FRIDAY

08:00 AM - 09:00 AM

BEGIN CHECK-IN Rheinische Hochschule, Campus Vogelsanger Straße

09:00 AM - 10:00 AM

7.1: AI AND QUALITATIVE RESEARCH
RH, Seminar 01

7.2: NEW INSIGHTS ON SATISFICING
RH, Seminar 02

7.3: DESIGNING INCLUSIVE AND ENGAGING SURVEYS
RH, Seminar 03

10:00 AM - 10:15 AM

BREAK RH, Lunch Hall/ Cafeteria

10:15 AM - 11:00 AM

8: KEYNOTE 2: PROF. DR. PETER LUGTIG RH, Auditorium

11:00 AM - 11:30 PM

9: AWARD CEREMONY RH, Auditorium

11:45 AM - 12:00 PM

BREAK RH, Lunch Hall/ Cafeteria

12:00 PM - 01:00 PM

10.1: SMART SURVEYS AND INTERACTIVE SURVEY FEATURES
RH, Seminar 01

10.2: DATA DONATION
RH, Seminar 02

SOCIAL MEDIA RECRUITMENT
RH, Seminar 03

10.4: CURATED SESSION: COLLECT, SHARE, ACT: THE POWER OF ACTIVATED KNOWLEDGE
RH, Auditorium

01:00 PM - 02:00 PM

LUNCH BREAK RH, Lunch Hall/ Cafeteria

02:00 PM - 03:00 PM

11.1: SAMPLING AND WEIGHTING
RH, Seminar 01

11.2: ENSURING PARTICIPATION
RH, Seminar 02

11.3: INFERENTIAL LEAP: FROM DIGITAL TRACE DATA TO MEASURING CONCEPTS
RH, Seminar 03

11.4: DGOF KI (AI) FORUM: INSPIRATION SESSION (HELD IN GERMAN)*
RH, Auditorium

03:00 PM - 03:15 PM

BREAK RH, Lunch Hall/ Cafeteria

03:15 PM - 04:15 PM

12.1: SURVEY RECRUITMENT
RH, Seminar 01

12.2: PUSH TO WEB AND MIXED MODE SURVEYS
RH, Seminar 02

12.3: METHODS, TOOLS, AND FRAMEWORKS - A BIRD'S VIEW ON DATA COLLECTION
RH, Seminar 03

WEDNESDAY, 25/FEB/2026
11:30 - 12:30 PM BEGIN WORKSHOP CHECK-IN
GESIS – LEIBNIZ-INSTITUT FÜR SOZIALWISSENSCHAFTEN KÖLN

WORKSHOP 1

12:30 - 03:00 PM
GESIS, WEST I

SESSION CHAIR: FREDERIK FUNKE

HANDS-ON WORKSHOP ON AUTOMATED QUALITATIVE INTERVIEWS

BRUNO RECHT
Userflix, Germany; bruno@getuserflix.com

TARGET GROUPS
Market research professionals and agencies, UX researchers and product teams, "People Who Do Research" (PWDR) in product organizations

IS THE WORKSHOP GEARED AT AN EXCLUSIVELY GERMAN OR AN INTERNATIONAL AUDIENCE?
International audience

WORKSHOP LANGUAGE
English

DESCRIPTION OF THE CONTENT OF THE WORKSHOP
his hands-on workshop demonstrates how autonomous voice AI moderators can conduct, transcribe, and analyze qualitative interviews at scale - bridging the traditional divide between qualitative depth and quantitative reach. Participants will experience the complete workflow: (1) Co-creating a research study through AI-assisted question development and visual stimulus integration; (2) Observing live autonomous voice-to-voice interviews with dynamic follow-up questions and multilingual capabilities across 50+ languages; (3) Exploring real-time analysis that extracts patterns across small and large sample sizes while maintaining traceability to individual voices. Through hands-on configuration of their own study, participants will critically evaluate when autonomous voice AI maintains research rigor versus when human moderators remain essential. Workshop covers end-to-end orchestration, quality validation through full transcript access, GDPR compliance, and integration with existing research workflows.

GOALS OF THE WORKSHOP PARTICIPANTS WILL:
(1) Familiarize themselves with autonomous voice AI moderation as an emerging research methodology—understanding the complete workflow from study co-creation to real-time analysis; (2) Develop evaluation frameworks for assessing AI-moderated interview quality using real pilot transcripts and comparison criteria; (3) Apply co-intelligence principles: learning when to leverage AI autonomy versus when human researcher judgment remains essential; (4) Explore current capabilities and limitations; (5) Gain implementable insights from real deployments: cost models, change management, quality validation, and GDPR compliance; (6) Envision the future of customer research: from quarterly bottleneck to continuous insight engine while maintaining methodological rigor.

NECESSARY PRIOR KNOWLEDGE OF PARTICIPANTS
Basic familiarity with qualitative research methods (moderated interviews, user research, or market research). No technical or AI expertise required. Workshop designed for research professionals, moderators, product teams conducting research, and organizations exploring research scaling.

LITERATURE THAT PARTICIPANTS NEED TO READ PRIOR TO PARTICIPATION
None required. Workshop is designed to be self-contained with all context provided during the session.

RECOMMENDED ADDITIONAL LITERATURE
Co-Intelligence - Ethan Mollick, Research that Scales - Kate Towsey

INFORMATION ABOUT THE INSTRUCTOR
Bruno Recht is CEO and co-founder of Userflix, which develops autonomous voice AI moderators for qualitative interviewing. He serves as guest lecturer for Human AI Interaction at Elisava University Barcelona. Previously, he worked as a Designer at Porsche. Userflix's validation includes pilot deployments with the Nielsen Norman Group, IKEA, and leading market research agencies. Recent recognition: marktforschung.de Innovation Award 2025 and RWTH Spin-off Award. Bruno brings a unique combination of design thinking, AI product development, and hands-on implementation experience with research organizations - bridging academic rigor with commercial deployment insights. Maximum number of participants 25-30 participants (to ensure everyone can engage hands-on with the platform and receive individual attention during the practical exercises).

WILL PARTICIPANTS NEED TO BRING THEIR OWN DEVICES IN ORDER TO BE ABLE TO ACCESS THE INTERNET? WILL THEY NEED TO BRING ANYTHING ELSE TO THE WORKSHOP?
Yes, participants should bring their own laptop or tablet with internet connection to access the Userflix platform during hands-on exercises. No software installation required (browser-based). Optional: participants may bring headphones for better audio experience during demo interviews.

WORKSHOP 2

12:30 - 03:00 PM
GESIS, WEST II

SESSION CHAIR: FREDERIK FUNKE

VIBE CODING FOR ONLINE RESEARCH: FROM FINDINGS TO INTERACTIVE, OPEN, AND PARTICIPATORY PORTALS

ALI REZA HUSSAINI1, ASLI TELLI2
1Universität Leipzig, Germany; 2Universität zu Köln;
ali_reza.hussaini@uni-leipzig.de

DURATION OF THE WORKSHOP
2.5 h

TARGET GROUPS
Applied researchers, data scientists, methodologists, and research communicators who want to convert datasets, literature syntheses, or analysis outputs into interactive dashboards and tools without deep full-stack expertise.

IS THE WORKSHOP GEARED AT AN EXCLUSIVELY GERMAN OR AN INTERNATIONAL AUDIENCE?
International Audience

WORKSHOP LANGUAGE
English

DESCRIPTION OF THE CONTENT OF THE WORKSHOP
This practical workshop will introduce researchers to vibe coding, a rapidly emerging AI-assisted approach that allows researchers to analyze, design, and share their work through natural language interaction, rather than conventional programming.

Vibe coding is a collaborative and iterative workflow where large language models generate either analytic insight, functional code, or web interfaces based on the goals and feedback of the researcher. In practice, it bridges the entire research cycle, from data analysis (qualitative and quantitative) to digital dissemination. It allows researchers to structure and interpret data in an interactive way, automate parts of their analysis, and then transform their results into interactive, open, and participatory online formats.

The workshop will explore ways in which emerging AI coding assistants—e.g., OpenAI Codex, Claude Sonnet Code—and cloud-based environments—Replit, Vercel, Windsurf—can support social scientists both in performing analytic tasks and creating lightweight public research portals. Guided demonstrations and teamwork exercises take participants through the steps involved in conceptualizing a design system for the research analysis and communication, experimenting with emergent forms of AI-assisted coding templates, and discussing ethical, privacy, and sustainability considerations. A short demo—the "Mapping Afghan Diaspora Organizations" portal—will illustrate how vibe coding supports both analytical insight and interactive dissemination within one coherent system.

Understand the principle behind vibe coding and its dual role in data analysis and dissemination of research findings.
- Use AI-driven tools to analyze or structure data using natural, plain-language
interaction; no prior coding experience is necessary.
- Create a simple interactive prototype that communicates research findings, using
Replit and provided templates.
- Reflect on the ethical and practical issues regarding AI-assisted research with respect to privacy, bias, transparency, and long-term maintenance.

NECESSARY PRIOR KNOWLEDGE OF PARTICIPANTS
1. No prior programming or "coding" experience is strictly necessary. 2. Familiarity with using Large Language Models (e.g., ChatGPT, Claude) is helpful but not required. 3. A willingness to experiment with new digital tools is essential.

INFORMATION ABOUT THE INSTRUCTOR
Ali Reza Hussani is a PhD candidate at the Institute of Communication Science and Journalism at Leipzig University. His research investigates online behaviors on social media, operating at the intersection of policy, evidence, and social impact. His background is in political science and conflict management. He completed a focused data science bootcamp at Spiced Academy Berlin to enhance his methodological toolkit. He is a passionate enthusiast for AI in academia, continuously adopting and experimenting with new AI-assisted workflows to make research more efficient and interactive.

MAXIMUM NUMBER OF PARTICIPANTS
20

WILL PARTICIPANTS NEED TO BRING THEIR OWN DEVICES IN ORDER TO BE ABLE TO ACCESS THE INTERNET? WILL THEY NEED TO BRING ANYTHING ELSE TO THE WORKSHOP?
Yes. Participants must bring their own laptop.

03:00 - 03:30 PM BREAK
GESIS – LEIBNIZ-INSTITUT FÜR SOZIALWISSENSCHAFTEN KÖLN

WORKSHOP 3

03:30 - 06:00 PM
GESIS, WEST I

SESSION CHAIR: FREDERIK FUNKE

AI-TECHNOLOGIES FOR QUALITATIVE DATA ANALYSIS

DAVID RANFTLER1, PAUL WESENDONK2
1David Ranftler Xelper, Germany; 2Paul Wesendonk Xelper, Germany; ranftler@xelper.de

DURATION OF THE WORKSHOP
2h

TARGET GROUPS
Researchers and practitioners who are interested in exploring AI technologies for qualitative data analysis.
No prior programming experience is required – participants just need curiosity and a laptop.

IS THE WORKSHOP GEARED AT AN EXCLUSIVELY GERMAN OR AN INTERNATIONAL AUDIENCE?
English-speaking

WORKSHOP LANGUAGE
English

DESCRIPTION OF THE CONTENT OF THE WORKSHOP
Artificial Intelligence is rapidly transforming how qualitative research is conducted – from automating analysis to novel approaches of results presentation. In this hands-on workshop, we will explore how large language models (LLMs) and embedding technologies can be used in qualitative research.

PARTICIPANTS WILL:
- Gain an understanding of the underlying principles of LLMs and embeddings.
- Explore real-world use cases in qualitative research, such as coding interview data, clustering open-ended responses, and generating actionable insights.
- Work through guided exercises demonstrating how AI can support coding, clustering, and summarizing textual data in market research settings.

We will also discuss methodological challenges, transparency, and limitations of AI-based approaches, and share benchmarking strategies for evaluating model performance in research contexts. The workshop concludes with an open Q&A and exchange session to discuss participants' experiences, opportunities, and concerns about integrating AI in qualitative research practice.

GOALS OF THE WORKSHOP
By the end of the workshop, participants will gain knowledge of LLMs and embeddings.

NECESSARY PRIOR KNOWLEDGE OF PARTICIPANTS
No prior experience with software development or AI tools is required.

INFORMATION ABOUT THE INSTRUCTORS
Founders of xelper, a startup providing AI solutions for qualitative market research.

PARTICIPANTS SHOULD BRING THEIR OWN LAPTOPS WITH INTERNET ACCESS.

KEYNOTES

THURSDAY, 26/FEB/2026:
09:00 - 10:00 AM RH, AUDITORIUM

OPENING AND KEYNOTE 1:

PRESENTATIONS

MORE EFFICIENCY, NEW PROBLEMS? AI IN EMPIRICAL RESEARCH

DR. KATHARINA SCHÜLLER
STAT-UP Statistical Consulting & Data Science GmbH, Germany;
katharina.schueller@stat-up.com

Artificial intelligence opens up numerous new possibilities in online research. Synthetic data, automated quality assurance, intelligent recruitment strategies, and new pattern recognition tools promise unprecedented efficiency – while also raising fundamental questions: How reliable are AI-generated samples? Can algorithmic methods really reduce nonresponse? And how can we prevent models from reproducing systematic biases on a large scale?

Using concrete examples, the presentation shows how AI can improve data quality, what pitfalls lurk in automated data generation, and what skills are needed to work with it responsibly. The goal: a realistic, inspiring look at the interplay between statistics, humans, and machines – practical and future-oriented.

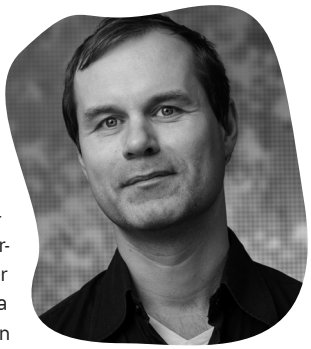
FRIDAY, 27 FEB 2026:
10:15 - 11:00 AM RH, AUDITORIUM

KEYNOTE 2:

PRESENTATIONS

WELCOME TO THE AGE OF DATA INTEGRATION. BUT HOW TO DO IT?

PROF. DR. PETER LUGTIG
Utrecht University, Netherlands, The; p.lugtig@uu.nl



Traditionally, much of the data within the social sciences and market research were purposely collected. In both qualitative research, surveys or experiments, the researcher is in control of the design of data collection. Research designs align with the goals of the study, and are targeted towards answering the most important research question(s) the researcher has.

Nowadays, more and more data are not designed, but rather 'found'. There is large variety in different kinds of found data: social media, online communication and (web) browsing data are often used in a variety of applications, but so are administrative data from governments, citizen science data, or publicly available (text) data from businesses for example. A key characteristic of found data is that the data were not collected with doing research in mind. Often, raw found data are not fit to answer a particular research question a researcher has. There is a need to process found data, (dis)aggregate them, and merge them with other datasets to be usable for research. Data integration is the process of merging or combining data from multiple sources to produce statistics.

One problem is that it is often unclear how to integrate data exactly. Integration methods are both research-question and data-source specific, making it difficult to establish a general methodology for how to integrate data. It also means that each time a statistic has to be computed, methods have to be (re)-evaluated making data integration a resource-intensive technique.

In this talk, Peter Lugtig will discuss how data-quality frameworks can be used to understand when and how to integrate data effectively. He will discuss several use cases, and show how already exisiting data quality frameworks can be used to understand how to integrate data for these specific use cases. Data quality frameworks can also be used in evaluating what datasources potentially would be useful when data integration is considered as a technique. He will also argue for the more frequent use of both designed data and found data within one study as a way to increase the quality of statistics produced by data integration.

THURSDAY, 26/FEB/2026
02:00 - 03:45 PM RH, SEMINAR 03
SESSION CHAIR: YANNICK RIEDER

GOR IMPACT AND INNOVATION AWARD

SCALING QUALITATIVE DEPTH: A LARGE-SCALE VALIDATION STUDY COMPARING AI-MODERATED INTERVIEWS AND CONVENTIONAL SURVEYS IN OTC PHARMA RESEARCH

SARAH GOTO1, THOMAS KOPF1, OLIVER TABINO2, JONATHAN HEINEMANN3, DAVID RANFTLER4, PAUL WESENDONK4
1MCM Klosterfrau Vertriebsgesellschaft mbH; 2Q Agentur für Forschung GmbH; 3horizoom-Panel – horizoom GmbH; 4xelper UG (haftungsbeschränkt)

OBJECTIVES Klosterfrau wanted to understand what consumers know about and how they evaluate different OTC pharmaceutical ingredients. Beyond this contextual question, another objective was methodological: understanding the similarities and differences between AI-moderated interviews via chat and conventional online surveys. More precise, the goal was to determine whether the two methods produce comparable results and where each method might provide unique value. Ultimately, Klosterfrau wanted to improve ingredient-level decision-making by combining quantitative results with qualitative consumer context.

METHOD & APPROACHES & INNOVATION In this project, 4 parties collaborated: Klosterfrau (industry partner), Q Agentur für Forschung (Agency), horizoom (Online Access Panel) and xelper (AI tool provider). We conducted a two-arm validation study with n=1,000 respondents in a conventional 15-minute online survey and n=1,000 respondents in a fully AI-moderated chat interview. Both arms used the same screener and identical set of ff40 OTC ingredients. The innovation of AI-moderated interviews lies in the combination of qualitative insights and quantiative reliability.

The AI-moderator probes, asks for clarifications, and encourages participants to articulate their reasoning. This conversational approach produces rich, contextual opinions at a volume typically unachievable in traditional qualitative setups. Fieldwork runs 1–2 weeks and captures both deterministic measurement and rich narratives. Importantly, this large-scale comparison study represents a methodological innovation in its own right.

It systematically benchmarks AI-moderated interviews and conventional surveys under identical experimental conditions. To the best of our knowledge, no prior research has conducted such a direct comparison – particularly at this scale and level of control.

RESULTS Full results will be available at GOR. Based on our hypotheses, we expect: - High comparability on core quantitative metrics (ingredient awareness, basic evaluations). - Higher efficiency of the conventional survey for deterministic items and structured ratings. - Strong added value from the AI-moderated interviews, generating richer narratives, decision rationales, misconceptions, and contextual associations that numbers cannot capture. The study will provide the first systematic, large-scale comparison of AI-moderated interviews versus conventional surveys.

IMPACT For Klosterfrau, the project fundamentally enhances how decisions about products and active ingredients are made. The AI-moderated interviews reveal how consumers become aware of certain ingredients, what they associate with them, where misunderstandings occur, and how these factors shape purchase decisions.

These insights directly inform ingredient communication, packaging and label claims, educational messaging, and the strategic prioritization of ingredients within the product portfolio. Beyond the immediate project outcomes, Klosterfrau and the broader research community benefit from the methodological validation itself. By empirically comparing AI-moderated interviews and conventional surveys, the study provides an evidence base for how AI-driven qualitative methods can be positioned within established research frameworks.

AGENTIC AI IN STRATEGIC DECISION MAKING

JAN MACIEJANSKI
Aequitas Group, Germany

OBJECTIVES – WHICH BUSINESS QUESTION DID THE CLIENT WANT ANSWERED?
Our client asked us to develop a strategy for the introduction of M365 Copilot. In addition to a quantitative employee survey of around 800 employees, additional employee interviews were to be conducted with the aim of determining user acceptance and preparing strategic action areas for the rollout to approximately 13,000 employees. For the generation, execution, and evaluation of these interviews, the client used the AI interview bot developed by us.

METHOD & APPROACHES & INNOVATION – HOW WERE THE INSIGHTS GATHERED?
The interview bot consists of a total of nine highly specialized AI agents, trained to conduct a real-time interview on the use of M365 Copilot. What makes this special is that the multi-agent model, like in a real conversation, can individually adapt to the mood, disposition, and

needs of the interview partner and, following the DICE methodology (Descriptive, Idiographic, Clarifying, Explanatory), can ask suitable follow-up questions.

RESULTS – WHAT ARE STRIKING AND IMPACTFUL INSIGHTS?
With our solution, we succeeded in reducing interview costs from 400 euros to 0.5 euros. In addition, all information is broken down into small knowledge units using special logics and stored in optimized knowledge databases. With these technologies, we can utilize all semantic and thematic connections and make them available in an evaluation bot. By using this AI, market researchers can analyze information in the shortest possible time and ask follow-up questions at any desired level of detail.

IMPACT – HOW DID THE PROJECT MOVE THE NEEDLE FOR THE CLIENT? WHAT WAS DONE DIFFERENTLY AFTERWARDS?
Following our project, the market research and innovation management departments became aware of us. In addition to further developing the interview bot, we are jointly planning to evolve the solution into a generic application for the generation, processing, and provision of corporate knowledge, which can be used in various areas of the company.

COMPARING AI-MODERATED, HUMAN-MODERATED, AND UNMODERATED USABILITY TESTING: INSIGHTS INTO QUALITY, USER PERCEPTION AND PRACTICAL IMPLEMENTATION

CHRISTIAN POTTIEZ1, MARTIN EINHORN1, MARKUS PANDREA2, DAVID RANFTLER3
1Porsche AG, Germany; 2Userlutions GmbH; 3xelper UG

OBJECTIVES:
This study examines three approaches to traditional, use-case-based usability testing: AI-moderated interviews, human-moderated interviews, and unmoderated testing. The aim is not only to compare the quality of insights obtained but also to analyze user perception and participant behavior during interaction with the tools. An additional goal for the research team was to gain practical experience in setting up surveys, managing workflows, and exploring analysis capabilities.

METHOD & APPROACHES & INNOVATION:
In a controlled experiment with nine participants, identical tasks were carried out under all three test conditions. Collected metrics included objectively measurable dimensions such as word count in open-ended responses, user evaluations of the methods/tools (e.g., enjoyment, willingness to participate again, usability), as well as qualitative assessment of responses regarding depth, emotional engagement, and relevance.

RESULTS:
Initial findings indicate that AI-based moderation offers significant advantages in scalability and cost but still shows usability shortcomings and limited applicability. It delivered high-quality responses and–contrary to our hypothesis–was broadly accepted, even among premium customers.

IMPACT:
The results contribute to the ongoing discussion about the maturity of AI-driven interview tools for UX research applications. Furthermore, they provide practical insights into their limitations and potential.

DEEP IMPLICT – A NEW MARKET RESEARCH FRAMEWORK – VALIDATED AT DFBS TOUGHEST CHALLENGE

FRANK BUCKLER
SUPRA, Germany

OBJECTIVES
WHICH BUSINESS QUESTION DID THE CLIENT WANT TO ANSWER?
The German Women's National Soccer Team (GWNST) needed reliable, budget-friendly insights to boost fan engagement and guide their marketing strategy leading up to UEFA Women's EURO 2025. Their core question was:

How can we increase the popularity, emotional connection, and reach of women's soccer using insights that go beyond traditional and AI-based research?

METHOD & APPROACHES & INNOVATION
HOW WERE THE INSIGHTS GATHERED?
The team pioneered a completely new methodology called Deep Implicit, designed to capture pure human intuition – something neither traditional research nor AI can reliably access.

The innovative approach included:
1. Guided audio-based alpha-state induction using techniques used in hypnosis and meditation and binaural beats to calm the rational mind and activate intuitive processing.
2. Projective questions asked during a relaxed state to elicit first, unfiltered System-1 thoughts.
3. Strict 30-second capture window to prevent rationalisation.
4. LLM-assisted multi-stage analysis:
Filter for genuine System-1 responses
Summarize intuitive statements

Interpret deeper psychological meaning
1.Quantitative validation via a 500-respondent causal driver analysis using Supra Causal AI to test whether intuitive findings causally explain real-world fan excitement.

This is the first-ever methodology to measure intuitive qualitative insight in a standardized and automatized form, filling a major gap in both AI and market research.

RESULTS
WHAT ARE STRIKING AND IMPACTFUL INSIGHTS?
The Deep Implicit process uncovered a new strategic truth that traditional research and AI had missed:
1. Fans become supporters when they truly identify with the players as relatable humans – not as polished athletes.
2. They also need to identify with the team's values – the club must "stand for something I support."

3. Authenticity and emotional storytelling – not scripted content – are the most powerful drivers of fan excitement.
4. Quantitative validation showed:
The new attribute "the club stands for something I support" was the #1 causal driver of fan excitement – surpassing all expert-designed items.
Authenticity ranked among the top three drivers.
These insights fundamentally reframed how women's soccer should communicate its identity.

Key takeaways:
- Without Deep Implicit, the Causal AI study would have missed to by far largest driver
- Without Deep Implicit, supporting LLMs would have come up with shallow inputs

The true power is in combining Human and Artificial Intelligence.
- Human intelligence is not only exert and domain knowledge, but the underleveraged intelligence of human is also called INTUITION
- Artificial intelligence is not just generative AI but causal AI, to explore ground truth in data.

IMPACT

HOW DID THE PROJECT MOVE THE NEEDLE FOR THE CLIENT? WHAT CHANGED AFTERWARDS?

The insights were implemented immediately in the GWNST's social media and communications strategy – shifting from scripted content to authentic, emotionally resonant, player-centered storytelling.

The effect was dramatic:
Instagram Performance (May / June)
After implementing the new strategy:
- Reach increased by +120% on average
Album posts: 296,878 / 542,928
Reels: 297,896 / 822,160
- Interactions nearly doubled
Albums: 9,908 / 19,206
Reels: 10,799 / 20,104

This represents a step-change in communication efficiency, achieved without additional budget – purely through better insight.
Beyond performance metrics, the project also:
- Provided the DFB with the first validated intuition-based insight tool
- Revealed a new strategic narrative for women's soccer
- Demonstrated how AI and research can be complemented – not replaced – by intuitive intelligence

FROM INSIGHTS TO IMPACT: A LIFE-CENTRIC, AI-DRIVEN APPROACH TO MODERN BRAND TRACKING

JÖRG MUNKES
GIM Gesellschaft für innovative Marktforschung mbH, Germany

Modern market research increasingly faces the requirement to deliver not just insights but actionable, impact-oriented strategies. Our contribution addresses the central question of how brand tracking can be advanced to explain when and why consumers choose brands, how these decisions are embedded in both situational and general life contexts, and how such understanding can be translated into growth-oriented actions.

Methodologically, the approach builds on a life-centric brand tracking framework that captures brand usage within contextualized real-life situations and is guided by the concept of Category Entry Points (CEP) as defined by Byron Sharp. By applying a reversed questioning logic – asking "which brand fulfills which need in which situation?" rather than capturing image associations – we can model situational decision-making with greater precision. In addition, real user profiles are enriched using the AI system Allon, which generates contextual information from consumers' living environments and creates synthetic consumer profiles. Combined with social and behavioral science expertise, this results in a data-driven, context-sensitive model of brand choice.

The results show that situational needs are key drivers of brand decisions and that a life-centric approach significantly sharpens the understanding of these decision dynamics. AI-generated profiles deepen the contextual knowledge of consumers' life situations and enable new activation points along the customer journey. The integration of tracking data, synthetic profiles, and AI-based context expansion increases the precision of target group modelling and makes consumers more effectively addressable.

For companies, this provides clear added value: instead of descriptive insights, they receive concrete communication and action strategies that are directly linked to situational consumer needs. Looking ahead, this approach positions brand tracking as a systematic, AI-supported steering instrument that identifies growth opportunities and aligns marketing activities with consumers' real-life contexts.

JURY MEMBERS OF 2026

DR. EVA AIZPURUA
Lloyds Banking Group and
University of Northern Iowa

PROF. DR. NICOLA DOERING
Ilmenau University of
Technology, Germany

PROF. DR. FLORIAN KEUSCH
University of Mannheim, Germany,
Jury Chair

DR. HENNING SILBER
University of Michigan, USA

PROF. DR. CHRISTOPH KERN
Ludwig-Maximilians-University
of Munich, Germany

CHAN ZHANG, PHD
Zhejiang University, China

AWARD CEREMONY:
FRIDAY 27.02.26, 11:00 - 11:45 AM

SESSION CHAIR: FLORIAN KEUSCH
LOCATION: RH, AUDITORIUM

The German Society for Online Research (DGOF) annually recognizes outstanding scientific contributions in online research through the DGOF Best Paper Award for a researcher or group of researchers.

The prize is awarded to a paper that provides a fundamental scientific contribution to the advancement of the methods of online research. Both theoretical/conceptual and empirical/methodological papers are considered for the award.

The award is worth 500 Euro and will be presented at the annual GOR conference. An abstract (and, if available, a preprint) of the award-winning paper will be posted to the DGOF website (www.dgof.de).

To be considered for the award, papers must have been published in an outlet that uses a peer-review process (e.g., peer-reviewed journal, full papers in peer-reviewed conference proceedings, refereed book chapter) at the time of submission. Papers written in German or English and published not earlier than 2023 (if the paper was published online-first, then the online-first publication date counts) were eligible to be submitted for the DGOF Best Paper Award 2025.

WINNER:

SIMSON, J., ET AL. (2025).

PREVENTING HARMFUL DATA PRACTICES BY USING PARTICIPATORY INPUT TO NAVIGATE THE MACHINE LEARNING MULTIVERSE. PROCEEDINGS OF THE 2025 CHI CONFERENCE ON HUMAN FACTORS IN COMPUTING SYSTEMS, 806, 1-30.

https://doi.org/10.1145/3706598.3713482

THURSDAY, 26/FEB/2026
01:30 - 02:30 PM RH, AUDITORIUM

4.1: POSTER SESSION

SESSION CHAIR: TOBIAS RETTIG

PRESENTATIONS

BEYOND THE FIRST QUESTIONNAIRE: RETAINING PARTICIPANTS IN AN APP-BASED HOUSEHOLD BUDGET SURVEY

MAREN FRITZ, FLORIAN KEUSCH
University of Mannheim, Germany; maren.fritz@uni-mannheim.de

RELEVANCE & RESEARCH QUESTION

Survey attrition is a common problem in household budget surveys (HBS) as such surveys impose a high burden on participants, asking them to report their expenses daily for a specific period. We study the attrition in an app-based HBS with three preceding questionnaires and a 14-day diary. Participants can drop out at each of the questionnaires or during the diary. If respondents drop out, the data obtained from them contains missing information. The research questions are: 1. At what stage do participants drop out of an app-based HBS? 2. Does the way the study is presented in the invitation letter influence the drop-out? 3. What individual characteristics correlate with drop-out at the different stages?

METHODS & DATA

In 2024, we drew a probability sample of 7049 individuals from the register of residents and invited them by mail. We included an experimental variation in the invitation letters to analyse whether stressing the effort associated with participation and whether mentioning a receipt scanning function has an influence on participation.

The survey consisted of three initial questionnaires regarding personal information, income and expenses, and information about the household. After having completed them, respondents were asked to continue with an expense diary. To be eligible for an incentive, participants had to manually enter their expenses or scan their receipts on at least seven days.

RESULTS

Participants drop out of the study continuously, even after they started data entry in the diary. The poster will present results on the attrition and drop out across the stages, and whether the three experimental groups differ significantly in their attrition rates across the stages. Additionally, the poster shows whether people with specific characteristics are more likely to drop out at a particular stage.

ADDED VALUE

This research is part of the Smart Survey Implementation (SSI) project, funded by EUROSTAT, which aims to enhance data collection for official statistics across Europe through digital innovation. This experiment addresses attrition in app-based HBS. It informs decisions about what actions can be taken to reduce attrition.

JOINT EVALUATION OF LLM AND HUMAN ANNOTATIONS WITH MULTITRAIT–MULTIERROR MODELS

GEORG AHNERT1, MAXIMILIAN KREUTNER1, ALEXANDRU CERNAT2, MARKUS STROHMAIER1,3,4
1University of Mannheim; 2University of Manchester; 3GESIS–Leibniz Institute for the Social Sciences; 4CSH Vienna; georg.ahnert@uni-mannheim.de

RELEVANCE & RESEARCH QUESTION

Large Language Models (LLMs) are increasingly used to for text annotation, and for simulating survey responses via silicon samples. To evaluate their validity, researchers commonly assess the alignment of generated annotations and survey responses with data from human participants. However, treating human data as ground-truth overlooks the various sources of measurement error inherent in human data collection.

We propose applying MultiTrait–MultiError (MTME) models to jointly analyze imperfect responses from LLMs and humans. Our goal is to assess the extent to which MTME models can qualify measurement quality in responses from LLMs and humans, and whether they can improve estimates of underlying traits.

METHODS & DATA

Our MTME models estimate latent trait and error factors based on multiple measures from humans and LLMs of the same trait, e.g., the readability of a text excerpt. LLM measures may vary by model size, model family, prompting strategy, or decoding approach.

In an initial study, we fit an MTME model on LLM-generated annotations of publication year and readability for 600 text excerpts from the CommonLit Ease of Readability Corpus (CLEAR) corpus. We evaluate four Qwen 3 LLMs (4B–32B). We include ground-truth data from the CLEAR corpus into our evaluation, purely for demonstrating the feasibility of the MTME approach.

RESULTS

Our initial results show that the measurement quality estimated by the MTME model partially matches correlations of individual LLMs with the CLEAR data. We find that the factor scores extracted from the MTME

model for readability have a higher correlation with the CLEAR data (0.76) than the responses of each individual LLM (≤ 0.70), indicating that MTME aggregation yields a more accurate estimate of the underlying trait. Next, we will incorporate human-coded annotations into the MTME model and conduct experiments involving other datasets, categorical variables, and LLM-generated survey responses.

ADDED VALUE

We demonstrate that MTME models can jointly estimate measurement quality in responses from humans and LLMs. By extracting latent traits that pool information across participants and models, we show that MTME-generated predictions can improve the estimation of underlying traits. MTME models are a promising method for the joint analysis of LLM and human data.

LESSONS LEARNED: UTILIZING SOCIAL MEDIA INFLUENCERS FOR TARGETED RECRUITMENT ON DISCRIMINATION IN THE GERMAN HEALTHCARE SYSTEM

ZAZA ZINDEL1,2, AYLIN MENGI1, ZERRIN SALIKUTLUK1,3, TAE KIM1
1German Centre for Integration and Migration Research (DeZIM), Germany; 2Bielefeld University, Germany; 3Humboldt-University, Germany; zaza.zindel@uni-bielefeld.de

RELEVANCE & RESEARCH QUESTION

Traditional sampling strategies often fail to recruit sufficient cases from small or marginalized populations, especially for sensitive topics where distrust or fear of repercussions are common. While social media recruitment is increasingly discussed in survey methodology, most work centers on targeted advertising rather than leveraging the "social" infrastructure of platforms: trusted creators embedded in tightly segmented communities. Influencers can function as credible intermediaries and, methodologically, as seeds in a semi-network (virtual snowball) recruitment process. This poster reports lessons learned from a feasibility study leveraging a network of "Medfluencers" and anti-racism advocates to recruit participants for research on discrimination in the German healthcare system. We ask: Is engaging social media influencers worthwhile for survey recruitment, and what methodological insights emerge regarding reach, conversion, and sample composition?

METHODS & DATA

We fielded an online survey on experiences of discrimination in the German healthcare system among patients and medical professionals. The questionnaire combined closed items with multiple open-ended text prompts to elicit detailed accounts. Recruitment centered on one primary seed influencer: a practicing medical doctor who is also a visible Muslim woman. She shared the survey link and mobilized her network of fellow influencers to repost, creating a virtual snowball mechanism through interconnected audiences. Data collection ran from February 1-25, 2025.

RESULTS

The seed post received 14,952 likes and 437 comments. The survey link was accessed 42,530 times; 27,710 individuals consented; 17,390 completed the questionnaire (62% post-consent completion). The campaign produced a strong multiplier effect over a short field period and reached key target groups: 3,505 Muslim participants and 2,979 respondents working in the German medical system. The sample skewed female (n = 16,803) and younger (mean age 35.2) than the average German population. High perceived source credibility appeared to facilitate trust and candid answering; 2,387 respondents volunteered for follow-up interviews.

ADDED VALUE

Although not suited for population inference, influencer sampling can efficiently generate large, highly motivated samples in under-researched groups. The poster concludes with concrete lessons learned on when and how influencer-seeded recruitment can be productively used in survey research, and which limitations researchers must communicate transparently when interpreting resulting data.

CLASSIFYING MORAL REASONING IN POLITICAL DISCOURSE: DEMONSTRATING INTERRATER RELIABILITY AND TESTING AN AI-BASED CLASSIFICATION APPROACH

FELIX SCHMIRLER, RUDOLF KERSCHREITER
Freie Universität Berlin, Germany; felix.schmirler@fu-berlin.de

RELEVANCE & RESEARCH QUESTION

Moral reasoning, whether people justify political positions through rules and duties (deontological reasoning) or through expected outcomes (consequentialist reasoning), is central to understanding how people deliberate in polarised political (online) debates. While experimental moral-dilemma research shows differences in rule-based vs outcome-based judgments depending on political ideology, it remains unclear whether such patterns manifest in real-world political communication. Capturing moral reasoning in naturalistic discourse could deepen our understanding of how polarization emerges and provide a foundation for designing interventions that adapt to people's reasoning styles. This study asks: Can moral reasoning in political discourse be reliably classified using NLP methods, and do we find evidence supporting fundings from experimental moral dilemma research?

METHODS & DATA

This study presents a validation of an approach to classify moral reasoning in political discourse using a large language model (LLM). A corpus of 576 sentences from Reddit discussions and German parliamentary speeches was pre-sampled using a novel extension of the Distributed Dictionary Representations (DDR) method (Garten et al., 2018), which identifies sentences with high cosine similarity to exemplary reasoning styles. Two expert raters then independently coded each sentence as deontological, consequentialist, or neutral and adapted the coding manual based on deliberation after each round.

RESULTS

Interrater reliability improved across three codebook iterations to excellent reliability (Krippendorff's = .56-.68 .92-.93). Agreement between human and AI-assigned labels based on cosine similarity were subsequently also sufficient to demonstrate the feasibility of classifying moral reasoning styles in large text corpora through LLMs (Krippendorff's = .70-.73). Building on this validation, this method

willbe applied to a larger multilingual corpus (ff12k sentences) to analyze ideological and temporal patterns in moral reasoning across political and cultural contexts.

ADDED VALUE

This work offers the first validated procedure for detecting deontological and consequentialist reasoning in naturalistic political communication, bridging experimental moral psychology and large-scale text analysis. It establishes a foundation for ongoing research using a multilingual, cross-country corpus to study ideological and temporal patterns in moral reasoning, with potential downstream applications for designing tailored, depolarizing communication interventions.

DYNAMIC SURVEYS FOR DYNAMIC LIFE COURSES: DEVELOPMENT OF A WEB-APP FOR SELF-ADMINISTERED LIFE HISTORY DATA COLLECTION

SEBASTIAN LANG1, HEIKE SPANGENBERG2, DAVID OHLENDORF2, HEIKO QUAST2, LEENA LAHSE2

1Leibniz Institute for Educational Trajectories (LIfBi), Germany; 2German Centre for Higher Education Research and Science Studies (DZHW), Germany; sebastian.lang@lifbi.de

RELEVANCE & RESEARCH QUESTION

Life history data is essential in social sciences, yet retrospective collection often suffers from memory errors, reducing data quality. The Life History Calendar (LHC) addresses some of these issues, but its traditional interviewer-based administration in surveys is costly and participation rates are declining. Self-administered web surveys (CAWI) offer a promising alternative. Our research asks: Can a dynamic, self-administered web application with an integrated LHC and continuous access improve data quality and reduce response burden compared to a classic retrospective collection?

METHODS & DATA

To explore this, we developed a web app that enables respondents to update their life courses continuously, aiming to enhance usability, minimize recall errors, and simplify panel maintenance. We designed two experiments where respondents are randomly assigned to treatment groups (continuous life history data collection with or without reminders) and a control group (classic retrospective collection). Preliminary results come from the first experiment, implemented as NEPS next add-on study using starting cohort three. We analyze response rates and the effect of the new LHC implementation on response burden using a fixed-effects panel regression model.

RESULTS

First results show no differences in initial participation across groups. Similarly, response rates (RR1) for completed interviews do not differ between the existing LHC and the new web-app implementation. Regarding response burden, we observe an increase immediately after the LHC section, but this increase does not differ across experimental groups.

ADDED VALUE

Our approach introduces a flexible, cost-efficient solution for self-administered life history data collection. The added flexibility in reporting life courses did not negatively affect participation or burden, which is a key prerequisite for the main experiment on continuous access. This innovation enables continuous updates, simplifies panel management, and aims to counteract attrition, offering a scalable model for improving longitudinal data quality in self-administered life history collection.

THURSDAY, 26/FEB/2026
01:30 - 02:30 PM RH, AUDITORIUM

4.2: POSTER SESSION

SESSION CHAIR: TOBIAS RETTIG

PRESENTATIONS

BEYOND ALGORITHMS: HOW TO IMPROVE MANUAL CLASSIFICATION OF VISUAL DATA OBTAINED IN SURVEYS

MARIA PAULA ACUÑA-PARDO, MELANIE REVILLA, LEYRE PADILLA LÓPEZ
RECSM- Universitat Pompeu Fabra, Spain; mariapaula.acuna@upf.edu

RELEVANCE & RESEARCH QUESTION

An increasing number of studies are requesting visual data within web surveys, arguing that they can enhance data quantity and quality and provide new insights. However, important challenges remain. This study focuses on one of these aspects: extracting relevant information from the visual data, a process called "classification".

Researchers are increasingly relying on automated methods using machine learning to classify visual data. Although these methods are becoming more powerful, they still cannot extract information in the same way as manual classification. Thus, the main objective of this study is to explain the challenges and solutions encountered while implementing manual classification in a complex case study on remote work homestations, where approximately 70 items must be classified based on three photos.

METHODS & DATA

A web survey will be conducted in the opt-in online panel Netquest in Spain (N = 1,200) in early December among remote workers. After answering conventional questions about their remote work conditions, respondents will be asked to upload two photos of their homestation and one of their main screen or laptop model information. All photos will be reviewed by the project ethics advisor to ensure no private information is visible. Manual classification will be implemented in accordance with detailed guidelines.

The two homestation photos will be classified jointly, and the device-model photo will be classified separately. Two researchers will share homestation classification, with approximately 10% of the photos coded by both of them to compute interrater reliability (IRR) indicators and identify potential systematic biases. One researcher will code the device-model photo, but a subsample will undergo double coding to assess IRR.

RESULTS

We expect to find differences between classifiers, identify problematic items, and detect the types of errors most likely to occur across classifiers.

ADDED VALUE

This work in progress focuses on discussing the challenges encountered when dealing with the classification of complex visual data collected in web surveys. We aim to provide practical, user-oriented guidelines that extend beyond the explanations usually found in academic papers, which often prioritize presenting results over detailing the underlying classification process.

AI FOR SURVEY DESIGN: GENERATING AND EVALUATING SURVEY QUESTIONS WITH LARGE LANGUAGE MODELS

ANNA FUCHS1, ANNA-CAROLINA HAENSCH1,2,3, WIEBKE WEBER1
1LMU Munich; 2Munich Center for Machine Learning; 3University of Maryland, College Park; anna.fuchs@stat.uni-muenchen.de

RELEVANCE & RESEARCH QUESTION:

Designing high-quality survey questions is a complex task. With the rapid development of large language models (LLMs), new possibilities have emerged for supporting this process through the automated generation of survey items. Despite growing interest in LLM tools within industry, published research in this area remains sparse, and little is known about the quality and characteristics of survey items generated by LLMs or the factors influencing their performance. This work provides the first in-depth analysis of LLM-based survey item generation and systematically evaluates how different design choices affect item quality.

METHODS & DATA:

Five LLMs, namely GPT-4o, GPT-4o-mini, GPT-oss-20B, LLaMA 3.1 8B, and LLaMA 3.1 70B, were used to generate survey items on four substantive domains: work, living conditions, national politics, and recent politics. We additionally evaluate three prompting strategies: zero-shot, role, and chain-of-thought prompting. To assess the quality of the generated survey items, we use the Survey Quality Predictor (SQP), a tool developed by survey methodologists for estimating the quality of attitudinal survey items based on codings of their formal and linguistic characteristics. To code these characteristics, we used an LLM-assisted procedure. The analysis allows us to evaluate not only overall quality but also around 60 specific survey item characteristics, offering a detailed view of how LLM-generated questions differ.

RESULTS:

The findings show striking differences in survey item characteristics across the different models and prompting techniques. Both the choice of model and the prompting technique employed influence the quality

of LLM-generated survey items. Closed-source GPT models generally produce more consistent items than open-source LLaMA models. The topics 'work' and 'national politics' yield survey items with the highest quality. Overall, chain-of-thought prompting achieved the best results. GPT-4o, GPT-4o-mini, and LLaMA 3.1 70B achieved similar item quality, although the LLaMA model showed greater variability.

ADDED VALUE:
For the GOR community, the study offers empirical evidence on how LLMs can (and cannot) be reliably integrated into questionnaire design workflows, providing a systematic basis for evaluating emerging AI tools in survey research and informing methodological decisions in applied settings.

WHO MOVES AND WHO DO WE LOSE? MOBILITY-SPECIFIC ATTRITION IN PANEL SURVEYS.

MARKUS SCHMADERER, TOBIAS GUMMER
GESIS, Germany; markus.schmaderer@gesis.org

RELEVANCE & RESEARCH QUESTION
Residential mobility is an important source of attrition in address-based panel surveys. That is, panelists cannot be invited to a survey because their new address is unknown. Current research into mobility specific attrition (MSA) is lacking in three aspects: (1) Because of a lack of meta-reviews and ambiguous definitions for MSA, there is limited insight in the magnitude of MSA. (2) research into the selectivity of MSA and its' contribution to attrition bias is sparse and almost exclusively based on panels of special populations, and (3) almost no research on MSA for self-administered panel surveys exists, which continue to displace traditional face-to-face-interviewing.

Consequently, in this study, we address two research questions: 1) How many respondents in panel surveys are mobile and how many of those attrite due to MSA. 2) Are certain subpopulations more prone to MSA than others?

METHODS & DATA
To answer RQ1 we conduct a meta-review by (1) systematically sampling panel surveys from three extensive data archives (ICPSR, CESSDA and GESIS data-archive). and then (2) analyzing their field documentation to gain a deep understanding of the prevalence of mobility and MSA in panel surveys with special regards to self-administered surveys.

For RQ2 we explore subpopulations at risk of MSA by employing machine learning algorithms (classification trees) on data from all available waves of FReDA (currently three waves; N = 42,787). FReDA is a probability-based self-administered mixed-mode panel study with biannual surveys using both web-based and paper-based questionnaires. FReDA's primary mode of contact is postal mail. Its sample base are German residents aged 18 to 49 years which were recruited by drawing a sample of 108,256 individuals from population registers of German municipalities (Bujard et al., 2025).

RESULTS
Analyses are planned for December 2025 and January 2026. Preliminary results will thus be available in February 2026.

ADDED VALUE
We identify the magnitude of MSA in panel surveys and whether it introduces systematic biases, thereby quantifying a potential source of error for panel researchers.

"MY (22M) GIRLFRIEND (23F) COMES HOME AND DOES NOTHING" – GENDERED PERCEPTIONS OF PAID AND HOUSEHOLD LABOR IN REDDIT RELATIONSHIP DISCUSSIONS OVER TIME

BIRGIT ZEYER-GLIOZZO1, JOHANNA HÖLZL2, GUNDULA ZOCH3,4, PHILIPP DOEBLER1
1TU Dortmund University, Germany; 2University of Mannheim, Germany; 3Carl von Ossietzky University Oldenburg, Germany; 4Leibniz Institute for Educational Trajectories (LIfBi), Germany; johanna.hoelzl@uni-mannheim.de

RELEVANCE & RESEARCH QUESTION
The COVID-19 pandemic has reignited longstanding questions about gender inequalities in paid and unpaid labor. While survey research has advanced our understanding of these disparities, it typically relies on predefined categories and is susceptible to social desirability bias, especially for sensitive topics. In contrast, online postings capture intimate relationship conflicts in great depth, however rarely include demographic information.

We leverage discussions in Reddit relationship communities that, due to unique community roles, include both rich descriptions of relationship conflicts around (un)paid labor and demographic details (age, gender). We first assess how well Large Language Models (LLMs) classify manifest and latent content in Reddit posts. Building on the best-performing approach, we examine how men and women discuss relationship conflicts around (un)paid labor before, during, and after the pandemic.

METHODS & DATA
Using GPT-family LLMs on 500,000 posts from the subreddits r/relationships and r/relationship_advice, we extract manifest demographic attributes and classify whether posts discuss romantic relationships, paid, and unpaid labor. We systematically vary model specifications (o3, 4o, 4.1), prompting strategies (zero-shot vs. few-shot), fine-tuned vs. base models, and context window lengths, evaluating each against human annotations. Using classifications from the best-performing approach, we apply Structural Topic Models to explore how men and women discuss (un)paid labor in romantic relationships, and how these discussions evolve over time (2011-2023).

RESULTS
Across specifications, LLMs excel at predicting manifest categories but struggle with latent sociological constructs. Even the best-performing approach of fine-tuned GPT-4.1 with few-shot prompting and detailed category descriptions achieves only moderate performance when classifying paid and unpaid labor. Substantively, preliminary findings indicate that women more often discuss mental health in work-related conflicts, while men more frequently emphasize career objectives.

ADDED VALUE

We apply LLM-based classifications to core sociological questions around gender inequalities. We add to the growing research body on LLMs' capabilities and limitations in classifying complex social-science constructs and offer new evidence on how gender disparities in paid and unpaid labor are reflected and negotiated in relationship conflicts. By combining demographic information with highly sensitive narratives, our dataset provides an empirical resource rarely available in either survey or social-media research.

A CHANGING LANGUAGE OF SUSTAINABILITY? GLOBAL ONLINE DISCOURSE ANALYSIS WITH A DEEP-DIVE ON GERMANY

SEBASTIAN TIEKE1, JAN DIRK KEMMING2

1Weber Shandwick, Germany; 2Fresenius University Koeln, Media School; stieke@webershandwick.com

BACKGROUND

The sustainability discourse has become one of the most dynamic and contested global public conversations. Terms such as ESG, Net Zero, Sustainability, Climate Targets and Decarbonisation are widely used but differ substantially in meaning and emotional connotation across regions (Chernyshova et al., 2025).

Prior research (Müller-Hansen et al., 2023; Mi & Zhan, 2023; Gupta et al., 2024) shows strong emotional patterning and polarisation in climate debates. Topic-modeling and semantic-network studies (Jaiswal et al., 2024; Lee et al., 2023) demonstrate how sustainability narratives can be algorithmically mapped, though usually without global comparison or predictive framing.

PURPOSE

This study provides a data-driven analysis to quantify semantic differences and assess whether discourse dynamics can serve as early indicators of social and policy developments. We examine regional variation in meaning and sentiment, the stabilising or polarising roles of key terms, how the German discourse diverges from global patterns, and whether semantic shifts produce predictive signals.

METHOD

Using Talkwalker, we analysed over 50 million posts from news outlets, social media and public platforms in more than 40 languages (Oct 2024–Oct 2025). Building on sentiment and ESG discourse methodologies, we extend existing approaches with a gravity model combining favourability and momentum. Methods include volume and engagement metrics, sentiment analysis, semantic clustering, trend identification, regional comparison and early-signal extraction, drawing conceptually on opinion-dynamics and DETM-based temporal modelling.

RESULTS

Globally, Sustainability acts as a stable anchor, Net Zero shows declining credibility and volatility, and Decarbonisation remains the operational core. Regional narratives diverge: North America polarises, Europe institutionalises, Asia operationalises, Africa and South America emphasise fairness and identity. In Germany, 63% of discourse centres on climate neutrality and emissions reduction. Engagement peaks around conflictive terms (Net Zero, Climate Targets), while regulatory topics (ISSB, CSRD, CSDDD) dominate momentum. Predictive indicators include early sentiment decline for Net Zero and rising relevance of circularity, life-cycle and nature-positive concepts.

ADDED VALUE

The study demonstrates how sustainability operates as a semantic system, how regional meanings diverge despite shared terminology, and how discourse data can signal emerging regulatory, communicative and societal developments.

References on demand.

THURSDAY, 26/FEB/2026
01:30 - 02:30 PM RH, AUDITORIUM

4.4: POSTER SESSION

SESSION CHAIR: TOBIAS RETTIG

PRESENTATIONS

TECHNOSTRESS AND BURNOUT IN DAILY ACADEMIC LIFE: AN EMPIRICAL INVESTIGATION OF STUDY-RELATED STRESSORS WITHIN THE STUDY DEMANDS AND RESOURCES MODEL

ANNIKA PUHL, IVONNE PREUSSER
Technische Hochschule Köln, Germany; annika.puhl@smail.th-koeln.de
Understanding Technostress in Higher Education: Insights from an Online Survey Using the Study Demands and Resources Model

RELEVANCE AND RESEARCH QUESTION
Digitalization in higher education through learning platforms, collaboration tools, and AI applications creates new stressors that manifest as technostress and can increase the risk of burnout. Critical discrepancies arise between students and technologies (P-TEL), their social environment (PP), and study organization (PO). This study examines how multidimensional technostress affects burnout symptoms, whether general perceived stress mediates this relationship, and whether study-related self-efficacy serves as a personal resource.

METHODS & DATA
The study is based on a quantitative online survey of students from German universities (N ≈ 210). Measures included technostress (Technostress-Misfit Scale with P-TEL, PP, and PO), general stress (PSQ), burnout (MBI-SS short version), and study-related self-efficacy (BSW-5-Rev). Analyses comprised reliability and factor analyses, multiple regressions, and moderated mediation analyses (PROCESS models 14 and 7), controlling for sociodemographic factors.

RESULTS
Technostress correlates strongly and positively with burnout symptoms. The misfit between students and institutional structures, platforms, and deadlines (PO misfit) is the strongest trigger for burnout. General stress partially mediates this relationship, while study-related self-efficacy represents a strong protective factor but does not significantly moderate the effects. Exploratory findings indicate higher burdens among women, bachelor's students, and individuals with lower technology skills.

ADDED VALUE
The study integrates technostress as a specific demand dimension into the Job Demands-Resources model and identifies organizational misalignments as central intervention points. It provides practical recommendations for universities: uniform digital platforms, transparent structures, and targeted promotion of self-efficacy.

Lesener, T., Pleiss, L. S., Gusy, B., & Wolter, C. (2020). The Study Demands-Resources Framework: An Empirical Introduction. International journal of environmental research and public health, 17(14), 5183. https://doi.org/10.3390/ijerph17145183

Schaufeli, W. B., Martinez, I. M., Pinto, A. M., Salanova, M., & Bakker, A. B. (2002). Burnout and engagement in university students: a cross-national study. Journal of Cross-Cultural Psychology, 33(5), 464–481. https://doi.org/10.1177/0022022102033005003

ESTIMATING ECONOMIC PREFERENCES FROM SEARCH QUERIES

MAXIMILIAN ALTHAUS, KEVIN BAUER, BERND SKIERA
Goethe University, Germany; m.althaus@econ.uni-frankfurt.de

RELEVANCE & RESEARCH QUESTION
Economic preferences are central to understanding consumer decision making, yet established measurement approaches remain costly and difficult to scale. At the same time, consumers generate rich digital traces in their day-to-day online behavior. This project asks whether modern language models can infer core economic preferences directly from search query histories.

METHODS & DATA
We recruit approximately 800 participants through the Datapods Platform. Each participant links a Google account and shares up to twenty years of pseudonymized search queries. They then complete incentivized tasks and matched survey items that elicit six preference parameters: risk preferences, time discounting, altruism, trust, positive reciprocity, and negative reciprocity. Large Language Models receive each participant's query history and generate preference predictions, which are evaluated against experimentally measured benchmarks.

RESULTS
The study is still ongoing. Initial trials have confirmed the functionality of the idea.

ADDED VALUE
The project offers the first large-scale test of preference prediction from search histories using incentivized ground truth. It informs the feasibility and limits of low-cost, behavior-based preference measurement, contributes to ongoing debates on digital profiling in markets, and highlights privacy-relevant implications of model-based inference from everyday online behavior.

TRAIT OR STATE? UNDERSTANDING MOTIVATIONAL DRIVERS OF STRAIGHTLINING IN A LONGITUDINAL PANEL SURVEY

ÇAGLA E. YILDIZ
GESIS - Leibniz Institute for the Social Sciences, Germany; cagla.yildiz@gesis.org

RELEVANCE & RESEARCH QUESTION

Straightlining—providing (nearly) identical responses in multi-item batteries—is a common indicator of satisficing in survey research. Compared with task difficulty and respondent ability, respondent motivation has received less systematic attention as a driver of satisficing. In longitudinal surveys, an important open question is whether motivational constructs reflect stable, trait-like characteristics or situational, state-like fluctuations across waves.

Survey research also employs diverse operationalizations of motivation (e.g., topic interest, personality traits, survey attitudes), yet these are rarely compared in terms of temporal stability or predictive power. This study therefore examines the extent to which motivational measures display trait- versus state-like variation and how these components relate to straightlining over time.

METHODS & DATA

We use data from the GESIS Panel.pop, a probability-based mixed-mode panel in Germany, drawing on nine annual waves of the Social and Political Participation Longitudinal Core Study. Repeated measures are available for political interest, Big Five traits (agreeableness, conscientiousness), survey attitudes, and straightlining. To assess stability, we estimated separate random-intercept models for each motivational indicator and computed intraclass correlation coefficients (ICCs). To predict straightlining, we applied a within–between decomposition and estimated a multilevel binomial logistic regression with respondent random intercepts, controlling for demographics, cohort, mode, and survey year.

RESULTS

Motivational indicators show considerable temporal stability. Political interest is the most trait-like (ICC = 0.76), followed by conscientiousness (0.64) and agreeableness (0.58). Survey attitudes display more moderate stability, with ICC values ranging from 0.53 (perceived burden) to 0.59 (survey value) and 0.62 (enjoyment). In the predictive model, political interest is the strongest determinant of straightlining: both higher average levels and within-person increases reduce straightlining. Survey attitudes show smaller, largely trait-level associations, and personality traits have modest effects.

ADDED VALUE

Findings indicate that straightlining is driven mainly by stable, between-person motivational differences, with political interest standing out as the strongest factor. By comparing multiple motivation measures and separating their trait and state components, the study provides practical insights for identifying respondents at risk of satisficing and for supporting data quality in longitudinal surveys. Extensions to additional satisficing indicators (e.g., item nonresponse, speeding) are planned.

COMPARING PROBABILITY AND NONPROBABILITY ONLINE SURVEYS: DATA QUALITY AND FIELDWORK PROCESSES

EMMA FÖSSING1,2, LUKAS OLBRICH1, STEFAN ZINS1, JÖRG DRECHSLER1,2,3
1Institute for Employment Research, Germany; 2Ludwig-Maximilians-Universität, Munich, Germany; 3University of Maryland, College Park, USA; emma.foessing@iab.de

OBJECTIVES

Which business question wanted the client to be answered?
Online surveys based on nonprobability samples have become increasingly popular due to their cost efficiency and rapid fieldwork. However, nonprobability samples have a poor reputation regarding their data quality compared to probability-based samples. This study investigates the extent to which probability and nonprobability samples differ with regard to data quality.

METHOD & APPROACHES & INNOVATION

How were the insights gathered?
We conduct one self-administered online probability recruited via postal invitation letters and four nonprobability online surveys administered by commercial online-access panel providers. All surveys use an identical questionnaire. We compare the resulting datasets in terms of key statistical indicators of data quality such as the survey duration, passing or failing of an attention check and specific screen times for longer item texts. In addition, we compare operational aspects of the fieldwork process between probability and nonprobability samples.

RESULTS

What are striking and impactful insights?
Preliminary findings indicate not only substantial differences between the probability and nonprobability datasets, but also among the nonprobability panels themselves. These differences highlight the methodological and practical challenges of relying on nonprobability data for inferential research.

IMPACT

How did the project move the needle for the client? What was done differently afterwards?
The study adds value by providing a systematic, multi-panel comparison of online survey methods covering a broad target population based on IAB administrative data and provides a strong data base for developing and testing statistical adjustment techniques to correct for biases.

THURSDAY, 26/FEB/2026
01:30 - 02:30 PM RH, AUDITORIUM

4.5: POSTER SESSION

SESSION CHAIR: TOBIAS RETTIG

PRESENTATIONS

DEVELOPING A MEASUREMENT OF MASCULINITY NORMS: INSIGHTS FROM THE MEN4DEM PROJECT

VERA LOMAZZI
University of Bergamo, Italy; vera.lomazzi@unibg.it

RELEVANCE & RESEARCH QUESTION

The literature identifies different types of masculinity. Hegemonic masculinity refers to the type of masculinity legitimating unequal gender relations (between men and women, between masculinity and femininity and among masculinities). Hegemonic masculine norms for the "ideal man" include traits and practices like being strong, successful, independent, unemotional, in control. Hypermasculinity, aggressivity, sexual prowess are generally glorified.

Differently, caring masculinity refers to the idea that, without rejecting masculinity, men are able to adopt what (traditionally) is seen as a feminine characteristic. Here, the emotional dimension is crucial. Recent research refers of men negotiating masculinity e.g., not feeling challenged 'as men' because of their caregiving roles but reinforce interest in 'manly' hobbies/sports suggesting processes in which they negotiate what aspects of masculinity do not fit with their identity and what they do.

Extreme-right and manosphere ideologies tend to privilege hegemonic masculinity, which legitimizes unequal gender relations and anti-democratic stands. But how to measure these norms to investigate their spread in the public opinion?

This contribution describes the development, both conceptually and empirically, of the measurement of hegemonic masculinity norms and explore their spread among the general population of six European countries.

METHODS & DATA

The contribution reports on the development of the new "Hegemonic Masculinity Norms Scale" drafted in the context of the MEN4DEM (Masculinities for the future of European democracy – Horizon Europe, GA n. 101177356), on the cross-cultural comparability challenges (including use of advance translation), results from the online survey experiment conducted to assess reliability and validity. The final measurement is then used to show the spread of hegemonic masculinity norms in Greece, Germany, Italy, Netherlands, Poland, and Sweden. (N=800 in each of the 6 countries; representative samples randomly assigned to exp.setting)

RESULTS

Fieldwork will finish in November 2025

ADDED VALUE

The "Hegemonic Masculinity Norms Scale" brings the following innovation: it combines a multidimensional perspective to masculinity with a continuum approach in its measurement, allowing for in-depth analysis of public opinion on masculinity, an aspect still largely unexplored. Methodology applied to develop the instrument in a cross-national setting can be of interest for cross-cultural survey scholars.

SOCIAL MEDIA SURVEYS IN EMERGING RISKS: MEASURING STRESS DURING BRADYSEISM IN CAMPI FLEGREI, ITALY

MARGHERITA SILAN1, ROCCO MAZZA2, MARTINA SALIERNO3, MANUELA SCIONI1
1University of Padova, Italy; 2University of Bari "Aldo Moro", Italy; 3University of Salerno, Italy; margherita.silan@unipd.it

RELEVANCE & RESEARCH QUESTION

In emerging risk contexts and situations requiring contingent data collection, social media surveys can play a crucial role in analysing population needs and stress levels, offering an innovative monitoring tool. Bradyseism is a phenomenon of slow ground deformation that occurs in the Campi Flegrei area, in the Naples proximity (South of Italy). This phenomenon can cause shallow earthquakes, damage to structures, and an increase in seismic activity. The Campi Flegrei area is characterized by high population density and strong socioeconomic inequalities. This study aims to assess stress levels in the population affected by bradyseism in the Campi Flegrei area.

METHODS & DATA

An online survey was conducted during a period of increased bradyseism activity in the Campi Flegrei area. Meta advertisement platform was used to recruit participants exposed to the phenomenon. The survey, lasting 11 days, collected over 600 completed questionnaires with an investment of 720 Euros. This approach allowed for targeted spatial sampling of the population. The study included residents in the arounds of Pozzuoli (high-risk area) and Ottaviano (control area).

RESULTS

The survey provided valuable information on stress levels among the affected population. The propensity to respond to a survey on social networks is proportional to the interest in the survey topic. This resulted in a higher number of responses from those affected by

the earthquakes, despite our efforts to achieve a balanced sample. The data collected offered insights into the differential impacts of the bradyseism phenomenon across various demographic groups, with a particular focus on gender-based vulnerabilities. Moreover, the stress level, measured using a ten-item psychometric scale, showed a high correlation with both the number of perceived earthquake tremors and their intensity.

ADDED VALUE

This study demonstrates the effectiveness of social media-based surveys in rapidly gathering data during emerging risk situations in a specific area. It provides a cost-effective and time-efficient method for assessing population needs and stress levels in volcanic contexts. The findings contribute to a better understanding of the socio-economic and gender-based impacts of bradyseism, potentially informing more targeted and inclusive public policies for disaster preparedness and response.

AUTOMATED POLITICAL STANCE IDENTIFICATION IN POLITICAL TEXTS

JUAN SALVADOR GOMEZ-CRUCES, YORICK SCHEFFLER, EWAN THOMAS-COLQUHOUN
Universität Potsdam/Hasso-Plattner Institut; juan.gomezcruces@hpi.de

RELEVANCE & RESEARCH QUESTION

This study addresses the growing need for transparent, theory-based methods to analyze political text using artificial intelligence. It asks whether Large Language Models (LLMs) can reliably identify political stances—such as ideological orientation, support for liberal democratic values, and populist rhetoric—without the need for manually labeled data.

METHODS & DATA

The paper employs a Natural Language Inference approach leveraging a retrained model, DEBATE (Burnham et al. 2024). In so doing, we build on expert survey codebooks from reputable sources in political science. The model is validated by comparing its outputs with expert-assigned scores to assess accuracy.

RESULTS

The findings show no statistically significant differences between the model's classifications and those of human experts. The model also demonstrates strong multilingual capabilities across English, German, Italian, Portuguese, and Spanish.

ADDED VALUE

This study introduces a cost-effective, replicable, and theoretically grounded approach for stance detection in political texts. By eliminating the need for data labeling and integrating the method into the forthcoming Automated Political Stance Identification (APSI) platform, it provides an accessible tool for researchers, policymakers, civil society, and the public seeking evidence-based insights into ideological and rhetorical patterns in politics.

LESSONS LEARNED FROM DEVELOPING INDICES FOR SYNDICATED STUDIES

JUSTUS RATHMANN, INNA BECHER, MATTHIAS KELLER, MARC HERTER
YouGov, Switzerland; justus.rathmann@yougov.ch

RELEVANCE & RESEARCH QUESTION

Addressing complex market research questions often requires the integration of data from multiple sources, the development of reliable indices, and coordinated collaboration across organizations. In syndicated studies, even market competitors may work together within a shared research framework. Achieving high quality outcomes demands a deep understanding of the underlying data as well as the expectations and needs of all stakeholders. Only then can indices be developed that are robust, transparent, and capable of supporting strategic decision making.

METHODS & DATA

For a syndicated study, data from two competing companies have been pooled to date. A new questionnaire was developed specifically for the study in which internal employees were surveyed. In addition, data from various customer satisfaction studies were integrated. Financial performance indicators from the partner companies and their suppliers were also included.

From these inputs, distinct non-overlapping indicators were constructed, complemented by a single comprehensive index that synthesizes all individual measures. The development process took place in close collaboration with the syndicated study partners. Moreover, suppliers and customers were invited to contribute their perspectives and assessments, ensuring that the resulting indicators reflect a broad and balanced understanding of the market context.

RESULTS

The analysis highlights several factors that are essential for the development of meaningful indicators. Although large volumes of data are often available, careful dimensionality reduction is necessary to maintain clarity and to avoid overly complex structures. Indicators tend to overlap, which can create unintended effects and lead to misplaced analytical focus. In addition, integrating data from various empirical surveys and KPIs requires appropriate statistical procedures. In our case, we apply, among other methods, inverse-variance weighting to ensure high reliability of the indicators.

ADDED VALUE

We illustrate how heterogeneous data, differing organizational perspectives, and collective development processes can be brought together to create a coherent and reliable measurement framework. The insights gained expand the understanding of how indices can be designed in collaborative settings and show how transparent methods increase acceptance and long-term usability. We offer practical guidance to develop robust indicators for complex syndicated studies.

GOODNIGHT, PRINCE OF DARKNESS: OZZY OSBOURNE'S DEATH AS A GLOBAL FACEBOOK EVENT

GAL YAVETZ
Bar-Ilan University, Israel; gal.yavetz@biu.ac.il

RELEVANCE & RESEARCH QUESTION

When news of Ozzy Osbourne's death broke in July 2025, millions of Facebook users brought collective grief together in a global digital ritual. This study examines how mourning, nostalgia, and fandom unfolded on social media, asking: how does celebrity death evolve into a networked media event? Building on media event theory (Dayan & Katz, 1992) and the concept of affective publics, the research explores how emotions, algorithms, and cultural memory intersect when a legendary figure dies "live" in the networked sphere.

METHODS & DATA

We collected 46,390 public Facebook posts published between 22 and 30 July 2025 that mentioned "Ozzy Osbourne". We then preprocessed the texts through tokenization, lemmatization, and stopword removal, and extracted bigrams using the gensim Phrases model. We applied Latent Dirichlet Allocation (LDA) topic modeling (Coherence = 0.73) to identify dominant themes and conducted sentiment analysis with TextBlob to measure polarity on a –1 to +1 scale. Finally, we visualized temporal and emotional dynamics to trace the evolution of discourse over time.

RESULTS

Four dominant topics emerged: Death & Legacy, Memorial & Tribute, Music & Fan Tributes, and Tribute & Farewell. Posting peaked on 23 July (13,866 entries) and declined sharply thereafter, mirroring the lifecycle of an online event. Sentiment was predominantly positive (mean polarity = 0.105), with Tribute & Farewell (50% positive) and Music & Fan Tributes (49% positive) dominating. Frequent collocations such as "black_sabbath", "prince_darkness", "rest_peace", and "farewell_concert" illustrate how users merged grief, admiration, and mythology into a shared cultural narrative.

ADDED VALUE

This study conceptualizes celebrity death as a hybrid of media spectacle and participatory ritual. By integrating computational text analysis with cultural interpretation, it shows how algorithmic publics collectively reframe loss into commemoration, blending journalism, fandom, and emotional performance. The findings extend Dayan and Katz's framework to the digital sphere, demonstrating how networked mourning transforms moments of death into global acts of mediated memory.

THURSDAY, 26/FEB/2026
10:15 - 11:15 AM RH, AUDITORIUM

GOR THESIS AWARD MASTER

SESSION CHAIR: OLAF WENZEL

PRESENTATIONS

MEASURING AMBIVALENT SEXISM IN LARGE LANGUAGE MODELS: A VALIDATION STUDY

JANA JUNG
University of Mannheim, Germany; jana.jung@uni-mannheim.de

RELEVANCE & RESEARCH QUESTION

Large language models (LLMs) often perpetuate gender biases and stereotypes learned from uncurated training data, underscoring the importance of reliable methods for measuring bias. Although several methods of measuring gender bias have been proposed, there are concerns about the ambiguities and inconsistencies in how these methods conceptualize and operationalize gender bias. A promising alternative to these existing methods is LLM psychometrics, which applies psychometric tests originally developed for humans to evaluate human-like characteristics of LLMs, such as personality. Psychometric tests have several advantages: they are grounded in psychological theory, have been rigorously validated, and provide standardized assessment tools. However, it remains unclear how these assessments can be meaningfully applied to LLMs. In this thesis, I explore whether the Ambivalent Sexism Inventory (ASI) [4] can be used to measure sexism in LLMs. To address this question, I propose a systematic validation approach grounded in established psychometric standards.

METHODS & DATA

2.1 Inducing Individuals Using Context data
To approximate psychometric testing conditions, we conceptualize an LLM as a representation of a population and induce individuals by prompting the model with different context information. Two types of contexts are used: human-chatbot interactions and personas. Using personas is a method that has been used in previous LLM psychometrics studies. However, this does not reflect how most users incorporate LLMs into their everyday lives. Therefore, we also use real-life interactions between users and LLM-powered chatbots as a context type. For each context type, n = 300 contexts are sampled from the Chatbot Arena Conversations dataset [5] and the Persona Hub dataset [6], respectively.

2.2 Ambivalent Sexism Inventory
The ASI consists of 22 items, such as, “Women exaggerate problems they have at work.” Answers are provided using a 6-point Likert scale ranging from 0 (disagree strongly) to 5 (agree strongly). The overall ASI score of one context is computed by averaging the answer scores of all items given that context.

2.3 Data Collection
Each item is prompted individually to mitigate the effects of item order. In addition to the item, the prompt contains the context, general instructions, and answer scale. Answer scores are extracted directly from a model’s text response. Data is collected from six state-of-the-art LLMs, including Llama 3.3 70B Instruct, Mistral 7B Instruct, and Qwen 2.5 7B Instruct.
2.4 Psychometric Quality Critera
The systematic validation is conducted by first evaluating reliability (i.e., the consistency of a test) using three criteria:
(1) Internal consistency (Cronbach’s alpha): How consistent are responses across all items of the ASI?
(2) Alternate-form reliability (Pearson correlation): How consistent are the ASI scores when rephrasing the items without changing their meaning?
(3) Option-order symmetry (Pearson correlation): How consistent are the ASI scores when randomly changing the order of answer options?
If reliability is deemed acceptable based on established psychometric interpretation thresholds, validity (i.e., the extent to which a test measures what it is supposed to measure) is evaluated in a second step based on the following three types of validity:
(1) Concurrent validity (Pearson correlation): Does the ASI score align with the amount of sexist language used in a downstream task (writing reference letters)?
(2) Convergent validity (Pearson correlation coefficient): Does the ASI score align with the sexism score of another established sexism scale, the Modern Sexism Scale [7]?
(3) Factorial Validity (Confirmatory factor analysis; CFA): Do the items group together in a way that makes sense based on the underlying theory?

These analyses are conducted for each of the six models and two context types.

RESULTS

For 10 out of the 12 model-context type combinations, the ASI displays low reliability, indicating low consistency and high measurement error. Only for two models - Llama 3.3 70B Instruct and Qwen 2.5 7B Instruct - reliability is deemed acceptable when using Persona Hub contexts. However, the validity evaluation for these two cases indicates low validity. Crucially, for both Llama 70B and Qwen, ASI scores do not significantly correlate with sexist behavior in the downstream task ($r = -0.1$, $p = .523$ and $r = 0.08$, $p = .612$ respectively). Based on these findings, the ASI is not considered valid for any of the six LLMs.

However, the results also indicate that the choice of context type influences evaluation outcomes.

ADDED VALUE

The findings of this thesis emphasize that tests developed and validated for humans should not be automatically assumed to be valid for LLMs. This underscores the importance of conducting validation studies before interpreting psychological test scores for LLMs, which has rarely been done in the field of LLM psychometrics [3].

However, the results also raise several important questions and issues on how to conduct such validations. What constitutes an "individual" in the context of LLMs? How should a sample of "individuals" be selected? These issues highlight the need to adapt the psychometric validation approach to the LLM domain in future studies.

[1] BLODGETT, S. L., LOPEZ, G., OLTEANU, A., SIM, R., AND WALLACH, H. Stereotyping Norwegian Salmon: An Inventory of Pitfalls in Fairness Benchmark Datasets. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers) (Aug. 2021).
[2] PELLERT, M., LECHNER, C. M., WAGNER, C., RAMMSTEDT, B., AND STROHMAIER, M. AI Psychometrics: Assessing the Psychological Profiles of Large Language Models Through Psychometric Inventories. Perspectives on Psychological Science 19, 5 (Sept. 2024).
[3] LÖHN, L., KIEHNE, N., LJAPUNOV, A., AND BALKE, W.-T. Is Machine Psychology here? On Requirements for Using Human Psychological Tests on Large Language Models. In Proceedings of the 17th International Natural Language Generation Conference (Sept. 2024).
[4] GLICK, P., AND FISKE, S. T. Hostile and Benevolent Sexism: Measuring Ambivalent Sexist Attitudes Toward Women. Psychology of Women Quarterly 21, 1 (Mar. 1997).
[5] ZHENG, L., CHIANG, W.-L., SHENG, Y., ZHUANG, S., WU, Z., ZHUANG, Y., LIN, Z., LI, Z., LI, D., XING, E. P., ZHANG, H., GONZALEZ, J. E., AND STOICA, I. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena, Dec. 2023. arXiv:2306.05685.
[6] GE, T., CHAN, X., WANG, X., YU, D., MI, H., AND YU, D. Scaling Synthetic Data Creation with 1,000,000,000 Personas, Sept. 2024. arXiv:2406.20094.
[7] SWIM, J. K., AIKIN, K. J., HALL, W. S., AND HUNTER, B. A. Sexism and racism: Old-fashioned and modern prejudices. Journal of Personality and Social Psychology 68, 2 (1995).

AI FOR SURVEY DESIGN: GENERATING AND EVALUATING SURVEY QUESTIONS WITH LARGE LANGUAGE MODELS

ANNA FUCHS
Ludwig-Maximilians-Universität Munich, Germany; anna.fuchs@stat.uni-muenchen.de

RELEVANCE & RESEARCH QUESTION

Designing high-quality survey questions is a complex task. With the rapid development of large language models (LLMs), new possibilities have emerged for supporting this process, particularly in the automated generation of survey items. Despite growing interest in LLM applications from industry, published research in this area remains limited, and little is known about the quality and characteristics of survey items generated by LLMs, as well as the factors influencing their performance. This work provides the first in-depth analysis of LLM-based survey item generation and systematically evaluates how different design choices, such as prompting technique, model choice, and fine-tuning, affect item quality.

METHODS & DATA

Five LLMs, namely GPT-4o, GPT-4o-mini, GPT-oss-20B, LLaMA 3.1 8B, and LLaMA 3.1 70B, generated survey items for 15 concepts across four domains: work, living conditions, national politics, and recent politics. For each concept, three prompting techniques (zero-shot, role, and chain-of-thought prompting) were applied. Additionally, the best performing model and prompting combination, namely GPT-4o-mini combined with chain-of-thought prompting, was fine-tuned on high-quality survey items to explore the effects of fine-tuning on the quality of generated items.

To assess the quality of the generated survey items, we use the Survey Quality Predictor (SQP; https://sqp.gesis.org), a validated tool for estimating the quality of attitudinal survey items. SQP predicts item quality, validity, and reliability based on a range of coded formal and linguistic characteristics of the survey items, e.g., question formulations, the response scale types, and the inclusion of an introduction text. To code these characteristics, we used an LLM-assisted procedure. A GPT-4.1-nano model was fine-tuned on examples of coded characteristics from the SQP database to automate the coding process, supplemented by manual revision.

The analysis allows us to evaluate not only overall quality but also around 60 specific survey item characteristics, offering a detailed view of how LLM-generated questions differ.

RESULTS

The findings show striking differences in survey characteristics across the different models and prompting techniques. As an example, the type of response scale strongly differed by model family. Closed-source GPT models consistently generate five-category, bipolar response scales with medium correspondence between numeric and verbal labels. They rarely include a 'don't know' option and typically begin with the negative end of the response scale.

LLaMA models show greater variation: they generate a wider range of response options, show greater inconsistency between numeric and verbal labels, differ in whether scales start with the positive or negative option, and the inclusion of a 'don't know' option varies by model size. The inclusion of an introduction text in the survey item depends strongly on the type of prompting technique used. Survey items generated with chain-of-thought prompting often included an introduction text.

Regarding the quality of the generated survey items, the findings show that the prompting technique employed is a primary factor influencing the quality of LLM-generated survey items. Chain-of-thought prompting leads to the most reliable outputs. Closed-source GPT models generally produce more consistent and higher-quality items than open-source LLaMA models. The open-source GPT-oss-20B model failed to complete the given task, i.e., it did not produce a usable survey item in 68% of the cases. The survey topics `work' and `national politics' generate survey items with higher quality compared to `living conditions' and `recent politics'. Among all configurations, GPT-4o-mini combined with chain-of-thought prompting achieved the best overall results. Fine-tuning on high-quality survey items added variety in survey item characteristics but did not lead to noticeable improvements in item quality.

ADDED VALUE

For the GOR community, the study offers empirical evidence on how LLMs can (and cannot) be reliably integrated into questionnaire design workflows, providing a systematic basis for evaluating emerging AI tools in survey research and informing methodological decisions in applied settings. In addition to highlighting the strengths and limitations of LLMs for survey item generation, the work helped to identify concrete weaknesses within the SQP-based evaluation pipeline, particularly regarding the coding of characteristics. The development of an LLM-assisted coding procedure contributes to future research in AI-supported survey design by laying the necessary groundwork for a fully automated pipeline that can code the SQP item attributes at scale.

ADAPTIVE CODE GENERATION FOR THE ANALYSIS OF DONATED DATA WITH LARGE LANGUAGE MODELS

MIGER SHKREPA
University of Mannheim, Germany; miger.shkrepa@gmail.com

RELEVANCE & RESEARCH QUESTION: In an increasingly digitalized world, people are generating large amounts of digital trace data daily as a result of the constant recording of their information and activity. These data contain information with the potential to facilitate human behavior studies due to their accessibility and fine-grained nature. Data donation has emerged as a promising approach to get access to such data from specific online platforms, such as Instagram. In a data donation study, people are invited to answer a survey and subsequently asked to request, download, and finally donate their online platform data to research. However, as raw data, these donated Data Download Packages (DDPs) might contain highly sensitive or personal information. Previous research on data donation has solved this issue by developing privacy-preserving methods to anonymize and aggregate data directly on participants' devices. Data donation workflows typically rely on scripts designed to extract and process relevant data from these structures.

Researchers who develop these extraction scripts often face the challenge that the data structure of DDPs is undocumented for most online platforms and can be subject to modifications from the entities that issue them. As such, the processing scripts that are developed to extract information from these DDPs face deprecation threats and require the researchers' manual intervention. This thesis investigates the feasibility of employing Large Language Models (LLMs) in automatically processing and extracting relevant information from these structures by generating code as an alternative to traditional, manually maintained data analysis tools and in an effort to minimize manual script development and adjustment. It also aims to make data donation research more accessible by examining this approach as a means for researchers without technical expertise to analyze and interpret structurally complex DDPs.

METHODS & DATA: This study evaluates six open-source LLMs across thirteen Instagram DDPs of varying size and structural complexity to examine their capabilities and limitations in this domain. The models are asked a series of data-specific queries designed to assess their abilities in interpreting the provided data package, retrieving correct information, and processing it accordingly. Two main experimental settings are implemented and compared, where context on the supplied information, response formatting, and data structure are provided to the models across different setups. In the first setting, this information is integrated through an external knowledge base using Retrieval-Augmented Generation (RAG), whereas in the second, it is directly provided to the models within the prompt. The outputs of the models in each setting are assessed in terms of accuracy, common error patterns, and code generation habits.

RESULTS: Although RAG is a methodology originally developed to reduce hallucination and improve models' responses, our findings reveal that, for this task, directly providing the context in prompts yields higher accuracy in comparison. Overall, the models' performance is unsatisfactory in both settings. Their shortcomings can be attributed to the overly complex structure of the packages used in this evaluation. Ambiguous and similar naming conventions are observed throughout the Instagram DDP structure and enhance the hallucination and inaccuracy of the models' outputs. In an effort to improve the models' performance, when manually designed and explicit instructions on how to navigate these packages are provided, the LLMs perform substantially better, with some achieving near-optimal results. These instructions would also be subject to changes in response to structural updates to the DDPs provided by the online platforms. Ironically, the very manual intervention that this research sought to reduce is necessary in achieving greater performance for the evaluated LLMs at this current stage.

ADDED VALUE: This research presents an evaluation of LLMs for adaptive code generation in donated data analysis tasks and explores their potential as an alternative approach in this domain. It lays the groundwork for easing the skill-based entry that data researchers without programming skills face in navigating the structural complexities and challenges inherent in data donation studies. By identifying the strengths and current limitations of LLMs in understanding and adapting to evolving data structures, this study helps set realistic expectations for their application in this field while highlighting the considerable room for future improvement.

REINFORCEMENT LEARNING FOR OPTIMISING THE VEHICLE ROUTING PROBLEM

ABIGAIL HAYES
University of Mannheim, Germany; abigail.hayes@uni-mannheim.de

RELEVANCE & RESEARCH QUESTION

The Travelling Salesman Problem (TSP) requires identifying the shortest route to visit a set of locations and return to the starting depot, given the locations and the distances between them. The Vehicle Routing Problem (VRP) is similar but with multiple vehicles, with the number of customers per route limited by vehicle capacities. This was first formulated in 1959 by Dantzig and Ramser [1] and has since been further extended, such as restricting the time window for customer visits.

The VRP is NP-hard, and so finding exact solutions becomes computationally intractable with increased problem instance size. As a result, heuristic and metaheuristic algorithms have replaced exact

calculations. More recently, routes are found using reinforcement learning (RL) approaches, requiring less task specific expertise than heuristic selection but with the computational cost of deep learning.

With each new RL model, the creators generally compare against either heuristics or the very earliest RL models from Nazari et al. [2] and Kool et al. [3]. Therefore, it is often unclear the extent to which new methods achieve good performance. Additionally, whilst heuristics require computation for each new instance, RL models instead have a high up-front cost due to their training which must be justified. This thesis evaluates a range of RL methods against heuristics for the Capacitated VRP. Specific attention is paid to complex problem variants (the Time Window variant) and complex problem instances (such as with more customers). It aims to determine whether RL methods should be used over established and theory-informed heuristics.

METHODS & DATA

Six RL models for the CVRP are compared, covering a range of architectures and optimisation procedures: Nazari [2], AM [3], AM PPO, POMO [4], Sym-NCO [5] and MDAM [6]. The Nazari model is unique in using a recurrent neural network, with all others based on a Transformer architecture. The Attention Model (AM) uses a Graph Attention Model with REINFORCE. This is further adapted to use proximal policy optimisation (AM PPO), to consider multiple possible solutions concurrently (Policy Optimisation with Multiple Optima (POMO)) or to exploit problem and solution symmetries by adapting the REINFORCE rewards (Sym-NCO).

The Multi-Decoder Attention Model (MDAM) is a further Transformer model with multiple decoders. Additionally, a two-step method is included using a heuristic to group the nodes and an AM TSP model to build each route. Only AM, AM PPO, Sym-NCO and MDAM are applied to the CVRP-TW. All RL models are provided with large quantities of random problem instances for training.

Initial evaluation of the models uses standard CVRP and CVRP-TW benchmarks. Most of the CVRP benchmark instances are randomly generated, whilst the CVRP-TW benchmark deliberately varies the location patterns, customers per vehicle and time windows. Further systematic test instances were created for the CVRP by varying the position of the depot, distribution and number of customers and maximum demand from a single customer. This variation induces varied instance difficulties.

A robust baseline is provided by generating solutions with a range of heuristic and metaheuristic algorithms. Evaluation on all datasets considers how often valid solutions are returned and compares the average solution distances. Additional considerations are the number of vehicles used in a solution and the computation time.

RESULTS

Whilst all apart from one heuristic approach finds valid solutions for all instances, multiple RL models fail for at least one problem instance. The Nazari model implementation specifically had validity issues which excluded it from all further evaluation.

Regarding the quality of the solutions, the heuristics and metaheuristics consistently provide solutions within 4% of the optimum for CVRP benchmark problems, whilst even the best RL methods are more than 10% worse. When it provides valid solutions, the AM TSP two-step model outperforms other RL models, likely due to multiple TSP training instances being included in a single CVRP training instance. AM and POMO deliver the best RL solutions but consistently fall behind the heuristics.

A similar pattern appears with the CVRP-TW problems, although Syn-NCO is instead consistently the best RL model. With the more complex problem variant, all models are prone to using more vehicles than the optimal solution, likely due to the increased difficulty of finding any valid solution. The gap in performance between the heuristics and the RL models often increases with the more complex instances e.g. with more customers.

A time limit of 60 seconds per instance is pre-specified for heuristics. For the RL methods, producing solutions for an individual instance is almost instantaneous, but training can require large amounts of time. The 4 hours 10 minutes for training and testing the AM (10 location) model is the equivalent of finding solutions to 250 instances using a heuristic method. When training on larger instances the situation is much worse, with a training time increase of 358% for POMO with 20 customers compared to 10.

ADDED VALUE

The results demonstrate the robustness of the heuristic and metaheuristic methods such that RL approaches could only be viable where the same model will be used extensively. The RL models with even longer training periods might meet the heuristic performance but would only justify the computation cost after 1000s of uses.

The surprisingly competitive performance of the two-step AM TSP approach signals the crucial role of method design. This is already apparent in the heuristics and metaheuristics but often overlooked for RL. In this instance, breaking down the problem into smaller steps and only using RL for the more difficult component enables the model to train more efficiently.

[1] Dantzig, G. B. and J. H. Ramser (1959). The Truck Dispatching Problem. Management Science 6 (1), 80–91

[2] Nazari, M., A. Oroojlooy, L. Snyder, and M. Takac (2018). Reinforcement Learning for Solving the Vehicle Routing Problem. In Proceedings of the 32nd International Conference on Neural Information Processing Systems

[3] Kool, W., H. van Hoof, and M. Welling (2019). Attention, Learn to Solve Routing Problems! In Proceedings of the 7th International Conference on Learning Representations

[4] Kwon, Y.-D., J. Choo, B. Kim, I. Yoon, Y. Gwon, and S. Min (2020). POMO: Policy Optimization with Multiple Optima for Reinforcement Learning. In Proceedings of the 34th Conference on Neural Information Processing Systems

[5] Kim, M., J. Park, and J. Park (2022). Sym-NCO: Leveraging Symmetricity for Neural Combinatorial Optimization. In Proceedings of the 36th Conference on Neural Information Processing Systems

[6] Xin, L., W. Song, Z. Cao, and J. Zhang (2021, May). Multi-Decoder Attention Model with Embedding Glimpse for Solving Vehicle Routing Problems. In Proceedings of the 35th AAAI Conference on Artificial Intelligence

THURSDAY, 26/FEB/2026
11:30 - 12:30 AM RH, AUDITORIUM

GOR THESIS AWARD PHD

SESSION CHAIR: OLAF WENZEL

PRESENTATIONS

WHO COUNTS? SURVEY DATA QUALITY IN THE AGE OF AI

LEAH VON DER HEYDE1,2
1LMU Munich, Germany; 2Munich Center for Machine Learning; leah.vonderheyde@stat.uni-muenchen.de

RELEVANCE & RESEARCH QUESTION

Large language models (LLMs) have been hoped to make survey research more efficient, while also improving survey data quality. However, as they are based on Internet data, LLMs may come with similar pitfalls as other digital data sources with regard to making inferences about human attitudes and behavior. As such, they not only have the potential to mitigate, but also to amplify existing biases. In my dissertation, I investigate whether and under which conditions LLMs can be leveraged in survey research by providing empirical evidence of the potentials and limits of their applications. Potential applications of LLMs span the entire survey life cycle – before, during, and after data collection – where an LLM could act as a research assistant, interviewer, or respondent. Potential challenges for data quality stem from LLMs' training data, alignment processes, and model architecture, as well as the research design. I focus on two major applications of LLMs covering both representational and measurement challenges and test these applications in challenging, previously unexamined contexts.

METHODS & DATA, RESULTS

Two studies address the most prominent discussion regarding LLM-based survey research: Using LLM-generated "synthetic samples". Coverage bias in LLM training data and alignment processes might affect the applicability of this approach. In one study, I test to what extent LLMs can estimate vote choice in Germany. To generate a synthetic sample of eligible voters in Germany, I create small profiles matching the individual characteristics of the 2017 German Longitudinal Election Study respondents. These "personas" include socio-demographic and attitudinal information known to be associated with voting behavior. Prompting GPT-3.5 with each persona in German, I ask the LLM to predict each respondents' vote choice in the 2017 German federal elections and compare these predictions to the survey-based estimates on the aggregate and subgroup levels. I find that GPT-3.5 does not predict citizens' vote choice accurately, exhibiting a bias towards the Green and Left parties, and making better predictions for more "typical" voter subgroups.

While the LLM is able to capture broad partisan tendencies, it tends to miss out on the multifaceted factors that sway individual voters. As a consequence, not only are LLM-synthetic samples not helpful for estimating how groups likely swinging an election, such as non-partisans, will vote, they also risk underestimating the popularity of parties without a strong partisan base. Such samples thus provide little added value over survey-based estimates. Furthermore, the results suggest that GPT-3.5 might not be reliable for estimating nuanced, subgroup-specific political attitudes.

In a second study, I extend the previous study to the entire European Union (EU), this time focusing on an outcome unobserved at the time of data collection: the results of the 2024 European Parliament elections. I create personas of 26,000 eligible voters in all 27 EU member states based on the Eurobarometer and compare the proprietary LLM GPT-4-Turbo with the open-source LLMs Llama-3.1 and Mistral. A week before the European elections in June 2024, I prompted the LLMs with the personas in English and asked them to predict each person's voting behavior, once based only socio-demographic information, and once also featuring attitudinal variables.

To investigate differences in LLMs' bias across languages, I selected six diverse EU member states for which I prompted the LLMs in the respective country's native language. After the elections' conclusion, I compare the aggregate predicted party vote shares to the official national-level results for each country. LLM-based predictions of future voting behavior largely fail – they overestimate turnout and are unable to accurately predict party popularity. Only providing socio-demographic information about individual voters further worsens the results. Finally, LLMs are especially bad at predicting voting behavior for Eastern European countries and countries with Slavic native languages, suggesting systematic contextual biases.

These findings emphasize the limited applicability of LLM-synthetic samples to public opinion prediction across contexts. Without further adaptation through, e.g., fine-tuning, LLMs appear infeasible for public opinion prediction not just in terms of accuracy, but also in terms of efficiency, highlighting a trade-off between the recency and level of detail of available survey data for synthetic samples.

In the third study, I investigate the usability of LLMs for classifying open-ended survey responses. Due to their linguistic capacities, it is likely that LLMs are an efficient alternative to time-consuming manual coding and the pre-training of supervised machine learning models, but it is unclear how the sparse, sometimes competing, existing findings generalize and how the quality of such classifications compares to established methods. I therefore test to what extent different LLMs can be used to code German open-ended survey responses on survey motivation from the GESIS Panel.pop Population Sample. I prompt GPT-4-Turbo, Llama-3.2, and Mistral NeMo in German with a predefined coding scheme and instruct them to classify each survey response, comparing zero- and few-shot prompting and fine-tuning. I evaluate

the LLMs' performance by contrasting its classifications with those made by human coders. Only fine-tuning achieves satisfactory levels of predictive accuracy. Performance differences between prompting approaches are conditional on the LLM used, as overall performance differs greatly between LLMs: GPT performs best in terms of accuracy, and few-shot prompting leads to the best performance. Disregarding fine-tuning, the prompting approach is not as important when using GPT, but makes a big difference for other LLMs. Further, the LLMs struggle especially with non-substantive catch-all categories, resulting in different distributions. The need for LLMs to be fine-tuned for this task implies that they are not the resource-efficient, easily accessible alternative researchers may have hoped.

I discuss my findings in the light of the ongoing developments in the rapidly evolving LLM research landscape and point to avenues for future work.

ADDED VALUE

This dissertation makes both methodological and applied contributions to survey research. It discusses types and sources of bias in LLM-based survey research and empirically tests their prevalence from multiple comparative angles. It also showcases concrete applications of LLMs across several steps in the research process and several substantive topics, explaining their practical implementation and highlighting their potentials and pitfalls. Overall, this dissertation thus provides guidance on ensuring data quality when using LLMs in survey methodology, and contributes to the larger discourse about LLMs as a social science research tool.

CORRECTING SELECTION BIAS IN NONPROBABILITY SAMPLES BY PSEUDO WEIGHTING

AN-CHIAO LIU
Utrecht University, Netherlands, The; a.liu2@uu.nl

RELEVANCE & RESEARCH QUESTION

Statistics are often estimated from a sample rather than from the entire population. If the inclusion probability of the sample is unknown to the researcher, that is, a nonprobability sample, naively treating the sample as a simple random sample may result in selection bias.

Attention to correcting selection bias is increasing due to the availability of new data sources, for example, online opt-in surveys and data donation. These data are often easy to collect and may be so-called "Big Data" considering the large inclusion fraction of the population.

This dissertation consists of four scientific papers, where two of them are published in influential journals and the other two are under review. In the first paper, a novel framework for correcting selection bias in nonprobability samples is proposed. It follows with the discussion of three practical challenges, and possible solutions to them are provided.

METHODS & DATA

In the framework paper, the general idea is to construct a set of unit weights for the nonprobability sample by borrowing the strength of a reference probability sample. If a proper set of weights is constructed, design-based estimators can be used for population parameter estimation given the weights. To evaluate the uncertainty of the estimated population parameter, a pseudo population bootstrap procedure is proposed, given different relations between the nonprobability sample and the probability sample.

Three practical challenges for pseudo-weighting are also discussed. Namely, the model selection for bias correction, the imbalanced samples, and the small area estimation. Simulation studies and applications on real data are shown in each paper to reflect the usage of the proposed framework and the possible solution of the practical challenges.

RESULTS

The proposed framework is flexible, and many kinds of probability estimation models can be used. The question is raised about how to select a proper model given the population parameter in question. After a series of performance measures were tested, we found that modeling the target variable when evaluating the performance of weights may be useful. The second challenge comes from the large size of the nonprobability sample. Since we often have a large nonprobability sample assisted with a small probability sample, we end up with an imbalanced combined sample, which can cause problems when estimating model parameters.

Several remedies for imbalanced samples are discussed, and the proposed framework is also adjusted accordingly. The results show that SMOTE is a promising technique for dealing with imbalanced samples. Finally, we look at the scenario where not only the population level estimates are of interest, but also subpopulation estimates. Several approaches to combine pseudo weights with small area estimation are discussed. Of all approaches, we found that combining a hierarchical Bayesian model with weights is a relatively stable estimation approach. If both population-level and area-level estimates are of interest, aligning the weighted estimates with estimated marginal totals may be a better option.

ADDED VALUE

This research provides practical suggestions on how to deal with possible selection bias in nonprobability samples, which is gaining more and more importance in the age of digitalization and low response rates.

FRIDAY, 27/FEB/2026
12:00 - 01:00 AM RH, AUDITORIUM

10.4: COLLECT, SHARE, ACT: THE POWER OF ACTIVATED KNOWLEDGE

SESSION CHAIR: LISA DUST

This session brings together diverse perspectives on how organisations can manage consumer or audience knowledge more effectively and translate insights into action. Experts from technology, media, and in-house research share practical experiences and strategic reflections on making insights accessible, connected, and truly impactful within organisations.

"MIND THE GAP!" ON THE IMPORTANCE OF DATA LITERACY AND KNOWLEDGE MANAGEMENT IN THE DIGITAL AGE

JAN ISENBART
ARD MEDIA/agma, Germany; jan.isenbart@ard-media.de

The level of professionalism in any branch can be determined, among other things, by looking at their practice of knowledge management. In marketing, advertising, and media, the performance seems rather poor. In addition to that, there is a growing gap between the sheer amount of data and studies available for practitioners and their abilities to judge these data for quality and relevance. This may lead to faulty-and possibly costly-decisions. This talk is intended as a wake-up call for the pitfalls underneath bad data sets for ill-informed practitioners.

FROM INSIGHTS TO IMPULSES FOR ACTION

ROSINA BARBANERA
Deutsche Welle, Germany; rosina.barbanera@dw.com

Organizations rarely suffer from a lack of data or lack of knowledge. But action on this data can be scarce.

Many research findings confirm what teams accross the organization already sense in their daily work. Still, these insights often fail to translate into concrete decisions. Let's talk open about this "action gap" and how it can be approached. For a large-scale audience reserach project the Market and Audience Insights team of DW redesigned the way how insights are activated internally. By building a dedicated insights platform and running a multi-level roadshow tailored to different stakeholder groups, the team created continuous impulses that made insights visible, relevant, and actionable.

In this presentation we'll discuss this approach, the role of a thorough stakeholder analysis, the fun in creating a home for your insights, the achieved impact, and what can be done differently next time.

VALUES-BASED CUSTOMER TARGETING IN THE AGE OF AI

CARINA FRISCH
Uranos GmbH, Germany; carina.frisch@uranos.io

In the age of artificial intelligence, companies have access to unprecedented amounts of behavioral data—yet more data does not automatically lead to better decisions. While AI can optimize what people click or buy, it often fails to understand why they act. This presentation argues how in an algorithmic world, values-based segmentation can provide deeper insights into increasingly fragmented societies shaped by personalized media ecosystems. Micromilieus help to identify deep motivational structures and cultural codes that drive decision-making. When integrated into AI platforms, dynamic insight pools can be created, that enable companies to truly understand their audiences and align strategy, communication, and innovation with authentic human meaning. The presentation also touches on the ethical risks of AI optimization, arguing that AI strategies should position human understanding as the true competitive advantage in the AI era.

FRIDAY, 27/FEB/2026
02:00 - 03:00 PM RH, AUDITORIUM

11.4: DGOF KI (AI) FORUM:

SESSION CHAIR: OLIVER TABINO, YANNICK RIEDER, GEORG WITTENBURG

Running before you can walk? First-hand, opinionated experiences on the most recent developments in AI. – Inetractive Roundtable Discussion. Join us for an insightful session where industry experts put innovative solutions to the test and explore how AI can unlock new opportunities in market research.

INSPIRATION SESSION
(HELD IN GERMAN)

THURSDAY, 26/FEB/2026

ORAL PRESENTATIONS

2.1: AI AND SURVEY RESEARCH

10:15 - 11:15 AM RH, SEMINAR 01

TALKING TO RESULTS: LLM-ENABLED DISCUSSIONS OF QUANTITATIVE SURVEY FINDINGS

SOPHIA MCDONNELL1, AARON HEINZ1, GEORG WITTENBURG2, ANJA LANGNESS3, NICOLE KLEEB3
1Verian, Germany; 2Inspirient; 3Bertelsmann Stiftung; sophia.mcdonnell@veriangroup.com

RELEVANCE & RESEARCH QUESTION

Large Language Models (LLMs) are increasingly used to transform quantitative survey data into accessible, actionable insights. Yet, their adoption in professional research settings is limited by concerns over accuracy, transparency, and the risk of hallucinated results, especially when advanced statistical methods are involved. This presentation addresses the central question: How can LLMs be reliably integrated into the workflow of survey analysis and reporting, ensuring both rigor and user trust and adding real value?

We showcase the Bertelsmann Stiftung's deployment of the 'SurveyMind' chatbot, which enables interactive, natural-language exploration of survey findings while meeting strict quality and security standards.

METHODS & DATA

Our approach combines precomputed, validated statistical analyses with a Retrieval-Augmented Generation framework. SurveyMind operates exclusively on anonymized survey data processed on secure servers, with no data used for external model training. The system accesses a pool of results that are pre-generated using a validated methodology, thus ensuring that the LLM never performs its own calculations but instead retrieves and explains trustworthy findings. We detail the iterative development process, including multi-stage user testing, expert ratings, and continuous feedback from researchers and clients at the Bertelsmann Stiftung.

RESULTS

SurveyMind has demonstrated that LLM-enabled chat interfaces can deliver accurate, context-aware textual outputs, such as executive summaries and press releases, directly from raw, interview-level survey data. The system supports nuanced queries, adapts to user feedback, and provides transparent links to underlying statistical analysis, significantly reducing turnaround times for survey deliverables and empowering both researchers and clients to explore findings independently. Moreover, a variety of further applications emerged during our client discussions, from pre-drafting press releases to checking data and comparing trends to presenting data more accessibly to specific target groups such as young people.

ADDED VALUE

This work provides a blueprint for integrating LLMs into quantitative research workflows without compromising on analytical rigor or data security. By combining technical safeguards with user-centered design, our chatbot solution bridges the gap between advanced automated analytics and practical communication needs. The Bertelsmann Stiftung's client perspective highlights real-world benefits, including enhanced efficiency, flexibility in reporting, and new ways to disseminate results.

TESTING THE PERFORMANCE AND BIAS OF LARGE LANGUAGE MODELS IN GENERATING SYNTHETIC SURVEY DATA

CHARLOTTE MUELLER, BELLA STRUMINSKAYA, PETER LUGTIG
Utrecht University, Netherlands, The; c.p.muller@uu.nl

RELEVANCE & RESEARCH QUESTION

The idea to generate so-called silicon survey samples with large language models (LLMs) has gained broad attention in both academic and market research as a promising timely and cost-efficient method of data collection. However, previous research has shown that LLMs are likely to reproduce limitations and biases found in their training data, including underrepresentation of certain subgroups and imbalanced topic coverage. Using survey data from a probability-based online panel in Netherlands, we conduct a large-scale analysis examining model performance across different item types (factual, attitudinal, behavioral) and social-science-related topics, to identify when and for whom synthetic approaches perform best. We further explore strategies to mitigate potential performance limitations, including few-shot leveraging of LLMs on longitudinal survey data.

METHODS & DATA

We compare existing survey data with LLM-generated synthetic data, using the 17th wave of LISS Panel fielded in 2024. We selected nine

survey items that cover key social-science topics, such as health, family, work, and political values. We chose the items based on their characteristics (factual, behavioral, attitudinal) as well as outcome type. We leverage the LLM agents based on two different few-shot learning tasks: a sociodemographic setup, which draws on seven individual background variables, and a panel-informed setup, which additionally incorporates previous responses. We compare the generated data across five different proprietary LLMs, including GPT-4.1, Gemini 2.5 Pro, Llama 2 Maverick, Deepseek-V3 and Mistral Medium 3.

RESULTS

Our first results show a deep lack of accuracy in the sociodemographic few-shot setup with an average accuracy of 0.2 and a strongly underestimated variance across items and topics. Prediction errors vary significantly across subgroups, particularly by age, showing differences not only in magnitude but also in the direction of errors. Adding longitudinal survey responses to the few-shot input substantially improves the prediction quality, showing a 40 to 60 percentage point increase in overall accuracy, correcting variance underestimation, and reducing subgroup disparities.

ADDED VALUE

Our research contributes to a more responsible application of silicon survey samples by providing practical guidance for researchers and survey practitioners to evaluate and improve AI-generated datasets.

TRANSCRIBING AND CODING VOICE ANSWERS OBTAINED IN WEB SURVEYS: COMPARING THREE LEADING AUTOMATIC SPEECH RECOGNITION TOOLS

MELANIE REVILLA1, CARLOS OCHOA1, JAN HÖHNE2, MICK COUPER3
1RECSM-UPF, Spain; 2DZHW, Leibniz University Hannover; 3University of Michigan; melanie.revilla@upf.edu

RELEVANCE & RESEARCH QUESTION

With the rise of smartphone use in web surveys, voice or oral answers have become a promising methodology for collecting rich data. Voice answers present both opportunities and challenges. This study addresses two of these challenges—labor-intensive manual transcription and coding of responses, by answering the following research questions: (RQ1) How do three leading Automatic Speech Recognition (ASR) tools –Google Cloud Speech-to-Text, OpenAI Whisper, and Vosk– perform across various dimensions? (RQ2) How similar or different are the codes of transcribed responses generated by a human and the OpenAI GPT-4o model?

METHODS & DATA

We used data collected in the Netquest opt-in online panel in Spain in February/March 2024. The questionnaire included over 80 questions, mainly about citizens' perceptions of nursing homes. This study focuses on one open-ended narrative question in which respondents were asked to explain why they selected a given answer in a prior closed question on the amount of information nursing homes provide to the general public. For this question, participants were initially asked to answer through voice recording. In a follow-up, respondents skipping the question were also offered to type in a text box. We extracted various aspects from the transcriptions and compared them across ASR tools and human vs GPT coding. After data cleaning, 859 panellists were used for analyses.

RESULTS

We found that each of the ASR tools has distinct merits and limits. Google sometimes fails to provide transcriptions, Whisper produces hallucinations (false transcriptions), and Vosk has clarity issues and high rates of incorrect words. Human and LLM-based coding also differ significantly. Thus, we recommend using several ASR tools and implementing human as well as LLM-based coding, as the latter offers additional information at minimal added cost.

ADDED VALUE

This study offers valuable insights into the feasibility of using voice answers. Depending on the most critical quality dimension for each study (e.g., maximizing the number of transcriptions or achieving the highest clarity), it provides guidance on selecting the most suitable ASR tool(s) and insights into the extent to which LLMs can assist with manual tasks like answer coding.

2.2: PARADATA AND METADATA

10:15 - 11:15 AM RH, SEMINAR 02

METADATA UPLIFT OF SURVEY DATA FOR RESEARCH DISCOVERY AND PROVENANCE

JON JOHNSON1, PAUL BRADSHAW2, SUPARNA DE3, DEIRDRE LUNGLEY4
1University College London; 2Scotcen; 3University of Surrey; 4University of Essex; jon.johnson@ucl.ac.uk

RELEVANCE & RESEARCH QUESTION

The profusion of data from the introduction of CAI has created three main problems for researchers, volume, complexity and understanding quality.

The disjointed nature of many survey data collections has fragmented this across many organisations, during in which much valuable information is lost or opaque to the researcher using data at the end of the data lifecycle.

METHODS & DATA

Focused and well contructed Machine Learning offers the possibility of automating at scale the available metadata resources into standardised metadata which can be made available in repositories for discover, and creating the detailed granular metadata that a researcher needs to evaluate quality of complex survey prior to data applications or access.

The collaboration between CLOSER, University of Essex, University of Surrey and Scotcen has been developing machine learning models, utilising the CLOSER Discovery metadata store to improve the timeliness and accuracy of metadata extraction to deliver high quality metadata.

RESULTS

Preliminary results will be presented on the success and challenges faced in taking complex survey instruments and rendering them into DDI-Lifecycle for ingest into repository platforms and the opportunites for further enhancement of these metadata resources for reuse of questions into the survey development pipeline.

ADDED VALUE

The abilty to create high quality reuable metadata across the survey specification, collection, management and dissemination lifeycle would bring efficiencies in terms of costs, improvements in quality, discoverability and understanding of these complex data resources.-

BEYOND THE QUESTIONNAIRE: LINKING PASSIVELY METERED PLATFORM DATA WITH SURVEYS FOR AUDIENCE PROFILING

DAVID GOLDSCHMIDT, LUKAS STEIN
Datapods GmbH, Germany; david@datapods.app

RELEVANCE & RESEARCH QUESTION

The integration of large-scale passively metered data with established survey methodologies has become a central development in contemporary market and social research. In particular, digital trace data originating from major platform operators such as Google, Meta and TikTok represents a highly promising source for enhancing sociological measurement, audience segmentation and modeling. However, substantial challenges remain with respect to obtaining continuous, consent-based access to such platform data, and to linking heterogeneous data types in a methodologically robust and privacy-compliant way.

METHODS & DATA

Datapods has established a novel approach with its own proprietary user panel that allows for the combination of survey methodologies to define socio-economic, value-based and demographic profiles with direct copies of the personal data from big tech companies. We first established the baseline for these profiles by utilizing common survey methodologies.

These measures serve as our ground truth for subsequent validation. In a second step, we linked these baseline profiles with corresponding behavioral data streams, including web-browsing histories, YouTube viewing histories and interaction logs on Instagram, Facebook and TikTok. We identified key indicators for different data types to be the most influential for the profile of the panelist and joined data across types to ensure a holistic picture about the user.

RESULTS

Results indicate that a subset of survey items can be replaced with digital trace data. Researchers can, in practice, rely on high-quality, consent-based personal data to assign users to pre-defined socio-demographic and value-based target group profiles, and to identify fundamental clusters and segments within the panel. The results suggest that passively collected platform data can function as a proxy for many conventional survey indicators.

ADDED VALUE

This passively metered, platform-data-based approach to user profiling and segmentation substantially enhances survey-centric designs and, for certain research questions, can partially substitute conventional survey data collection. It enables more granular behavioral indicators and provides a scalable solution for continuous audience measurement and sociological analysis.

VISUALIZING THE ANSWERING PROCESS: EXPLORING MODE DIFFERENCES WITH RESPONDENT-LEVEL PARADATA FROM THE IAB ESTABLISHMENT PANEL

CORINNA KÖNIG1, MARIEKE VOLKERT1, JOSEPH SAKSHAUG1,2
1Institute for Employment Research, Germany; 2LMU Munich; corinna.koenig@iab.de

RELEVANCE & RESEARCH QUESTION

Understanding how respondents interact with survey instruments is crucial for facilitating the response process and improving data quality. Especially for establishments there is still a lack of insights into their response behavior. By analyzing respondent-level paradata we aim to explore the answering process in detail.

We investigate how establishments navigate through the online questionnaire of the IAB Establishment Panel, focusing on differences between survey modes (CAI versus Web) and samples (panel versus refreshment). Following this, we investigate whether distinct response patterns can be identified and if there is a need for a tailored response process by utilizing important establishment characteristics.

METHODS & DATA

We analyze detailed respondent-side paradata from the IAB Establishment Panel, conducted annually by the Institute for Employment Research (IAB). Since 2018, the survey has been implemented in a mixed-mode design with computer-assisted personal interviewing (CAI) and an online mode (Web) using identical software. In 2022, we collected paradata logging every click, answer, and timestamp at the second level. After creating an audit trail for each respondent, we identify appropriate paradata indicators and apply cluster analysis to identify groups of establishments with similar navigation and response behaviors.

RESULTS

The visualization of paradata via audit trails reveals differences in navigation behavior. Some establishments follow a straightforward sequence, while others loop back or perform multiple checks before submission. Paradata indicators reveal that Web respondents take longer, more breaks, use more tree view, edit more answers and drop-out more often than CAI respondents. Preliminary results of the clustering analysis show that we can identify two clusters for each

combination of sample and mode. Cluster 1 seems to include the linear respondents, while cluster 2 includes all other respondents with more conspicuous response behavior.

ADDED VALUE

By combining visualization and clustering of establishments response processes, we provide an empirical approach of utilizing paradata. Our results help to understand how establishments navigate and respond to a survey. Simultaneously, we give recommendations for the survey design and evaluate mixed-mode establishment surveys.

2.3: MEDIA STUDIES

10:15 - 11:15 AM RH, SEMINAR 03

DEPLOYING ONLINE EXPERIMENTS TO INVESTIGATE CONTENT CREDIBILITY IN SENSOR-BASED JOURNALISM

IRINA BOBOSCHKO, CLAUDIA LOEBBECKE
University of Cologne, Media and Technology Management, Germany; claudia.loebbecke@uni-koeln.de

RELEVANCE & RESEARCH QUESTION

The emerging field of sensor-based journalism relies on data beyond human reach collected by sensors (Diakopoulos, 2019; Loebbecke & Boboschko, 2020). Research on sensor-based journalism (Boboschko & Loebbecke, 2025) studies the impact of identity cues and outlet reputation on content credibility (Sundar, 1999). Communication and media studies (Boller et al., 1990; Wathen & Burkell, 2002) analyze how testimonial-based 'argument strength' drives journalistic content credibility. Aiming to complement both research streams, we ask how argument strength influences content credibility in the context of sensor-based journalism.

METHODS & DATA

This study deploys a between-subjects online experiment (N= 853) followed by multi-group covariance-based structural equation modeling. Two treatment groups read an article on traffic affecting air pollution in London, one drawing evidence from sensor data, the other from testimonials. As endogenous latent variables, we measure argument strength with four items and content credibility with five items. Measurement items, wording of all items, descriptive statistics, standardized factor loadings, squared multiple correlations for each indicator, construct-level reliability, and convergent validity are available upon request.

RESULTS

Sensor-based journalism fosters argument strength and credibility formation; statistical details and interpretations are available upon request. Controlled online experiments allow for realistically simulating (journalistic) media consumption.

ADDED VALUE

Promoting research in sensor-based journalism in times of AI-based hallucinations – an increasingly relevant phenomena in today's democracies.

WAR, ANXIETY, AND DIGITAL BEHAVIOR: HOW ARMED CONFLICT RESHAPES ONLINE MEDIA CONSUMPTION AND SOCIAL MEDIA ENGAGEMENT

VLAD VASILIU1, HANEEN SHIBLI2, GAL YAVETZ3
1The Max Stern Yezreel Valley College, Emek Yezreel, Israel; 2University of Washington, Seattle, WA, USA; 3Bar-Ilan University, Ramat Gan, Israel; vladv@yvc.ac.il

RELEVANCE & RESEARCH QUESTION

Armed conflicts fundamentally disrupt daily life, yet their impact on digital media behavior remains understudied, particularly across different population groups. This study examines how the October 2023 Israel-Hamas war affected media consumption and social media activity among Jewish and Arab citizens of Israel, investigating the relationship between anxiety and online behavior changes. The research addresses: How did the war affect social media activity across populations? Did anxiety levels differ between groups? What is the relationship between anxiety and social media behavior during conflict?

METHODS & DATA

We conducted representative online panel surveys with 505 Jewish and 146 Arab respondents through Ipanel during October-November 2023. The survey measured retrospective media consumption (before vs. during war), social media activity changes, and generalized anxiety using the GAD-7 scale. Media consumption was categorized into four daily time brackets. Social media activity changes were measured on a five-point scale. We employed correlation analyses examining relationships between anxiety and social media behavior across population groups.

RESULTS

Jewish respondents showed dramatic media consumption increases, with heavy users (4+ hours/day) rising from 19.2% to 38.6%, while Arab respondents showed minimal change. Approximately 31% of Jewish and 6% of Arab respondents reported increased social media activity. No significant anxiety differences emerged between populations (Jewish M=9.38, Arab M=8.67). Critically, opposing anxiety-behavior patterns emerged: among Arabs, higher anxiety correlated with decreased social media activity ($r=-.237$, $p<.01$), while Jewish respondents showed non-significant positive trends ($r=.059$), suggesting anxiety drives increased engagement in majority populations but withdrawal in minorities during intergroup conflict.

ADDED VALUE

This research challenges assumptions that populations respond uniformly to crisis events, revealing that majority and minority groups exhibit fundamentally different digital coping mechanisms during intergroup conflicts. The opposing anxiety-behavior relationships

suggest social media serves different psychological functions across populations—as an information-seeking and community tool for majorities versus a stress source for minorities. These findings have critical implications for targeted crisis communication, mental health interventions in conflict zones, and understanding how digital platforms may amplify psychological disparities during emergencies. Results indicate researchers and practitioners must account for population-specific responses when designing crisis communication strategies and evaluating social media's role in conflicts.

GENERATIVE AI IN MEDIA 2025

KRISTINA HAGEN, LISA LAUTENBACH, ALINA SCHROEDER, CLAUDIA ROSENKÖTTER, FRANZISKA RIEDER
Annalect/OMG Solutions GmbH; lisa.lautenbach@omc.com

RELEVANCE & RESEARCH QUESTION

Generative AI has emerged in recent years as a key technology shaping both consumers' everyday lives and the media and advertising industry. It competes with major search engines such as Google in searching information and opens up new efficient and profitable opportunities for advertisers to engage consumers – for example, through AI-generated advertising, synthetic brand influencers or AI shopping agents.

This study addresses three core questions: (1) How do consumers in Germany use generative AI in daily life and how have their attitudes evolved over time? (2) What opportunities and risks are perceived regarding GenAI? (3) To what extent do usage and attitudes toward GenAI influence the acceptance of AI-generated advertising?

METHODS & DATA

The research applies a mixed-method design comprising three modules:
- Module 1: In-depth interviews (n=18) with non-users, occasional users, and heavy users to explore attitudes and everyday usage of GenAI as well as perceptions of AI-generated advertising by blind-testing AI-generated TV commercials (A/B test).
- Module 2: Online survey (n≈4,000; ages 16–69; online-representative) to quantify acceptance and perception of AI-generated advertising by blind-testing AI-generated TV commercials (A/B test).
- Module 3: Tracking study (three annual waves since 2023; n≈1,000 per wave; ages 16–69; online-representative) to analyze temporal changes in usage, attitudes and acceptance of GenAI.

RESULTS

Preliminary findings indicate a steady increase in familiarity and regular usage. GenAI is perceived as indispensable for information gathering. While efficiency is highly valued, concerns about data privacy, misinformation and job displacement persist but do not significantly affect the rate of adoption and usage. AI-generated advertising is considered forward-looking but evokes mixed reactions: younger, tech-savvy users show higher acceptance, whereas older cohorts remain skeptical. Synthetic influencers face the strongest resistance, while AI-generated TV commercials and shopping assistants receive comparatively higher acceptance.
Final results will be available in January 2026.

ADDED VALUE

The study provides empirically grounded insights for advertisers to strategically leverage the potential of generative AI. It identifies target groups open to AI-based advertising formats and highlights acceptance barriers. These findings support the development of effective communication strategies in an increasingly AI-driven media landscape.

2.4: INNOVATION IN MEASURE-MENT INSTRUMENTS

10:15 - 11:15 AM RH, SEMINAR 04

IMPROVING MEASUREMENT OF MIGRATION PREFERENCES: A CHOICE-BASED CONJOINT APPROACH TO STUDYING REFUGEE RESETTLEMENT DECISIONS

ARMIN KÜCHLER, MARVIN BÜRMANN
Bielefeld University, Germany; armin.kuechler@uni-bielefeld.de

RELEVANCE & RESEARCH QUESTION

Refugees in first host countries often face adverse conditions including limited legal rights, restricted employment access, and inadequate social services. While resettlement programs offer long-term settlement opportunities for particularly vulnerable refugees, little is known about how these individuals evaluate and prioritize different aspects of potential destination countries. Traditional survey items fall short in capturing refugees' complex decision-making processes regarding resettlement. This study addresses this gap by employing a choice-based conjoint experiment to reveal how refugees eligible for resettlement weight different factors in their settlement decisions.

METHODS & DATA

The study uses data from an online survey (n = 375) conducted jointly by BAMF-FZ and the University of Bielefeld, targeting particularly vulnerable refugees eligible for resettlement procedures. Respondents complete up to four paired comparisons of hypothetical destination countries that vary across eight key characteristics: family presence, diaspora size, crime rates, language skills, attitudes toward refugees, political systems, labor market access, and lifestyle familiarity. After

each choice, respondents evaluate their perceived opportunities for building a future in both countries, enabling analysis of both relative preferences and absolute settlement potential assessments.

RESULTS

The conjoint approach successfully reveals how refugees trade off different aspects of potential host countries when making settlement decisions. Preliminary findings indicate that attitudes toward refugees and political systems emerge as the most influential factors in resettlement preferences. Notably, the relative importance of these factors varies considerably across refugee groups, highlighting the heterogeneity in decision-making priorities shaped by diverse experiences and backgrounds.

ADDED VALUE

This study advances the methodological and substantive understanding of refugee migration preferences. By focusing on refugees in their first host countries, the study provides reliable insights into real-world decision-making under precarious conditions. Thus, it offers valuable lessons for researchers targeting vulnerable groups through choice-based conjoint experiments.

FEAR IN FLIGHT: MEASURING DIGITAL RISK PERCEPTION AND EMOTIONAL RESPONSES TO AVIATION SAFETY IN ROMANIA

ANTONIO AMUZA
University of Bucharest, Romania; antonio.amuza@fjsc.ro

RELEVANCE & RESEARCH QUESTION

The perception of flight safety represents an exemplary case for studying how digital media mediate fear, risk, and trust. While aviation remains one of the safest modes of transport, emotional reactions to incidents are often disproportionate, amplified by online news and algorithmic visibility. This paper asks: How does exposure to digital media shape fear of flying and perceived aviation risk? and To what extent can online sentiment predict self-reported anxiety in surveys? The study addresses an underexplored link between risk perception, media consumption, and emotion measurement in digital environments.

METHODS & DATA

I integrate two complementary data sources. First, a national online survey (N = 921) which measures emotional, cognitive, and behavioral components of flight anxiety, including trust in institutions, information sources, and prior flying experience. Second, we analyze a corpus of 43,000 online news articles containing the keyword "aviation" (2022–2023), collected from Romanian mainstream media. Using topic modeling (LDA), sentiment analysis (NRC and BERT-based classifiers), and correlation with survey variables, we identify symbolic frames of fear ("catastrophe," "loss of control," "safety reassurance"). All analyses were conducted in R.

RESULTS

Findings reveal that individuals with higher exposure to negatively valenced news content report significantly higher levels of aviation anxiety ($p < .01$). Topic models show a predominance of emotional frames emphasizing loss, danger, and uncertainty, while trust in technological expertise moderates these effects. The integrated model combining sentiment scores and survey data predicts fear-of-flying.

ADDED VALUE

This study advances methodological innovation in online research by combining survey data with large-scale digital trace data. It demonstrates how affective communication and digital media coverage can shape collective perceptions of risk, offering new insights for both computational social science and public communication of safety. The proposed framework can be extended to other domains of perceived technological risk (AI, climate, or health).

INDUSTRY AND OCCUPATION CODING: A COMPARISON OF OFFICE-BASED CODING AND A CLOSED-LIST APPROACH

CRISTIAN DOMARCHI1, OLGA MASLOVSKAYA1, CURTIS JESSOP2, LISA CALDERWOOD3, MATT BROWN3
1University of Southampton, United Kingdom; 2National Centre for Social Research (NatCen), United Kingdom; 3Centre for Longitudinal Studies, University College London, United Kingdom; C.Domarchi@soton.ac.uk

RELEVANCE & RESEARCH QUESTION

Social surveys often collect information about industry and occupation. The traditional 'gold-standard' approach has been to capture this information by asking open questions about job title and duties which are later classified into standardised coding systems by expert coders. The shift towards self-completion surveys makes this more challenging as respondents often provide insufficient information to facilitate accurate coding. In addition, office-based coding is expensive and time consuming.

Closed-list questions, where respondents choose from pre-defined categories, could reduce response burden and coding costs and potentially also remove ambiguities inherent to open-ended responses. However, a challenge is that category labels can be difficult to interpret. This paper assesses whether closed-list questions for industry and occupation can produce industry and occupation classifications comparable to those derived from manually coded open text responses.

METHODS & DATA

We use data from two waves of the NatCen Panel survey, a UK probability-based online panel, which employs a mixed-mode design combining online self-administration with telephone interviews. In these two waves, the panel supplemented its standard industry and occupation open-text questions (which are manually coded) with closed-list questions. We compare agreement between the two methods, using both descriptive analysis and regression models.

RESULTS

For industry, the mean agreement rate between the closed-list and 1-digit manual codes was 64%, with variation across industries. Agreement was lower for occupation, at 56% for 1-digit codes and 46% for 2-digit codes, with significant differences across occupation types. Agreement rates were also influenced by sociodemographic and the length of open-text entries, with shorter descriptions generally leading to higher agreement rates.

ADDED VALUE

This study provides a first empirical assessment of the quality of occupation and industry data collected using a closed-list approach, offering evidence on the method's potential and limitations. The findings help identify potential improvements for collecting industry and occupation data in online surveys.

3.1: APP-BASED DATA COLLECTION

11:30 - 12:30 AM RH, SEMINAR 01

WHY ARE PEOPLE UNWILLING TO PARTICIPATE IN SMARTPHONE APP DATA COLLECTION? RESULTS FROM QUALITATIVE IN-DEPTH INTERVIEWS

ALEXANDER WENZ
University of Mannheim, Germany; a.wenz@uni-mannheim.de

RELEVANCE & RESEARCH QUESTION

Smartphones have become promising tools for collecting digital behavioral, sensor, and survey data in the social sciences. However, the recruitment of study participants who are willing to install a smartphone app and fully participate throughout the study period remains a challenge. This paper reports on results from qualitative in-depth interviews to better understand the mechanisms underlying the decision to participate in smartphone app data collection.

The study addresses the following research questions: (1) What difficulties and risks do people perceive in smartphone app data collection? (2) How do people perceive the collection of different forms of data, in particular digital behavioral, sensor, and survey data? (3) Under what conditions would people be more willing to participate and adhere in smartphone app data collection? (4) Which strategies for increasing participation and adherence might work best for whom?

METHODS & DATA

In-depth interviews with n = 30 participants with different sociodemographic characteristics (age, gender, education) and types of smartphone use are conducted in November/December 2025. Participants are presented with a hypothetical research app that involves the collection of survey, GPS, Internet browsing, and app usage data. They are subsequently asked a series of open-ended questions and probes based on a semi-structured discussion guide. The interview data are analyzed using a thematic analysis approach.

RESULTS

The study will examine reasons for and against participating in smartphone app data collection and the extent to which these vary across the different data components. The potential reasons include those related to the study design, its perceived relevance to science and society, participants' interest in the study topic, and their concerns about privacy and data security. In addition, it will be investigated how different study designs, such as different monetary and non-monetary incentive structures, affect people's willingness to participate, and for which population subgroups these might be more effective.

ADDED VALUE

The study provides insights into people's decision-making processes regarding their participation in smartphone app data collection and aims to inform researchers about how to best design and implement smartphone-based studies.

READY, SET, GO! DATA COLLECTION FOR THE HOUSEHOLD BUDGET SURVEY WITH AN APP

MAAIKE KOMPIER, JANELLE VAN DEN HEUVEL, JELMER DE GROOT
Statistics Netherlands, Netherlands, The; me.kompier@cbs.nl

RELEVANCE & RESEARCH QUESTION

Household budget surveys provide an excellent opportunity for the implementation of smart services. The response burden for reporting expenditures in a questionnaire is high and is likely to result in underreporting. Over the past years, Statistics Netherlands has worked on implementing smart features in the household budget survey, aiming to reduce response burden and increase data quality. The fieldwork with the app will start at the end of 2025.

In this presentation, the main findings and efforts of the qualitative and quantitative tests done in 2025 will be presented.

METHODS & DATA

Five tests were conducted using the StatNL app for the household budget survey: two usability tests, a field test with a fresh sample and two internal StatNL tests with the mobile application for the household budget survey. The usability tests included cognitive interviews and talk-aloud methods to gather in-depth qualitative insights in the entire process from invitation letter to final participation using the app. In the field test, we experimented with different interviewer roles and gathered data in a realistic setting to gain quantitative insights in the responses, which was enriched with qualitative evaluation questionnaires and telephone interviews. The two internal tests each served a different goal: with the first, receipts were collected for training and validating the machine learning algorithm, whereas the second set-up quick iterative evaluation rounds of the final redesigns of the app.

RESULTS

A timeline of the different projects that were completed before the implementation will be presented and key findings will be highlighted. The qualitative tests demonstrated over and over that abilities of respondents vary widely. With the results of the field and internal tests the application and algorithms were further developed and improved.

ADDED VALUE

In this presentation we outline the work and testing needed for successful implementation of an app for the household budget survey. We demonstrate the need to engage a multidisciplinary team of experts to cover all disciplines involved in such an innovative project.

THE EFFECT OF CONTROL OVER DATA COLLECTION ON WILLINGNESS TO PARTICIPATE IN APP-BASED DATA COLLECTION

THIJS CARRIERE, BELLA STRUMINSKAYA, LAURA BOESCHOTEN
Utrecht University, Department of Methodology and Statistics; t.c.carriere@uu.nl

RELEVANCE & RESEARCH QUESTION

Providing participants with control over the data collection process has been linked to increased willingness to participate in app-based studies. As the mechanism for this relation is not yet well understood, it is unclear under which conditions and for whom this effect holds. We use the Leverage Saliency Theory of survey participation (LST; Groves et al. 2000) which states that the interplay of personal leverages and the saliency of study characteristics influence whether someone participate in a study to answer the research question: 'How does control over data collection influence willingness to participate in app-based data collection?'.

METHODS & DATA

In December 2024, we conducted a vignette study in the Centerpanel, a longitudinal panel of the general population in the Netherlands. In a vignette study (n=1200, response rate: 86.8%), participants rated whether several study characteristics influenced their decision to participate in app-based data collection. Participants were presented with a hypothetical study description and mock-up research app screen, and indicated how willing they would be to download this app. Participants were randomly assigned to be presented with either 7 control and transparency features in the app description and app mock-up screen or no control features in the presented app at all.

RESULTS

Participants generally rated control features as important in their decision to participate in app-based studies. The data retention period, control over third-party sharing, and the study's relevance are rated as most important. When controlling for speeding and attention check, we find a main effect of control over data collection on willingness to download the app ($t = 3.85$, $p < .001$). However, we find no significant interactions between experimental conditions and rated importance of control features. Therefore, we find no support for the LST.

ADDED VALUE

Our findings show what type of control and transparency features people find most important when participating in app-based studies, and replicates the effect of control on willingness to participate. Therefore, it can be valuable for researchers to invest time and resources in developing control functionalities when conducting app-based research, as this might cause participants to be more willing to adopt a research app.

3.2: VIDEO AND IMAGES IN SURVEY RESEARCH

11:30 - 12:30PM RH, SEMINAR 02

A PICTURE IS WORTH A THOUSAND WORDS: FACTORS INFLUENCING THE QUALITY OF PHOTOS RECEIVED THROUGH AN ONLINE SURVEY

PATRICIA IGLESIAS1, JESSICA DAIKELER2, FIONA DRAXLER3
1Centre d'Estudis Demogràfics, Spain; 2GESIS - Leibniz Institute for the Social Sciences; 3University of Mannheim; piglesias@ced.uab.es

RELEVANCE & RESEARCH QUESTION

Photos, which can be easily captured with smartphones, offer new opportunities to improve survey data quality by replacing or complementing conventional questions. However, potential advantages may vary by respondent characteristics, including age, gender, education, photo-taking and -sharing frequency, self-assessed verbal, mathematical, and spatial skills, and comfort with new technologies. This study addresses the following research question: To what extent do such individual characteristics affect the quality of photos of books at home submitted through an online mobile survey?The topic was selected because the number of books is a well-established proxy for cultural and socioeconomic capital in social sciences, yet its measurement through conventional questions is challenging.

METHODS & DATA

Participants were asked for information on the books in their home through conventional survey questions and/or by submitting photos. This information covered the number of books, their intended audience (illiterate children, literate children and teenagers, and general audiences), language, and storage. Quality was evaluated with indicators tailored to book information, drawing on literature about data quality in survey methodology and in computer vision. The survey, conducted in 2023, used Netquest's opt-in panel in Spain. The target population was parents of children in primary school. Of 1,270 individuals reaching the questions on books, 703 were asked for photos, and 238 provided at least one.

RESULTS

Photo quality was not systematically affected by most of the studied variables. Although older participants submitted more photos, extracting book information from their photos was less feasible, and they experienced more capture and submission issues. The findings

suggest that photo-based surveys can be collected across diverse populations; however, age may be an exception in contexts like books, where finer details are required to extract information consistently.

ADDED VALUE

Evidence on the quality of photos submitted through online surveys remains limited, especially when photos address a relevant social science topic, such as counting books and using that information to characterize respondents. This study contributes to the literature by showing that photos can be requested from broad audiences with comparable quality across respondents, and it offers practical guidance for researchers collecting photos.

ENHANCING PARTICIPATION IN VISUAL DATA COLLECTION IN ONLINE SURVEYS: EVIDENCE FROM AN EXPERIMENTAL STUDY ABOUT REMOTE WORK ENVIRONMENTS

LEYRE PADILLA, MELANIE REVILLA
RECSM - Universitat Pompeu Fabra, Spain; leyre.padilla@upf.edu

RELEVANCE & RESEARCH QUESTION

Collecting visual data through web surveys offers a promising way to obtain richer and more accurate information. Yet participation in image-based tasks remains low, and evidence on how to motivate respondents while maintaining data quality is limited. This study examines whether different strategies can help researchers achieve higher participation when requesting photos of remote work environments in web surveys: (1) offering an extra incentive specifically for sharing photos, beyond the standard survey participation incentive, (2) adding a follow-up prompt immediately after the initial photo request emphasizing the importance of sharing the photos, and (3) sending a reminder email to respondents not sharing the photos initially. Furthermore, we investigate whether the timing of incentive announcement (either at the initial request or only in the reminder) also affects participation.

METHODS & DATA

An experiment is being conducted in the opt-in online panel Netquest in Spain (N = 1,200) among adults who have worked remotely for at least seven hours per week in the past two months. Respondents are randomly assigned to one of three groups: (1) Control – asked to upload three photos of their home workspace without extra incentive; (2) Incentive – offered 10 extra panel points for uploading all photos; and (3) IncentiveReminder – offered the same extra incentive but only in the reminder. All groups get the follow-up prompt. Descriptive analyses will be used to assess differences in participation.

RESULTS

The survey is currently being programmed. Data collection is planned for early December. Results are expected by early February 2026. We expect that extra incentives, follow-up prompts and reminders will all increase photo submission rates, and that announcing the incentive initially will be more effective. We will test this by comparing different participation indicators, such as break-off or item nonresponse defined in different ways, across experimental groups and across both conventional and visual formats.

ADDED VALUE

This study provides evidence on how different strategies might help improving participation in photos requests. Findings will help improve the design of online surveys combining textual and visual data while balancing respondent burden and data richness.

VIDEO-INTERVIEWS IN MIXED-MODE PANEL SURVEYS: SELECTIVE FEASIBILITY AND DATA QUALITY TRADE-OFFS

JULIA WITTON1, CARINA CORNESSE1,2, MARKUS GRABKA1, SABINE ZINN1,3
1DIW Berlin, Germany; 2GESIS Leibniz Institute for the Social Sciences, Germany; 3Humboldt University of Berlin, Germany; jwitton@diw.de

RELEVANCE & RESEARCH QUESTION

As survey methodologists seek innovative approaches to address declining response rates and evolving technological landscapes, Computer-assisted live video interviewing (CALVI) emerges as a promising hybrid mode combining the personal interaction of Computer-assisted personal interviews (CAPI) with the convenience of remote participation as given in Computer-assisted web interviews (CAWI). This study addresses the critical question: Is CALVI a feasible and useful add-on in a German general population mixed-mode household panel survey? Understanding CALVI's viability is essential for survey researchers considering technological adaptations to maintain data quality while accommodating respondent preferences.

METHODS & DATA

We implemented CALVI using a randomized controlled experimental design within the 2024 Innovation Sample of the Socio-economic Panel Study (SOEP-IS): (E1) 1,261 households from the established panel (previously using CAPI) randomly assigned to CALVI versus (C1) 1,106 households continuing with traditional face-to-face interviews; and (E2) 409 households from the 2023 refreshment sample (recruited in CAWI) assigned to CALVI versus (C2) 1,513 households continuing with web self-completion. Both experimental groups retained fallback options to their previous data collection modes. We analyzed participation rates, technical implementation success, and data quality indicators.

RESULTS

Among 1,670 households invited to CALVI, 376 video interviews were conducted. Technical implementation proved largely successful, though 23% of interviews experienced connectivity issues, particularly when interviewers worked remotely. Unit nonresponse among CALVI-invited households is significantly higher compared to established modes, with pronounced effects among former CAWI participants (response rate in E1: 56% vs. C1: 61%; E2: 45% vs. C2: 74%). CALVI attracted specific demographic segments: young, highly educated, high-income, full-time employed urban participants from West Germany with reliable internet access. However, CALVI participants demonstrated superior data quality metrics, including lowest speeding rates, minimal item nonresponse, and highest data linkage consent rates.

ADDED VALUE

This research provides empirical evidence for CALVI's selective feasibility in mixed-mode designs, highlighting both opportunities and limitations. While CALVI may reduce overall participation rates, willing participants

generate high data quality and high data linkage consent rates. CALVI serves specific demographic segments effectively, informing strategic decisions about incorporating video interviewing in surveys.

3.3: MARKET AND CUSTOMER RESEARCH

11:30 - 12:30 PM RH, SEMINAR 03

EXPERIMENTAL EVIDENCE ON HOW QUESTIONNAIRE STRUCTURE AFFECTS AWARENESS REPORTING

JUSTUS RATHMANN, INNA BECHER, MARC HERTER
YouGov, Switzerland; justus.rathmann@yougov.ch

RELEVANCE & RESEARCH QUESTION

In brand awareness research, even universally recognized brands that should reach maximum awareness (for example, Google among internet users) often fall short in actual surveys. This suggests that true awareness may be systematically underestimated. Participants in online research panels frequently take part in surveys and are highly familiar with questionnaire formats.

Therefore, respondents may adopt behaviors that reduce cognitive effort or optimize earnings per hour. Our research examines whether this underreporting reflects conscious strategic choices or whether it arises more implicitly through mechanisms such as fatigue.

METHODS & DATA

We embedded a survey experiment in a client study among 996 driving license holders in Switzerland from the nationally representative YouGov panel. Respondents were randomly assigned to one of four experimental groups. We used a 2x2-design with list length (either 8 or 24 brands) and a behavioral nudge (vs. no nudge), stating that the questionnaire length and payout were independent of their awareness of the specific brands) as the experimental conditions. Audi, BMW and Mercedes-Benz can be assumed to have complete awareness among the target population. If the nudge increased reported awareness, this would indicate deliberate effort reduction. If the shorter list increased awareness, this would point to implicit mechanisms such as fatigue. We estimated the effects of the experimental conditions using a linear regression with the number of these brands being reported as recognized as the dependent variable.

RESULTS

On average, respondents recognized 2.85 of the three brands. The linear regression indicates that the shorter list leads to significantly higher reported awareness among the three brands in this condition ($p < 0.05$). The behavioral reminder does not show a positive effect on reported brand awareness.

ADDED VALUE

The effect of list length indicates that respondents do not intentionally underreport awareness. Instead, the results suggest that underreporting is driven primarily by implicit mechanisms such as cognitive fatigue and reduced attentional effort. This implies that survey designers can improve data quality by using shorter and/or splitting brand lists for subgroups wherever possible. By minimizing mental load, researchers can obtain awareness estimates that more closely reflect true recognition levels.

FROM REFLECTION TO INTUITION: INTEGRATING SYSTEM 1 AND SYSTEM 2 MEASURES IN BEHAVIOURAL CAMPAIGN EVALUATION

EMMANUEL GUIZAR ROSALES
YouGov, Switzerland; emmanuel.guizar@yougov.ch

RELEVANCE & RESEARCH QUESTION

Behavioural campaigns – communication initiatives aimed at changing behaviours – play a vital role in both social and market research. Their evaluation requires a holistic approach integrating behavioural, cognitive, and emotional indicators. This is particularly challenging when topics are sensitive to social desirability or strategic responding, such as traffic safety or pricing research. Traditional explicit self-reports (System 2) provide valuable insights into reflective attitudes/intentions but are prone to response bias and often overlook intuitive or automatic aspects of behaviour. Implicit methods complement them by capturing unconscious processes (System 1). This presentatoin examines how explicit and implicit methods can be combined to obtain a more comprehensive picture of behavioural campaign effects. We explore this in the context of a Swiss road-safety campaign addressing cyclists' "illusion of control" – the tendency to overestimate control in traffic situations and underestimate associated risks.

METHODS & DATA

The evaluation follows a quasi-experimental repeated cross-sectional design combining explicit and implicit measures. Four online panel surveys are conducted between 2025 and 2027 in Switzerland (intervention) and Germany (control), each with 1'000 regular bicycle or e-bike users. Three outcome categories are assessed: knowledge, attitudes/intention, and behaviour. Explicit measures capture self-reported knowledge, perceived risk, and behavioural intentions. Implicit measures employ a reaction-time-based Single Association Test (SAT) with scenario vignettes covering four risk-prone situations: cycling without a helmet, mobile phone usage while cycling, running a red light, and riding under the influence of alcohol. These vignettes assess implicit risk perception and control illusions. The design enables pre/post comparisons between Switzerland and Germany to infer potential causal effects of the campaign over time.

RESULTS

Preliminary findings show that explicit and implicit measures can be integrated within one online survey. Scenario-based SATs complement self-reports by revealing intuitive risk perceptions and control beliefs. Data from two waves (pre- and first post-campaign) provide first insights into early campaign effects.

ADDED VALUE

The study demonstrates a scalable framework for embedding implicit measures in large-scale online surveys. Combining scenario vignettes with reaction-time methods offers a practical blueprint for evaluating sensitive behavioural topics with higher ecological validity and depth.

MEASURING THE BEAUTY OF PRODUCTS: THE PRODUCT AESTHETICS INVENTORY (PAI)

MEINALD THIELSCH1, MAIKE RAMRATH1, BORIS FORTHMANN2, HENNING BRAU3, SIMON EISBACH4

1University of Wuppertal, Germany; 2University of Münster, Germany; 3BSH Hausgeräte GmbH, Germany; 4Provinzial Versicherung, Germany; thielsch@uni-wuppertal.de

RELEVANCE & RESEARCH QUESTION

Aesthetics is quickly processed, has a multitude of consequences for attitudes, recommendations, and purchase behavior—and thus is paramount for the success of products. However, differentiated and validated evaluation tools are lacking, particularly for interactive products. In order to close this gap, we aimed to develop and validate the Product Aesthetics Inventory (PAI) and a short version (PAI-S).

METHODS & DATA

Items were developed in a pre-study (N=6 design experts, N=4 product users). The resulting item pool with 54 questions was tested in an online survey on different types of household appliances (alongside with several validation measures; N=6,002).

In Study 1, data from n=3,000 of these participants were used to determine the number of factors using exploratory graph analysis. In Study 2, the found structure was validated in a confirmatory factor analysis with the remaining n=3,002 participants. A third web-based study (N=1,028) was conducted to further determine construct validity and generalizability among different products such as IT products, home entertainment devices, and power tools.

RESULTS

The final PAI consists of 34 questions covering eight dimensions: visual aesthetics, operating elements, brand logo, feedback sounds, operating noises, haptics, interaction aesthetics, and impression. A higher-order factor of product aesthetics can be determined both with the full PAI and with the 8-item short version PAI-S. Both versions demonstrate excellent reliability, as estimated by Cronbach's alpha.

The decrease in reliability from the full scale (.97) to the short scale (.89) is acceptable. Further, all eight subscales of the PAI achieved good to excellent reliability (range: .85-.92). Validity was confirmed by corresponding correlations with other established scales, intention measures, and overall judgments. Further, Study 3 confirms and extends these results with respect to reliability and validity. Finally, in all three studies, strict measurement invariance was achieved across different product classes for the majority of subscales.

ADDED VALUE

The PAI meaningfully supplements the measurement spectrum of product evaluation beyond classic usability and brand perception. Questionnaire templates, evaluation guides and interpretation aids are freely available online (https://doi.org/10.5281/zenodo.6478042). Further, we will complement this by discussing additional options of response scaling and benchmarks for the PAI.

3.4: ONLINE RESEARCH ON YOUTH AND MENTAL HEALTH

11:30AM - 12:30PM RH, SEMINAR 04

PRESENTATIONS

ENSURING THE VOICES OF YOUNG PEOPLE ARE HEARD: AN INNOVATIVE APPLICATION OF RESPONDENT-DRIVEN SAMPLING WITH PROBABILITY-BASED SEEDS IN THE NATCEN PANEL

LUCIANO PERFETTI VILLA1, OLGA MASLOVSKAYA1, CARINA CORNESSE2, CURTIS JESSOP3

1University of Southampton, United Kingdom; 2GESIS – Leibniz Institute for the Social Sciences; 3National Centre for Social Research

RELEVANCE & RESEARCH QUESTION

Certain population sub-groups are consistently under-represented or excluded from survey samples, or they may appear in insufficient frequencies. In the UK, where a named sample frame is often unavailable, current strategies for boosting samples can be prohibitively costly or impractical, particularly in self-completion survey formats. As survey research in the UK strives to become more inclusive, it is essential to explore effective methods for incorporating and enhancing the representation of these under-represented subgroups.

METHODS & DATA

We employed respondent-driven sampling (RDS) to recruit young adults aged 18 to 24 from the UK, using the NatCen online probability-based panel. Participants completed a brief online questionnaire and could recruit up to five peers using unique digital codes, with incentives for participation and successful recruitment. This process spanned multiple waves, aiming for a sample of about 1,500 participants and allowing for an assessment of recruitment cooperation and network characteristics in a self-completion survey format.

RESULTS

We examined the patterns and predictors of recruitment cooperation in RDS using probability-based seeds. The study also assessed the extent of non-response at various recruitment stages and explored how socio-demographic and network factors influence recruitment success. Intra-chain correlation revealed a strong clustering of respondents by ethnicity, while other characteristics, such as gender, had a less pronounced impact on recruitment correlation. Additionally, the study evaluated the effect of seed composition on recruitment outcomes and considered the implications of violating key RDS assumptions for the quality of the inferences made. Notably, incorporating seeds from outside the target population led to the recruitment of a significant proportion of young adults within the sample.

ADDED VALUE

This study aims to enhance the evidence base for using RDS as a survey data collection method, both in the UK and internationally, by integrating empirical findings with methodological diagnostics. It also introduces a novel approach designed to amplify the voices of young people, a demographic that remains under-represented in social surveys. This perspective has the potential to make surveys more inclusive and methodologically robust.

YOUTH LONELINESS EPIDEMIC: REAL TREND OR SURVEY ARTIFACT?

FRANCESCO BERLINGIERI1, BÉATRICE D'HOMBRES1, CATERINA MAURI2
1European Commission - Joint Research Centre; 2Vrije Universiteit Brussel, Belgium

RELEVANCE & RESEARCH

Question Major health organizations have raised alarms about rising youth loneliness (WHO, 2025), with reports suggesting young people experience the highest loneliness rates and saw the sharpest increases between 2018 and 2022 (OECD, 2024). However, reported prevalence varies dramatically across surveys—from 24% to 50.5% for young adults in different European studies—raising a critical question: are we witnessing a genuine loneliness epidemic or methodological artifacts? While genuine increases are plausible given documented declines in face-to-face social contact (OECD, 2025), the magnitude may be partially explained by methodological shifts from traditional face-to-face to web-based data collection. This study examines how survey mode influences loneliness measurement, with particular focus systematic variation by age, potentially distorting our understanding of youth loneliness trends.

METHODS & DATA

This study quantifies data collection mode effects by comparing loneliness reports across web-based, face-to-face, and telephone collection modes in the Survey on Income and Living Conditions. The analysis focuses on identifying age-specific patterns in how survey mode affects loneliness measurement, examining whether the magnitude of mode effects differs between younger and older respondents. To address potential limitations due to mode self-selection, we use a variaty of techniques including controlling for respondent's health status and limiting the analysis to a subsample of countries. Results from other surveys are included for rubustness (SOEP, Germany). Our research additionaly addresses the role of stigma (national aggregates by age group from the EU Loneliness Survey) to explain observed patterns.

RESULTS

The findings reveal striking age-specific mode effects: young adults report significantly more loneliness in online surveys compared to face-to-face interviews, while older adults exhibit the opposite pattern, reporting higher loneliness levels in face-to-face settings. These divergent patterns suggest that stigma associated with loneliness and social desirability bias operate differently across the lifespan.

ADDED

Value By demonstrating that mode effects vary systematically by age, this research helps disentangle genuine increases in youth loneliness from measurement artifacts introduced by the widespread transition to web surveys. The study contributes to research in the field of survey methodology by documenting how demographic characteristics interact with mode effects for (stigmatized) subjective measures.

DIGITAL MENTAL HEALTH IN WARTIME: AGE, GENDER, AND SOCIOECONOMIC PREDICTORS OF AI THERAPY ACCEPTANCE

AVIA BEN ARI1, VLAD VASILIU2, GAL YAVETZ3
1Technion – Israel Institute of Technology, Haifa, Israel; 2The Max Stern Yezreel Valley College, Emek Yizrael, Israel; 3Bar-Ilan University, Ramat Gan, Israel

RELEVANCE & RESEARCH QUESTION

Despite unprecedented psychological distress in conflict zones and growing treatment gaps, little is known about population readiness to adopt AI-based therapeutic interventions. Following the October 7th attack, Israel experienced severe mental health crisis with 31.5% reporting treatment needs but only 18.3% seeking help, against healthcare strain and six-month waiting periods. This study examines AI therapy acceptance predictors, addressing: How age and gender influence perceived treatment needs and actual help-seeking? What is the relationship between age and AI therapy readiness? Do socioeconomic factors predict beliefs about AI therapy accessibility?

METHODS & DATA

We conducted a representative online panel survey of 502 Jewish adults in Israel through Ipanel in June 2025. The survey measured demographic characteristics, perceived need for psychological treatment, actual treatment-seeking behavior, and readiness to seek AI-based therapy using a validated 6-item scale (=.854). Respondents ranged from ages 18-87 (M=42.66, SD=16.33) with balanced gender distribution. We employed chi-squared tests for demographic-treatment

relationships, Pearson correlations for age-readiness associations, independent samples t-tests for age group comparisons (18-40 vs. 41-87), and Spearman correlations for income-related beliefs.

RESULTS

Significant correlations emerged between age and perceived treatment need (^{2}=103.88, p=.002) and between gender and perceived need (^{2}=7.57, p=.006), with women reporting higher needs (37.2%) than men (25.8%). Age moderated gender effects on actual help-seeking in older adults (^{2}=4.00, p=.045). AI therapy readiness showed negative correlation with age (r=-.114, p=.005), with younger adults (M=16.59) demonstrating higher readiness than older adults (M=15.16; t=2.787, p=.003). Higher household income correlated with decreased belief that AI therapy serves those unable to afford treatment (r=-.135, p=.002).

ADDED VALUE

This research reveals critical demographic and socioeconomic barriers to AI mental health intervention adoption in crisis contexts. Findings challenge assumptions about universal digital health acceptance, showing those most likely to benefit—older adults and lower-income populations—demonstrate lowest readiness.

The substantial gap between perceived need (31.5%) and actual help-seeking (18.3%) highlights intervention opportunities, yet age-based resistance suggests technology alone cannot bridge treatment gaps without addressing acceptance barriers. Results have implications for designing age-appropriate AI interventions, targeting implementation strategies for crisis populations, and understanding how socioeconomic factors shape perceptions of digital therapeutic alternatives during healthcare strain.

5.1: DATA QUALITY AND MEASUREMENT ERROR I

2:30 - 3:30 PM RH, SEMINAR 01

I MISBEHAVE, BUT ONLY ONCE IN A WHILE: HOW FACE-SAVING STRATEGIES CAN REDUCE SOCIALLY DESIRABLE RESPONDING IN ONLINE SURVEY RESEARCH

EMMA ZAAL, YFKE ONGENA, JOHN HOEKS
University of Groningen, Netherlands, The; e.l.zaal@rug.nl

RELEVANCE & RESEARCH QUESTION

Online research methods are imperative for collecting accurate data on behaviors and cognitions. However, when respondents answer sensitive questions, they tend to provide socially desirable answers - overreporting desirable norms (e.g., exercising), underreporting norm violations (e.g., prejudicial attitudes), or skipping items.

Reducing this social desirability bias (SDB) is essential for improving data validity. To reduce desirable responding, we conducted two survey experiments employing face-saving strategies: survey formulations that soften or justify norm violations. We addressed two questions: (1) To what extent does the effectiveness of face-saving answer options depend on the question format?; (2) Do respondents attend to face-saving preambles, and does engagement reduce desirable responses?

METHODS & DATA

Two online survey experiments were conducted among residents in a large Dutch city (N study 1 = 529; study 2 data collection currently ongoing). Study 1 examined face-saving strategies across four question formats: open; yes-no; apply-does not apply, and; very often-never. Experimental manipulations varied (a) the presence of a face-saving preamble and (b) the availability of face-saving answer options. Sensitive items were about volunteering, disturbances, safety, and poverty. Reading time of question introductions was recorded to assess respondents' attention.

Study 2 replicated question formats but introduced new items and different operationalizations of face-saving answer options and question introduction engagement. Similar reductions in SDB are expected for study 2.

RESULTS

In study 1, face-saving answer options consistently increased the selection of socially undesirable responses across all closed question formats and domains, indicating reduced bias. In contrast, preambles did not meaningfully shift response patterns, even when participants spent sufficient time reading them. Data collection for study 2 will be completed in December 2025.

ADDED VALUE

This study offers the first systematic comparison of face-saving strategies across multiple online question formats. By demonstrating that face-saving answer options reduce SDB - while preambles did not - study 1 clarified the mechanisms behind face-saving techniques and offered practical guidance for online survey design. The oral presentation will integrate findings from both experiments, highlighting how face-saving strategies can strengthen the validity of online surveys on sensitive topics and thereby advance online research methodologies.

'AND YET': THE EFFECTIVENESS OF PROBING QUESTIONS IN REDUCING ITEM NONRESPONSE TO FINANCIAL QUESTIONS

EVGENIIA SHMELEVA, DANIIL LEBEDEV, EKATERINA BRAGINA
NRU HSE, Russian Federation; eshmeleva@hse.ru

RELEVANCE & RESEARCH QUESTION

High rates of item nonresponse to questions on income and expenditures compromise data quality, leading to sampling bias and limiting the generalizability of findings. This study investigates the effectiveness of follow-up (probing) questions in reducing nonresponse to financial questions within a Russian context and identifies the profiles of non-respondents.

METHODS & DATA

Using data from the 6th wave of the HSE University's "Economic Behavior of Households" survey (N=6000), the analysis employs a two-stage approach combining the Random Forest method with the Boruta algorithm and logistic regressions.

RESULTS

Results indicate that probing questions successfully converted 36-41% of initial nonresponses into substantive answers, decreasing the overall nonresponse rate from 6-17% to 4-10%. Key predictors of nonresponse were found to be unawareness of household expenditures (a 3- to 5-fold increase in odds), poor health, lack of savings, and residence in small towns. The technique proved most effective for respondents with lower education levels, no savings, and those from small towns.

ADDED VALUE

The findings demonstrate that while probing is a valuable tool, its primary mechanism is reducing cognitive complexity rather than mitigating question sensitivity. Based on this evidence, the paper offers practical recommendations for improving the design of surveys that include financial questions.

HAVING ACES UP YOUR SLEEVE: DEVELOPING AND VALIDATING ATTENTION CHECKS EMBEDDED SUBTLY (ACES) TO IMPROVE IDENTIFICATION OF INATTENTIVE PARTICIPANTS

MAREK MUSZY SKI
Institute of Philosophy and Sociology of the Polish Academy of Sciences, Poland; marek.muszynski@ifispan.edu.pl

RELEVANCE & RESEARCH QUESTION

Careless or insufficient-effort responding (C/IER) is a major threat to data quality in online surveys. Existing detection approaches face substantial limitations: indicator-based methods (e.g., straightlining indices) require subjective threshold-setting, while model-based approaches rely on strong and often unrealistic assumptions and are difficult to implement in applied research.

Attention checks offer an objective-to-score and easy-to-use alternative. However, traditional attention checks, such as instructed-response items (Please select strongly agree) or bogus items (Orange is a fruit), suffer from limited validity, and high respondent reactivity (Daikeler et al., 2024; Gummer et al., 2021; Silber et al., 2022). This project addresses the lack of attention checks that would mimic ordinary questionnaire items to limit reactivity while reliably identifying inattentive respondents. The central research question is: How can we design attention checks that outperform existing approaches in validity and non-reactivity?

METHODS & DATA

First, 460 candidate ACES were developed across diverse domains (e.g., personality, technology, political attitudes), drawing on the concept of frequency/infrequency items (Kay & Saucier, 2023), and tested on 1498 respondents. Items were evaluated using distributional properties and socio-demographic invariance.

The selected ACES were validated in a between-subjects experiment (N = 880) comparing four conditions: (1) ACES, (2) IRIs, (3) bogus items, and (4) a no-check control group. The questionnaire included measures of perceived clarity and seriousness, as well as open-ended feedback on attention checks.The findings were replicated in another survey (N = 1113) targeting less experienced online panel members. All studies used non-probability quota samples (age, gender, education) from Polish online research panels.

RESULTS

ACES showed stronger associations with independent indicators of inattentiveness (e.g., response times, straightlining indices) and demonstrated higher classification accuracy (careless-not careless) than IRIs and bogus items. Respondents generally did not recognize ACES as attention checks, despite being highly familiar with traditional checks.

ADDED VALUE

This project delivers the first validated ACES set and provides empirical evidence that ACES improve detection of C/IER in comparison to other

attention checks while minimizing respondent reactivity. This enables using ACES in probability samples or interviewer-based surveys where traditional attention checks were not used due to their coarseness and potential for reactivity.

5.2: ONLINE PANELS I

02:30 - 03:30PM RH, SEMINAR 02

COMPARING PROBABILITY, OPT-IN, AND SYNTHETIC PANELS: A CASE STUDY FROM THE NETHERLANDS

CARSTEN BROICH1, NADICA STANKOVIKJ2
1Norstat, Netherlands, The; 2Lifepanel; carsten.broich@norstat.nl

RELEVANCE & RESEARCH QUESTION

The growth of nonprobability online panels and the emergence of synthetic survey respondents have created new opportunities and uncertainties for social measurement. While probability samples remain the reference standard, opt-in and synthetic data sources offer faster fieldwork and lower cost but may introduce unknown biases. This study asks: How comparable are attitudes measured across a probability panel, an opt-in panel, and a synthetic dataset?

METHODS & DATA

Three parallel surveys (≈500 completes each) were administered using identical instruments.
1. Probability sample: Lifepanel Netherlands, recruited via random-digit-dial/SMS.
2. Opt-in sample: Norstat's proprietary nonprobability online panel.
3. Synthetic sample: Personas generated algorithmically to approximate Dutch demographic structures.

The questionnaire measured perceptions of the national situation, attitudes toward elections, and interest in sports. Analyses include demographic comparison with CBS benchmarks, item nonresponse, variance structures, and inter-item correlations. Calibration experiments test post-stratification and raking on age, gender, education, and region to evaluate alignment potential.

RESULTS

The probability panel demonstrates the expected demographic balance and serves as the comparative baseline. The opt-in panel aligns closely with probability results after weighting, although unweighted data show overrepresentation of younger, higher-education groups. Attitudinal means are largely consistent across the two empirical samples, with modest discrepancies in political trust and evaluations of the national direction.

The synthetic dataset approximates mean values for several attitude items but exhibits compressed variance and weakened correlation patterns, indicating insufficient behavioral realism. Some synthetic respondents show inconsistent response structures not observed among human participants. Calibration improves demographic similarity but does not correct these structural limitations, suggesting that synthetic data are constrained more by model assumptions than by post-survey adjustment.

ADDED VALUE

This is one of the first empirical comparisons integrating probability, opt-in, and synthetic survey data within a single national framework. The study provides practical guidance on when synthetic respondents can complement empirical data (e.g., instrument testing) and where their limitations lie. It also clarifies the degree to which calibration can bridge differences between probability and nonprobability data but highlights fundamental constraints for synthetic datasets. The findings contribute to methodological best practices as synthetic data become increasingly visible in survey research.

OPTIMIZING PANEL CONSENT USING REPEATED REQUESTS WHILE EXPERIMENTALLY VARYING REQUEST PLACEMENT AND PANEL CONSENT INCENTIVES

SEBASTIAN HÜLLE, BENJAMIN BAISCH, MUSTAFA COBAN, MARCEL MÜLLER, LUKAS OLBRICH, STEFAN SCHWARZ
Institute for Employment Research (IAB), Germany; Sebastian.Huelle@iab.de

RELEVANCE & RESEARCH QUESTION

High panel consent rates are essential for reducing panel attrition and limiting the risk of panel consent bias in panel surveys. This study investigates how panel consent rates can be optimized by applying three innovative survey design features covering a repeated request for panel consent within the questionnaire while experimentally varying the placement of requests (beginning vs. end) and the incentive for panel consent.

METHODS & DATA

Analyses are based on the recruitment wave of the third cohort of the "Online Panel for Labour Market Research" (OPAL) and cover about 7.200 cases (about 12% are classified as partial interviews). In our design of repeated requests for panel consent within the questionnaire, respondents who do not provide consent at the first request are followed up with a second request to reconsider their decision. The survey experiment comprises four experimental groups that differ in the placement of the first panel consent request (beginning vs. end) and the incentive for panel consent at the first (0€ vs. 5€) and the second request (5€ vs. 10€).

RESULTS

Concerning the first request, an early request is more successful than a late placement: When offering no incentive, the panel consent rate is higher when asked at the beginning rather than asking at the end. When offering a 5€ incentive, the panel consent rate is higher when asking at

the beginning rather than at the end. Due to the second request, the cumulated panel consent rate increases by 5 to 10 percentage points across experimental groups. The highest cumulative panel consent rate after two requests has the design with the first request at the beginning while offering 5€ and offering 10€ at the second request at the questionnaire end.

ADDED VALUE

This paper provides evidence, that the highest panel consent rates are realized with a placement at the beginning of the questionnaire, which questions the traditional placement at the end. The panel consent rate can be significantly improved when implementing a second request within the questionnaire. Results show that the placement of a panel consent request can be more relevant than incentivizing.

YOU'VE GOT MAIL: DOES SENDING THANK-YOU POSTCARDS INCREASE RESPONSE IN A PROBABILITY-BASED ONLINE PANEL?

ANNE BALZ2, JULIAN DIEFENBACHER1, BARBARA FELDERER1, PHIL-ADRIAN KLOTZ3, JANNIS KÜCK3, ALINE MOHR3, TOBIAS RETTIG2, MARTIN SPINDLER4

1GESIS, Germany; 2University of Mannheim, Germany; 3Heinrich Heine Universität Düsseldorf, Germany; 4University of Hamburg, Germany; barbara.felderer@gesis.org

RELEVANCE & RESEARCH QUESTION

Survey nonresponse is one of the major challenges to survey data quality. While various treatments, e.g., monetary incentives, have been tested to increase response rates, the evidence regarding differential effects of the treatments for different population groups is rather thin. For example, it is known that incentives work well overall, but it is unclear whether a less costly form of appreciation (like a postcard) would achieve the same or a greater effect for certain population groups or whether some groups do not need any treatment at all.

METHODS & DATA

An experiment to increase survey response will be fielded in mid-December 2025 among the newly recruited participants of the German Internet Panel (GIP). After the first panel wave, before the second panel wave, panelists will randomly be assigned to four experimental groups:
1.) receiving a "Thank-you" postcard from the GIP team
2.) receiving a handwritten "Thank-you" postcard with the same text as
3.) receiving a postcard that states that they have been credited an extra of 5 Euro as a "Thank-you" for being in the panel
4.) control group
The postcards are not connected to the invitation to the second panel wave but are a general expression of appreciation for the panelist's participation in the study.

RESULTS

Results will be presented including:
1.) the overall effect of the treatment on nonresponse in wave 2
2.) possible interaction effects of personal characteristics and the treatment on nonresponse in wave 2. The personal characteristics include basic socio-demography, the BIG 5 personality traits, and the motivation to participate in the survey.

ADDED VALUE

Looking at heterogeneous effects of the treatment on nonresponse, the findings will inform survey practitioners on developing a targeted design to increase response rates and more specially, to increase response rates for specific population subgroups.

5.3: AI AND SOCIETY

02:30 - 03:30 PM RH, SEMINAR 04

INFORMATION-SEEKING IN THE AGE OF GENERATIVE AI: FACTORS THAT INFLUENCE THE BEHAVIOURAL INTENTION OF MEDIA STUDENTS TO USE CHATGPT

MOHAMMAD MAFIZUL ISLAM

Hochschule Darmstadt, Germany; mohammadmafizul.islam@stud.h-da.de

RELEVANCE & RESEARCH QUESTION

The widespread use of generative AI tools such as ChatGPT is reshaping how students search for, evaluate, and apply information. This shift is especially critical for media students, who will become future communicators, journalists, and researchers. However, little is known about the specific psychological, social, and technical factors that influence their adoption of generative AI as an academic tool. This study investigates the behavioural intention of media students to use ChatGPT for academic information-seeking. It focuses on understanding how constructs such as performance expectancy, trust, perceived humanness, and availability influence their decision-making.

METHODS & DATA

The study employed a concurrent mixed-methods design. Quantitative data were collected via an online survey (n = 103) distributed among media students at a German university of applied sciences. Constructs from the UTAUT2 model were adapted and extended to include trust, perceived humanness, and availability. The data were analysed using PLS-SEM. In parallel, six semi-structured interviews were conducted and analysed using concept-driven coding in MAXQDA 24. The integration of qualitative and quantitative data provides both generalisability and depth.

RESULTS

Quantitative findings reveal that performance expectancy and perceived humanness significantly predict behavioural intention to use ChatGPT, while effort expectancy, hedonic motivation, and trust (measured as privacy-related trust) do not. Availability was found to be a key practical driver but did not moderate use behaviour as expected.

Qualitative results provide a nuanced picture: students appreciate ChatGPT's speed and interactive design but consistently question its factual accuracy. Epistemic trust, not privacy, emerges as the decisive factor in usage patterns, leading to widespread verification practices. Human-like responses increase engagement but do not guarantee trust.

ADDED VALUE

This research offers one of the first empirically grounded models of generative AI adoption among media students using an extended UTAUT2 framework. It introduces and validates the novel constructs of perceived humanness and availability in technology adoption theory. Practically, the study provides educators, developers, and policymakers with insights into how students use and critically assess AI tools in academic contexts. It highlights the urgent need for AI literacy initiatives that emphasise epistemic caution alongside technical competence.

EXPLORING DIFFERENCES IN CHATGPT ADOPTION AND USAGE IN SPAIN: CONTRASTING SURVEY AND METERED DATA FINDINGS

MELANIE REVILLA, CARLOS OCHOA
RECSM-UPF, Spain; melanie.revilla@upf.edu

RELEVANCE & RESEARCH QUESTION

Metered data—a type of digital trace data collected through a tracking application (a "meter") installed by participants on their browsing devices (PCs, smartphones, or tablets), which records at least the URLs of visited web pages—have attracted growing interest due to their granularity and continuity. However, metered data are also subject to errors, which may differ from those found in survey data.

Thus, the main goal of this study is to examine the extent to which results obtained from an online survey differ from those collected via a meter, providing new empirical evidence on a timely and relevant topic with a significant longitudinal dimension: ChatGPT adoption and engagement. Specifically, we aim to advance existing research by investigating the factors associated with variation in the discrepancies between survey-based and metered data—that is, to assess when such differences are likely to be larger or smaller (e.g., we expect larger discrepancies for behaviors that occurred further in the past).

METHODS & DATA

To achieve this, we use data from the Netquest opt-in online panel (www.netquest.com) in Spain, comparing various indicators of ChatGPT adoption and engagement between February 2023 and April 2025 across two independent samples of 2,100 panellists each. One sample responded to an online survey, while the other provided metered data, which was used to construct comparable variables to those in the survey. Profiling variables were available for both samples, including socio-demographics, technology ownership at home, and devices used. We employ both descriptive analyses and regression models to compare the two samples and examine whether different factors (e.g., the time elapsed since an event) influence the observed differences.

RESULTS

Preliminary results suggest that differences between the survey and metered samples can sometimes be substantial, although the magnitude of these differences varies depending on the specific concept being measured. Different factors, such as the time elapsed since the event, also play a role. Final results are still forthcoming.

ADDED VALUE

This research advances understanding of how different data collection methods influence findings, particularly in longitudinal studies, while also offering new insights into ChatGPT's use in Spanish society.

WHAT DO WE TALK ABOUT WHEN WE TALK TO LLMS?

DENIS BONNAY1, JUHI KULSHRESTHA2, OLIVEIRA MARCOS3, ORKAN DOLAY4
1Université Paris Nanterre, France; 2Aalto University, Finland; 3Vrije Universiteit Amsterdam; 4Bilendi; denis.bonnay@parisnanterre.fr

RELEVANCE & RESEARCH QUESTION

Commercial LLMs are now part of everyday online life, but we still know strikingly little about what people actually do with them in practice. Here, we present empirical insights upon the content of messages that people exchange with chatbots, such as ChatGPT.

There still are few consensual results in the area. A very recent study by OpenAI (Chatterji et al., 2025) found surprising results that partly contradicted earlier research, demonstrating limited gender and education differences and very few "personal" interactions between users and ChatGPT. We use extensive GDPR-complaint data to address two questions:
RQ1: Can these recent findings regarding topic distribution and gender/education differences be replicated?
RQ2: How personal do conversations with LLMs get, and do they become more personal over time?

METHODS & DATA

Our data covers 5 months of conversation records from panel members in Brazil, Germany, Mexico and Spain who agreed to share their internet activity on laptop and/or mobile device (01.06.25–31.10.25; N = 45,200 participants). We collect both HTML streams and in-app contents for six major AI platforms: ChatGPT, Claude, Copilot, Gemini, Meta and Perplexity. We examine the context of the conversations using LLM-based classifiers. In particular, we reproduce the same prompts and data input as those used in Chatterji et al. (2025) for comparability (RQ1).

RESULTS

Available mid-January 2026 – we are aware this is rather late but we thought the data was exciting enough to try and share at GOR!

ADDED VALUE

Our data uniquely combines multiple AI system sources and reliable sociodemographics information, putting us in a good position to better understand and assess the divergences in previous studies which were limited in terms of LLM sources and user qualification. LLM-based classifiers enable fine-grained classification on high volumes of data, allowing for new approaches to RQ2 besides mere content classification. We believe the extent to which people get personal

with LLMs is underestimated when it is assessed merely through topic classification, since topics other than "self expression" and "relationships", such as practical guidance or multimedia topics, may also involve personal engagement with the systems.

6.1: DATA QUALITY AND MEASUREMENT ERROR II

04:00 - 05:00PM RH, SEMINAR 01

ASSESSING TRENDS IN TURNOUT BIAS IN SOCIAL SCIENCE SURVEYS: EVIDENCE FROM THE EUROPEAN SOCIAL SURVEY AND GERMAN SURVEY PROGRAMS

SASKIA BARTHOLOMÄUS1,2, ELIAS NAUMANN1, TOBIAS GUMMER1,2
1GESIS - Leibniz Institute for the Social Sciences, Germany; 2University of Mannheim; saskia.bartholomaeus@gesis.org

RELEVANCE & RESEARCH QUESTION

Social science surveys frequently overestimate voter turnout due to measurement and nonresponse errors, which undermine the validity of research on the causes and consequences of political disengagement. As turnout bias may differ across countries and over time, both cross-national and longitudinal comparisons are challenged.

Despite these concerns, there is no comprehensive longitudinal and cross-national comparison of turnout bias. Consequently, it remains unclear to what extent turnout bias is shaped by contextual factors or by survey design. To close this gap, we examine (1) the prevalence and development, (2) the contextual factors and (3) the survey design features associated with turnout bias in European social science surveys since 2000.

METHODS & DATA

We analyze data from the European Social Survey (ESS) and a unique data set of German Survey Programs (GSP) conducted between 2000 and 2023. First, we run separate OLS regression models using either absolute or relative turnout bias as the dependent variable and the year of data collection as the independent variable for each country in the ESS and for each survey program in the GSP. Second, we estimate fixed effects and mixed effects models using absolute or relative turnout bias in the ESS and GSP as the dependent variable. As independent variables, we include contextual factors. Third, we add variables capturing variations in survey design as independent variables.

RESULTS

Our findings reveal that the extent of turnout bias varies between countries and has increased over time, posing a significant challenge for both cross-national and longitudinal research. We identify several survey design features that could mitigate turnout bias which are in line with previous literature. Moreover, we discuss methodological innovations aimed at reducing turnout bias by targeting nonvoters before or during data collection through tailored survey designs.

ADDED VALUE

The persistence of measurement and nonresponse errors, along with the lack of validated turnout data, are a constraint for social science surveys. This study offers the first comprehensive longitudinal and cross-national comparison of turnout bias. Our results underscore the urgent need for methodological innovations to ensure the validity and comparability of data on political disengagement.

VALIDATING A 6-ITEM SCALE FOR MEASURING PERCEIVED RESPONSE BURDEN IN ESTABLISHMENT SURVEYS

ANDRÉ PIRRALHA1, JOE SAKSHAUG2
1IAB, Germany; 2IAB, Germany; University of Munich, Germany; andre.pirralha@iab.de

RELEVANCE & RESEARCH QUESTION

Response burden is a significant challenge in establishment surveys, threatening data quality and survey participation. However, the field lacks validated instruments to measure perceived response burden. This study addresses this gap by developing and validating a 6-item, binary (Yes/No) response burden scale.

Our central research question is whether this scale achieves measurement equivalence across different levels of objective burden (questionnaire length) and stability over time (longitudinally).

METHODS & DATA

We utilize data from an experiment embedded within three quarterly follow-up waves (2023-2024) of the IAB Job Vacancy Survey (IAB-JVS), a large-scale German establishment survey. Establishments (n=3,888) were randomly assigned to receive either a short (2-Page) or a longer (4-Page) follow-up questionnaire. We test for measurement invariance using multi-group Confirmatory Factor Analysis (CFA) adapted for binary indicators (WLSMV estimator), following Wu and Estabrook (2016). We assess both cross-sectional invariance (between experimental groups) and longitudinal invariance (across the three waves).

RESULTS

The scale demonstrates strong construct validity: respondents in the 4-page condition reported significantly higher perceived burden across all items (e.g., "High number of questions"). The analysis confirms full scalar invariance across the 2-page and 4-page experimental groups in each wave (e.g., Q1: $\Delta CFI < 0.01$). This indicates the scale measures the same latent construct equivalently regardless of objective burden.

Furthermore, the scale achieved full longitudinal scalar invariance across the three waves, demonstrating its temporal stability even as quarterly questionnaire content changed.

ADDED VALUE

This study provides practicioners with a validated, concise instrument to monitor perceived burden in establishment surveys, Based on the confirmation of cross-sectional and longitudinal scalar invariance, researchers can now confidently use this scale to track burden trends over time and accurately evaluate the impact of questionnaire design interventions. Hopefully, our work provides a reliable tool for comparative analysis, supporting efforts to improve data quality and respondent engagement.

THE EFFECTS OF PANEL CONDITIONING ON RESPONSE BEHAVIOR ACROSS DIFFERENT COHORTS: BIAS IN THE CORE DISCUSSION NETWORK

SANTIAGO ANDRÉS ALVAREZ TOBAR
University of Mannheim, Germany; santiago.alvarez.tobar@students.uni-mannheim.de

RESEARCH QUESTION

Panel conditioning: changes in response behavior caused by repeated survey participation, is a central methodological concern in online panels. Research has identified both positive and negative conditioning effects, but little is known about how these processes unfold in egocentric social network surveys, where name-generator items create opportunities for satisficing. In this study I ask: (1) How does repeated participation affect the likelihood of motivated misreporting in these filter questions? (2) To what extent is this relationship mediated by respondents' reported network size, that is, the number of alters named in the generator?

METHODS

I use data from the 12th wave of the online probability-based LISS panel, drawing on the Core Discussion Network module, which includes a name generator and follow-up questions on alter characteristics. Panel experience operates as the independent variable, and Motivated misreporting in filter questions as the dependent variable, while network size serves as the mediator. I estimate causal mediation models using Poisson and logistic regression with 5,000 bootstrap resamples and control for sociodemographic and survey-evaluation variables associated with panel attrition.

RESULTS

The results reveal two opposing mechanisms: a direct and indirect effect. Indirectly, respondents with greater survey experience report larger discussion networks, which increases their likelihood of misreporting in the filter questions, to avoid the tie-strength assessments. Directly, however, more experienced participants are less likely to engage in motivated misreporting when network size is held constant, suggesting reduced satisficing due to increased familiarity with online survey tasks. Because these pathways counteract each other, the total effect of panel experience on misreporting is small and statistically nonsignificant.

ADDED VALUE

This study demonstrates that panel conditioning in online surveys operates through simultaneous, opposing mechanisms that remain hidden to conventional response-quality diagnostics. The findings highlight the need to consider how question order, task burden, and instrument structure interact with respondent experience in modules involving name generators. By applying mediation analysis, the study provides a framework for detecting hidden behavioral mechanisms and offers practical guidance for improving the design and interpretation of longitudinal online surveys.

6.2: ONLINE PANELS II

04:00 - 05:00PM RH, SEMINAR 02

HANDLING THE RECRUITMENT PROCESS FOR A PROBABILITY ONLINE PANEL IN-HOUSE: INSIGHTS AND LESSONS FROM THE 2025 GERMAN INTERNET PANEL RECRUITMENT

TOBIAS RETTIG, CAROLIN BAHM, ANNE BALZ, BENJAMIN GRÖBE, STEFANIE SCHMIDT
University of Mannheim, Germany; tobias.rettig@uni-mannheim.de

RELEVANCE & RESEARCH QUESTION

In this contribution we will report on the 2025 recruitment of new respondents for the GIP. This includes a general overview of the recruitment procedure and results, unexpected obstacles we encountered along the way, lessons learned, and things to look out for when handling sample recruitment in-house. As we observe a general trend of moving more processes in-house among academic survey projects, often to reduce costs, this contribution will be of interest for a wide audience of survey practitioners who (plan to) handle their own recruitment of respondents.

METHODS & DATA

In September and October 2025, we sent out 7,000 invitations to prospective new respondents for the GIP. About 2/3 went to a random sample from the population registers of 135 municipalities drawn by GESIS and 1/3 to addresses sampled from a commercial database. We report on the process of handling the recruitment of a probability sample in-house, obstacles we encountered and possible solutions to these, and recruitment results such as response rates.

RESULTS

During the recruitment process, we encountered a number of obstacles, such as uncooperative municipalities, coordinating the printing and sending of invitation letters and reminders in-house,

print quality, handling the prepaid cash incentives, and a higher-than-expected number of invitations being returned by the postal service as undeliverable, presumably due to incorrect addresses, particularly in the commercial address database sample. While the fieldwork is still ongoing, preliminary results indicate that about 1,900 recruitment interviews have been started and about 1,100 to 1,200 new respondents will be recruited into the panel, while about 800 of the 7,000 invitations have been returned undeliverable.

ADDED VALUE

We provide a hands-on report with concrete advice for conducting a survey recruitment and avoiding potential obstacles and costly errors. In addition, we report recent information on the quality of obtained address data, response rates, and recruitment success for a probability online panel in the ever-changing survey climate in Germany, which will be of interest to any practitioners planning the recruitment of a probability sample.

METHODS TO MAXIMIZE THE PANEL CONSENT RATE IN THE RECRUITMENT WAVE OF A NEW WEB PANEL

HÜLLE SEBASTIAN1, LUKAS OLBRICH1, LUISA HAMMER1, KATIA GALLEGOS-TORRES1,2, YULIYA KOSYAKOVA1,3, JOSEPH W. SAKSHAUG1,4,5

1Institute for Employment Research (IAB), Germany; 2ZEW Mannheim; 3University of Bamberg; 4LMU Munich; 5University of Mannheim; lukas.olbrich@iab.de

RELEVANCE & RESEARCH QUESTION

Panel consent is the permission given by respondents to be re-contacted for future panel waves. The lower the panel consent rate, the larger the initial panel attrition and the higher the risk of panel consent bias, which threatens data quality.

This study investigates whether incentives for panel consent and repeated requests for panel consent can increase panel consent rates.

METHODS & DATA

We conducted an experiment in the first wave of the new online panel survey "International Mobility Panel of Migrants in Germany" (IMPa) on two innovative design features: (1) A repeated request for panel consent within the questionnaire: Respondents who do not provide consent at the first request are immediately followed up with a second request to reconsider their decision. (2) Incentivizing panel consent: Depending on the randomized experimental group, respondents are (not) offered an additional incentive (5€) for providing panel consent at the first and / or second request. In group 1, no incentive for panel consent is offered at neither request. In group 2, respondents are offered an incentive for providing panel consent at both requests. In group 3, panel non-consenters of the first request are offered an incentive for panel consent at the second request.

RESULTS

Our analyses (N≈41.000) show that 1) panel consent rates at the first request are higher if incentives are offered; 2) the second request significantly increases the cumulated panel consent rate; 3) the second-request effect is highest for group 3, where incentives are offered only at the second request; 4) the cumulative panel consent rate is highest for groups 2 and 3. Regarding the cost effectiveness, group 3 leads to panel consent rates as high as for group 2, while requiring costs per panel consenting respondent is close to group 1. We will also provide results on how the experimental design affects wave 2 response rates.

ADDED VALUE

This paper introduces and assesses two innovative survey design features to maximize panel consent rates. Furthermore, we analyse the costs associated with each design and derive recommendations for panel consent request designs.

BETWEEN THE WAVES: HOW ADDITIONAL STUDIES SHAPE PANEL PARTICIPATION TRAJECTORIES.

JOHANNES LEMCKE, TIM KUTTIG, ALAN NOVAES TUMP

Robert Koch-Institut, Germany; lemckej@rki.de

RELEVANCE & RESEARCH QUESTION

In recent years, probability-based mixed-mode panels have become a common tool in empirical research. Most panels rely on continuous, regular surveys that ask participants about core topics at fixed intervals (in a similar way to classic longitudinal panels). In addition, some of these panels offer additional or ad hoc studies. These studies allow internal or external researchers to conduct additional and in-depth surveys. The Robert Koch Institute (RKI) Panel 'Health in Germany', set up in 2024, operates exactly in line with this logic. However, for research purposes, the question arises as to what extent additional studies (at irregular intervals on different topics) influence the non-response of further (regular) surveys. In this presentation we will present first findings within the RKI Panel.

METHODS & DATA

The analysis draws on online survey response metrics from panel waves of the RKI Panel, complemented by information on invitations to and participation in an ad hoc survey conducted between the regular waves. A comparison is made between two randomly selected groups of panel members invited versus not invited to the analyzed ad hoc survey. Multivariate models control for sociodemographic characteristics to isolate potential effects attributable to the ad hoc survey.

RESULTS

Preliminary findings show that participants who were invited to the ad hoc survey exhibit marginally lower sub-group response rates for the following wave. However, controlling for sociodemographic characteristics in a multivariate model, we see no substantial effect regarding potential non-response for participants of the ad hoc survey. Thus, participants show stable response patterns overall. Data collection is still ongoing. Further results on more ad hoc surveys and following waves will be presented at the conference.

ADDED VALUE

The findings contribute to an emerging evidence on the effects of additional survey burden in mixed-mode probability panels. By investigating invitation effects, the study offers practical insights

for panel management, especially regarding contact frequency and respondent burden in newly established panels such as the RKI Panel.

6.3: MODELLING PEOPLE, INFORMING POLICY: NEW APPROACHES IN THE AI ERA

04:00 - 05:00PM RH, SEMINAR 03

THE LAST INTERVIEW – A CONCEPT TO CREATE A DIGITAL DOPPELGANGER

HOLGER LÜTTERS1, ANDRÉ WOLFF2, DAVID RANFTLER3
1HTW Berlin, Germany; 2Splendid Research; 3Xelper; holger.luetters@htw-berlin.de

RELEVANCE & RESEARCH QUESTION

This paper explores the potential of large-language-model–based "digital twins" as a novel instrument for empirical social research. The Last Interview investigates whether an AI agent, trained on an extended multi-stage quantitative survey complemented by a single qualitative session, can faithfully reproduce the attitudes and reasoning patterns of the original respondent. We address three central questions:

- Which types of data are most suitable for constructing a digital twin?
- How accurately can a data-driven synthetic twin mirror a participant's measurable opinions across multiple quantitative dimensions?
- What methodological and ethical implications arise when such an agent is used beyond its initial research scope?

METHODS & DATA

The study was conducted at HTW Berlin with volunteer students who consented to provide comprehensive personal data. Each participant completed sequential quantitative interview stages—covering demographics, personality traits, value orientations, decision preferences, and behavioral indicators—followed by one in-depth qualitative interview to contextualize the numeric responses. These combined data sets served as the training corpus for an advanced large-language-model architecture, producing individualized conversational agents ("Digital Doppelganger").

RESULTS

We present the first practical implementation of this concept. Digital Doppelganger are able to answer survey questions in a manner closely aligned with its human counterpart, demonstrating high agreement across quantitative measures. Notably, the Doppelganger occasionally produced reasoned extrapolations beyond the original questionnaire, suggesting emergent interpretative capacities despite being trained primarily on structured survey data.

ADDED VALUE

This project introduces a hybrid methodology that fuses extensive quantitative measurement with state-of-the-art generative modeling to create person-specific agents which can used for research purposes and beyond. It illustrates how Digital Doppelganger may support longitudinal analysis without repeated data collection while foregrounding critical issues of informed consent, data sovereignty, and identity representation. The Last Interview provides a replicable framework and empirical evidence for extending quantitative social science into the era of personalized AI. Although we consider the current approach to be in an early developmental stage with some dead end ideas, the project has already prompted reflections far beyond its initial market-research intent, revealing broader scientific and ethical implications than originally anticipated.

PERSONAS++ – SLICING & DICING THE RESULT SPACE OF A SURVEY

GEORG WITTENBURG, GUILLLAUME AIMETTI
Inspirient, Germany; georg.wittenburg@inspirient.com

RELEVANCE & RESEARCH QUESTION

Advances in computational methods, in particular recent advances in Artificial Intelligence (AI), have vastly reduced the manual effort required to derive results from any given survey dataset. This equally applies to structured, quantitative, interview-level data, but also to qualitative data. For the former, statistical and visualization methods may now be applied automatically; for the latter, sentiments, topics and codes are now easily extracted. As an industry, we're thus experiencing the commoditization of results. This gives rise to the new questions of how to efficiently work with this overly abundant set of results, how to focus on what matters, how to tell signal from noise.

METHODS & DATA

In this talk, we introduce the concept of the result space, which we define as the set of all possible results that can be derived from a given dataset of interview-level raw survey data by (automatically) applying current analytical methods. Based on our practical work across dozens of surveys over the past years, we propose alternatives to structuring this space, e.g., by variable, by methodology, or by significance of result; we look into alternatives for sorting and ranking results; and we discuss ways for measuring relations between results.

RESULTS

To strongly anchor the rather theoretical aspects of our work to everyday practical use, we illustrate the specific applicability of these concepts on real-world survey datasets, and on specific questions that we can now answer: What are the Top 3 things to know among all results relating a given sub-demographic? Of all the regression analyses, which ones stand out and why? Is there anything I overlooked in this summary that I wrote? We further demonstrate practicality by showcasing the system used to automatically derive the result spaces.

ADDED VALUE

Certain slices through the result space of a survey have already proven their practical value: Personas, for example, allow zeroing in on the particular wants, needs, and opinions of a sub-demographic of particular interest. With the toolkit presented in this talk, we generalize this concept, thereby providing the means to more deeply and more effectively investigate increasingly abundant survey results.

THE EU-ALMPO PROJECT: RETHINKING ALMPS THROUGH AI-DRIVEN ANALYSIS AND POLICY INNOVATION

SILVIA CASTELLAZZI, MARIA JULIANA CHARRY CAMARGO, SIRO CIARIMBOLI, FLAVIA PESCE
Institute for Social Research (IRS), Italy; mjcharrycamargo@irsonline.it

RELEVANCE & RESEARCH QUESTION

Digital transformation, Evidence-based policymaking
Amid rapid technological innovation and digital transformation, EU-ALMPO addresses the need for more agile, inclusive, and responsive to evolving skill mismatches labour market interventions. Anchoring policy design in data, machine-learning analytics, and stakeholder co-creation, the project objective is the creation of the EU Active Labour Market Policy Observatory – an AI-enabled digital hub that enhances the design, implementation, and evaluation of ALMPs across Member States.

By integrating advanced AI tools into a centralised digital platform, the Observatory supports evidence-based decisions and fosters knowledge exchange among policymakers, researchers, and labour-market actors.

METHODS & DATA

Analytical framework, Participatory validation, Comparative policy evaluation
Funded under Horizon Europe, EU-ALMPO has completed the WP1, which developed the analytical framework underpinning the Observatory. The framework analysed existing ALMP systems, identified structural gaps, and assessed the effectiveness of policies in addressing skills mismatches. It also provides a conceptual bridge - a translation layer - for its integration into the project's AI-supported system for policy-makers across the EU and beyond. Methodologically, it combines a literature review, a meta-evaluation of ALMPs, and participatory validation with stakeholders from several EU countries.

RESULTS

Skills mismatch analysis, Servuction model
Serving both diagnostic and prescriptive functions for skill mismatch analysis, the framework also deepens reflection on the implications of generative technologies for policy innovation. Inspired by the Servuction Model - bridging together front-end and back-end - it develops a user-oriented perspective and translate the knowledge and content developed in the project into actionable items that are valuable and useful for the policy-makers involved.

ADDED VALUE

AI-policy integration, Adaptive governance
EU-ALMPO represents a pioneering intersection between labour-market policy and AI. Its analytical framework and reflections around the ways to bridge policy design and AI tools, offers a data-driven, adaptive, and inclusive approach that strengthens Europe's capacity to respond to future labour-market transformations. While the focus of the project is related to policy making in the area of skills and labour market, the project represents an innovative ground to bridge policy and technology and is thus relevant for potentially other policy areas as well.

FRIDAY, 27/FEB/2026

ORAL PRESENTATIONS

7.1: AI AND QUALITATIVE RESEARCH

09:00 - 10:00AM RH, SEMINAR 01

AI-CONDUCTED USER RESEARCH: FROM WEEKS TO HOURS THROUGH AUTONOMOUS INTERVIEWING

BRUNO RECHT, ERNST MAXIMILIAN MÜNKER
Userflix, Germany; bruno@getuserflix.com

RELEVANCE & RESEARCH QUESTION

Qualitative user and market research remains essential for customer-centric product development, yet traditional moderated interviews are time-intensive and expensive. A single 1-hour interview requires approximately 5 hours of researcher time for recruitment, coordination, preparation, execution, and analysis. At industry rates of €500-750 per interview, comprehensive research projects become prohibitively expensive, especially for SMEs. This study examines: How can AI-powered autonomous interviewing maintain research quality while achieving dramatic improvements in speed, cost, and scale?

METHODS & DATA

We developed Userflix, an end-to-end AI platform for qualitative research automation utilizing large language models fine-tuned for research methodology. The system implements: (1) AI-guided study setup through conversational project briefing, (2) real-time audio-to-audio interviews with dynamic follow-up questions, (3) visual stimuli presentation, (4) automatic transcription and analysis, and (5) automated insight extraction with traceability to source interviews. Evaluation pilots (Q3-Q4 2025) are ongoing with Nielsen Norman Group (UX research methodology assessment), Innofact and Skopos (agency workflow integration), and IKEA (multilingual European research).

Partners are systematically comparing AI versus human interview quality, transcript depth, and participant experience. Early evaluation feedback demonstrates 95% time reduction (weeks to hours) and 90% cost reduction (€36/hour vs €500-750/interview). The AI successfully conducts multilingual interviews, generates contextual follow-up questions, and enables unprecedented scale (50-500 interviews vs traditional 8-12), allowing statistical pattern recognition in qualitative data. Nielsen Norman Group is assessing methodology soundness against established standards. Agency partners report high participant comfort, with some showing greater openness on sensitive topics with AI interviewers. The platform's 24/7 availability increased completion rates by 40% compared to scheduled interviews. Key advantages include parallel execution, elimination of interviewer bias, and consistent quality.

ADDED VALUE

This research demonstrates that AI can augment human researchers by handling routine execution, enabling focus on strategic interpretation. The "quantified qualitative" approach—conducting 50-500 interviews instead of 8-12—bridges qualitative depth with quantitative validation, addressing the longstanding trade-off between scale and depth.

For the online research community, this represents making comprehensive qualitative research accessible to broader audiences while elevating professional researchers to strategic roles. Evaluation results will provide evidence-based guidance for AI research tool adoption and quality standards.

AUGMENTING QUALITATIVE RESEARCH WITH AI: TOPIC MODELING WITH AGENTIC RAG

GERION SPIELBERGER1, FLORIAN ARTINGER2, JOCHEN REB3, RUDOLF KERSCHREITER1
1Freie Universität Berlin, Germany; 2Deutsche Hochschule, Germany; 3Lee Kong Chian School of Business, Singapore Management University, Singapore; gerion.spielberger@fu-berlin.de

RELEVANCE & RESEARCH QUESTION

Large Language Models (LLMs) increasingly shape qualitative and computational social science research, yet their use for text data analysis using topic modeling remains limited by low transparency, unstable outputs, and prompt sensitivity. Traditional approaches such as LDA often produce overlapping, generic topics, whereas LLM prompting lacks consistency and reproducibility.

We introduce Agentic Retrieval-Augmented Generation (Agentic RAG) - a multi-step, agent based LLM pipeline designed to improve efficiency, transparency, consistency, and theoretical alignment in qualitative text analysis. Our study addresses two research questions: (1) How does Agentic RAG perform compared to LDA and LLM prompting in terms of topic validity, granularity, and reliability across datasets? and (2) How can Agentic RAG be extended to enable theory advancement through "lens-based" retrieval?

METHODS & DATA

We benchmark Agentic RAG against LDA and LLM prompting using three heterogeneous datasets: (i) the 20 Newsgroups corpus (online communication), (ii) the VAXX Twitter/X dataset (data on vaccine hesitancy), and (iii) a qualitative interview corpus from an organizational research context. Agentic RAG is implemented as a model-agnostic, agent-based pipeline that orchestrates retrieval, data analysis, and topic generation. In our analysis, Agentic RAG was applied to produce topics using different GPT models (GPT-3.5, GPT-4o, GPT-5). We evaluate all methods using standardized metrics: topic validity, topic overlap, and inter-round semantic reliability, computed via cosine similarity measures that extend prior topic quality metrics.

Across datasets, Agentic RAG consistently yields high-validity topics with minimal redundancy compared to both LDA and LLM prompting. Whereas LDA and LLM prompting perform well only on specific datasets, Agentic RAG maintains performance across heterogeneous data architectures, while being more transparent and efficient. Based on these results, we derive a structured trade-off table that summarizes the strengths and limitations of all approaches, providing qualitative and computational scholars with clear guidance for selecting an appropriate text analysis method.

ADDED VALUE

Our findings demonstrate that Agentic RAG offers a scalable, transparent, and reproducible approach for qualitative text analysis. The method strengthens the rigor of LLM-based qualitative research by enabling more stable outputs, explicit retrieval reasoning, and broader options for assessing topic quality.

REINVENTING ONLINE QUALITATIVE METHODS: LESSONS FROM AN AI-ASSISTED STUDY ON PATHWAYS OUT OF LONELINESS

ANNA SCHNEIDER1, OLE WESTHOFF1, ORKAN DOLAY2
1Hochschule Trier, Germany; 2Bilendi&respondi; a.schneider@hochschule-trier.de

RELEVANCE & RESEARCH QUESTION

Loneliness has emerged as a growing social and public health concern that increasingly affects younger age groups. In response, the state government of North Rhine-Westphalia has initiated multiple initiatives and established a competence network to counteract loneliness. Against this backdrop, the present study examines the role of digital technologies in both the emergence and alleviation of loneliness. The research focuses on three interconnected key questions:

1) How do technological environments, ranging from face-to-face-communication tools to digital social platforms, shape experiences of social connectedness and emotional well-being, and to what extend may they contribute to or mitigate feelings of loneliness?
2) What role do so called third-places play in individuals' perceptions of social belonging and connectedness?
3) How are digital media used to build and maintain social relationships and under what conditions are digitally mediated interactions transferred into offline, contexts?

METHODS & DATA

The study employs a qualitative research design based on more than 150 participants, and was conducted using BARI, the qualitative AI developed by Bilendi. Participants engaged via WhatsApp or Facebook Messenger over roughly one week. BARI supported almost the entire research process, including project flow, moderation, data analysis and reporting. The AI-based moderation is methodologically notable as the absence of a human interviewer may foster greater openness when discussing sensitive topics such as loneliness, potentially reducing social desirability bias. This setup allowed collecting rich narrative data while simultaneously enabling an empirical assessment of the methodological implications of AI supported qualitative research.

Beyond substantive insights into perceptions and experiences of loneliness, the presentation will highlight methodological findings regarding the strengths, weaknesses and challenges of AI-assisted qualitative research. The integration of participant feedback and researcher reflection will be shown to play a central role in improving the AI´s performance and refining its methodological contribution to future research.

ADDED VALUE

The study provides dual added value: empirically, it offers new insights into how digital media and social spaces shape loneliness; methodologically, it delivers one of the first systematic assessments of AI-moderated qualitative fieldwork, demonstrating its potential and its limitations for scalable, participant-centered online research.

7.2: NEW INSIGHTS ON SATISFICING

09:00 - 10:00 AM RH, SEMINAR 02

IS 'DON'T KNOW' GOOD ENOUGH? MAXIMIZING VS SATISFICING DECISION-MAKING TENDENCY AS A PREDICTOR OF SURVEY SATISFICING

HANNAH SCHWÄRZEL
Technical University Darmstadt, Germany; schwaerzel@ifs.tu-darmstadt.de

RELEVANCE & RESEARCH QUESTION

Respondents who do not go through the question answer process optimally and instead exhibit satisficing behavior are a longstanding problem for survey researchers. A growing body of studies examines stable, time-invariant, predictors of survey satisficing behavior, such as

personality traits. These predictors introduce the possibility to measure the potential to satisfice before actual survey satisficing behavior occurs. This study wants to add to this research by introducing a potentially stable and reliable predictor of survey satisficing. Building on the notion that the question-answer process is a decision-making process and satisficing behavior is caused by low aspiration decisions, the effect of a decision-making tendency to maximize (in opposition to satisfice) on survey satisficing behavior is modelled.

METHODS & DATA

Data was gathered in October 2024 from 2,911 respondents within the Bilendi non-probability online access panel. Due to its short length, the generalizability of items and reduced dimensions of the scale, the modified maximizing scale by Lai (2010) is applied in this study to measure maximizing with its opposite pole satisficing. As dependent variables, "don't know" responding and midpoint choosing in four single choice questions was measured. To explain each of them, two multilevel models with questions on level 1 and persons on level 2 were calculated to grasp four questions in one model. Langer's (2020) extension of the McKelvey & Zavoina (1975) Pseudo R^2 for multilevel logistic regression models was obtained for the baseline models.

Respondents who score medium to high on the maximizing scale exhibit significantly less "don't know" responding and midpoint choosing than those scoring lower. However, the magnitude of the effect is small. The maximizing scale explains only a small amount of between-person variance in the examined satisficing behaviors.

ADDED VALUE

By affecting satisficing behavior in surveys, maximizing is not only a source of bias in survey data. The short and compact scale can be integrated in surveys to measure a respondent's potential to exhibit satisficing behavior before it occurs. With the help of LLMs, consecutive survey questions could be tailored to said potential.

MEASURING RESPONSE EFFORT AND SATISFICING WITH PARADATA: A PROCESS-BASED APPROACH IN THE CZECH GGS II

DANIEL DVORAK
Masaryk university, Czechia; 462749@mail.muni.cz

RELEVANCE & RESEARCH QUESTION

Assessing the quality of online survey data increasingly requires understanding how responses are produced rather than merely what is reported. Traditional outcome-based indicators such as nonresponse or internal consistency fail to capture the cognitive and motivational processes that shape respondents' behaviour. This study therefore introduces a process-oriented framework that uses paradata—digital traces of respondent actions—to measure engagement and data quality. The research asks: How can behavioural traces such as response times, edits, and navigation patterns be transformed into valid indicators of response effort, satisficing, inattentiveness, and cognitive load?

METHODS & DATA

The analysis uses event-level Blaise 5 paradata from the CAWI component of CZ-GGS II, encompassing approximately 3,500 completed interviews. Raw XML logs containing timestamps, navigation events, and answer edits were parsed into event-, block-, and respondent-level datasets. Five standardised indices (0–100 scales) were constructed: Satisficing, Cognitive Load, Respondent Effort, Inattentive Responding, and Behavioural Process Validation (BPV). Each combines timing, editing, and navigation measures into interpretable dimensions of process quality. Regression models (OLS/Logit) examine associations between these indices, demographic and technological characteristics, and classical quality outcomes such as item nonresponse and completion time.

RESULTS

The study builds directly on the author's earlier device-effects analysis of the Czech GGS, extending it from technological determinants to behavioural mechanisms of response generation. Preliminary findings confirm the distinctiveness and validity of the indices.

Higher Effort and BPV scores correlate with longer, more consistent response behaviour and lower item nonresponse, whereas higher Satisficing and Inattentiveness indicate faster, less engaged answering. Smartphone users show higher satisficing and reduced behavioural engagement compared to desktop users, while older and higher-educated respondents exhibit greater effort and more deliberate response trajectories. Final results will be completed by the end of 2025.

ADDED VALUE

By converting behavioural traces into quantifiable process indicators, the study redefines paradata as a methodological tool rather than a technical by-product. The resulting process-based framework offers a replicable blueprint for measuring engagement and response quality across survey waves, countries, and questionnaire designs implemented in Blaise 5. This approach enhances transparency, comparability, and methodological innovation within the GGP and broader demographic research.

A DATA-DRIVEN APPROACH FOR DETECTING SPEEDING BEHAVIOR IN ONLINE SURVEYS

ALAN NOVAES TUMP, TIM KUTTIG, JOHANNES LEMCKE
Robert Koch Institute, Germany; novaes-tumpa@rki.de

RELEVANCE & RESEARCH QUESTION

Online surveys provide a great opportunity for researchers to collect response times alongside the participants' answers. Extremely short response times—known as speeding—can indicate careless responding. Previous studies often identified speeding behavior using fixed cutoffs, for example those derived from average reading speeds reported in the literature. Although empirically motivated, these thresholds overlook other important cognitive demands of the questions and differences between respondents.

This study introduces a probabilistic mixture modeling approach to identify speeding behavior in the "Health in Germany" probability-based panel of the Robert Koch Institute (RKI). We validate this approach by comparing its classifications against established methods and by analyzing correlations with other indicators of data quality.

METHODS & DATA

Response times from the CAWI participants of the 2024 regular panel (Nff30.000) wave were analyzed using a shifted lognormal–uniform mixture model. As is standard practice in response-time analysis, the lognormal component represents regular, attention-based responses. The uniform component captures implausibly short response times ("speeding"). The shift parameter models the minimal realistic answering time. Hierarchical model specifications allow for variation across survey items and respondents. Fixed effects allow to estimate how model features such as speeding likelihood and minimal attention-based answering time vary as function of characteristics the participant (e.g., age) and the item (e.g., number of words).

RESULTS

Preliminary results show that the model accurately reproduces the empirical distribution of response times and allows for the calculation of speeding probabilities per response, participant and question. Speeding probabilities vary substantially across participants, suggesting that individual differences are the dominant source of speeding behavior, while item-level differences are smaller but still substantial.

Further analyses, to be completed before the conference, will examine how participant and item characteristics correlate with model parameters, how speeding behavior correlates with other indicators of data quality, and how the model performs in out-of-sample prediction.

ADDED VALUE

This study demonstrates the practical value of response-time modeling for improving data quality diagnostics in online panels. By quantifying speeding probabilities instead of applying fixed cutoffs, the method supports more data-driven cleaning and better understanding of response behavior.

7.3: DESIGNING INCLUSIVE AND ENGAGING SURVEYS

09:00 - 10:00AM RH, SEMINAR 03

ACCESSIBILITY AND INCLUSIVITY IN SELF-COMPLETION SURVEYS: AN EVIDENCE REVIEW

CRISTIAN DOMARCHI1, NHLANHLA NDEBELE2, OLGA MASLOVSKAYA1, PETER LYNN3, RORY FIZTGERALD2, RUXANDRA COMANARU2

1University of Southampton, United Kingdom; 2City St George's, University of London; 3Institute for Social and Economic Research, University of Essex, United Kingdom; C.Domarchi@soton.ac.uk

RELEVANCE & RESEARCH QUESTION

Survey research aims to understand social issues and inform effective public policy. For results to be accurate and equitable, surveys must inclusively represent diverse population sub-groups. Excluding these groups can lead to biased data and policies that perpetuate inequalities. Consequently, inclusivity is now a core principle for major statistical bodies in the UK like the UK Statistics Authority. This has led to a "respondent-centred design" approach, which argues that making surveys accessible for marginalised groups often benefits all respondents (Wilson and Dickinson 2021). However, achieving greater inclusivity involves practical trade-offs, as measures like targeted procedures, alternative response modes, or survey questionnaire translations or adaptations, can often be resource intensive. Evidence of inclusivity practices implemented as part of probability-based self-administered surveys is scarce, and research is required to determine best practice recommendations.

METHODS & DATA

This evidence review highlights measures that aim to increase participation for harder-to-survey population sub-groups in self-administered surveys, while maintaining the goal of obtaining high-quality, representative data. We focus on two main population subgroups: (1) individuals with disabilities and impairments and (2) individuals with literacy and/or language limitations.

RESULTS

This evidence review identifies general recommendations for recruitment practices to facilitate the inclusion of these frequently excluded sub-groups. It also highlights the cost trade-offs involved in implementing these methods, beyond the ethical imperative for inclusivity.

ADDED VALUE

This study addresses an under-researched area by providing evidence-based, practical recommendations for enhancing participation, accessibility and inclusivity in large-scale surveys.

EFFECTIVENESS OF THE KNOCK-TO-NUDGE APPROACH FOR ESTABLISHING CONTACT WITH RESPONDENTS: EVIDENCE FROM THE NATIONAL READERSHIP SURVEY (PAMCO) AND NATIONAL SURVEY FOR WALES (NSW) IN THE UK

OLGA MASLOVSKAYA, CRISTIAN DOMARCHI, PETER W.F. SMITH
University of Southampton, United Kingdom; om206@soton.ac.uk

RELEVANCE & RESEARCH QUESTION

Knock-to-nudge is an innovative method of household contact, first introduced during the COVID-19 pandemic when face-to-face interviewing was not possible. In this approach, interviewers visit households and encourage sampled units to participate in a survey through a remote survey mode (either web or telephone) at a later date. Interviewers also can collect contact information, such as telephone numbers or email addresses, or conduct within-household selection of individuals on the doorstep if required. This approach continued to be used post-pandemic in a number of surveys, but there remains a knowledge gap regarding its advantages and limitations. It is still unclear whether knock-to-nudge approach leads to improvements in sample composition and data quality.

METHODS & DATA

We analysed data from two UK surveys: the National Readership Survey (PAMCo) and the National Survey for Wales (NSW), each of which employed different versions of the knock-to-nudge approach. Our aim was to determine whether this method improves survey participation and sample composition, and to assess how incorporating participants recruited via knock-to-nudge impacts on data quality and responses to substantive questions. We investigate these effects using descriptive analyses, statistical tests, and logistic regression models.

RESULTS

Our findings demonstrate that knock-to-nudge is associated with: (1) a significant increase in response rates, (2) improved sample composition, (3) higher item non-response, and (4) significant differences in responses to substantive survey questions.

ADDED VALUE

This study contributes to the under-researched area of knock-to-nudge methods. The results indicate that, when carefully designed and implemented, this approach can enhance recruitment efforts and improve sample composition of the resulting samples in surveys. However, its viability as a universal solution for mixed-mode surveys depends on whether these methodological benefits outweigh the potential compromises in data quality and the additional implementation costs.

HOW DO RESPONDENTS EVALUATE A CHATBOT-LIKE SURVEY DESIGN? AN EXPERIMENTAL COMPARISON WITH A WEB SURVEY DESIGN

MAREK FUCHS, STELLA CZAK, ANKE METZLER
Technical University of Darmstadt, Germany; marek.fuchs@tu-darmstadt.de

RELEVANCE & RESEARCH QUESTION

Web surveys efficiently collect data on attitudes and behaviors, but often face challenges like satisficing behavior. The increasing prevalence of respondents using smartphones to answer surveys has brought about additional design challanges. The application of a messenger design as a web survey interface offers the opportunity to mitigate some of the drawbacks of a responsive web survey desing. Recent studies have demonstrated that a chatbot-like survey design may provide higher quality responses and greater engagement, albeit with longer response times. This study explores the respondents' evaluation of using a messenger interface in a web survey setting.

METHODS & DATA

In 2025, a sample of 2.123 members of a non-probability online access panel in Germany answered a survey on the topic of "vacation". The sample was cross-stratified by age and gender and limited to respondents aged 18-74. In a field-experiment employing a between-subjects designs respondents were randomly assigned to either a web survey design or to a chatbot design that mimics a messenger interface. In addition to survey duration, we assess the respondents' evaluation concerning user experience, perceived social presence, perceived flow, ease of use and general satisfaction.

RESULTS

Overall respondents in the chatbot condition were less satisfied with the survey and it took them longer to answer the questions. Also, they experienced lower levels of flow and ease of use. There was no significant difference in the user experience and the survey related social presence was lower only for respondents using a mobile device.

Older respondents, females and respondents with a higher education degree seem to evaluate the chatbot design more preferable than younger, male and respondents with lower levels of education. However, the preference for the web survey design is generally confirmed for all respondent groups irrespective of age, education and gender and also for respondents using a desktop or a mobile device.

ADDED VALUE

This study contributes to an assessment of using a messenger interface for the administration of survey questions. We discuss the results in light of the recent trend towards an application of a chatbot-like interface in AI supported surveys.

10.1: SMART SURVEYS AND INTERACTIVE SURVEY FEATURES

12:00 - 01:00PM RH, SEMINAR 01

ALEXA, START THE INTERVIEW! RESPONDENTS' EXPERIENCE WITH SMART SPEAKER INTERVIEWS COMPARED TO WEB SURVEYS

ANKE METZLER1, CEYDA ÇAVU O LU DEVECI2, MAREK FUCHS1
1Technical University of Darmstadt, Germany; 2Former Postdoc at Technical University of Darmstadt, Germany; metzler@ifs.tu-darmstadt.de

RELEVANCE & RESEARCH QUESTION

In recent decades, there has been a shift from interviewer-administered surveys to self-administered modes. While this transition improves efficiency and reduces costs, it also raises concerns about response burden. Text-based web surveys are often perceived as tedious and burdensome due to the lack of social interaction. The growing presence of the Internet of Things (IoT) indroduces new opportunities to address these limitations.

Voice assistants, in particular, enable an oral, conversational mode of data collection that may enhance social engagment while maintaining moderate costs.

However, little is known about how respondents perceive and experience interviews conducted through voice assistants. This study uses a smart speaker to administer survey questions and compares respondents' experiences in the smart speaker interview with those in a web survey.

METHODS & DATA

A laboratory experiment was conducted in summer 2025 with 245 participants recruited in the city center of Darmstadt. Using a within-subjects design, each participant completed both a web survey and a smart speaker interview containing the same questions. Immediately afterwards, each mode was evaluated separately. Paradata (e.g., response times, interruptions) and self-reported evaluation measures (e.g., flow, ease of use and user experience) were collected to assess participants' experiences.

RESULTS

Preliminary results indicate that respondents' general satisfaction is significantly lower in the smart speaker interview than in the web survey. Paradata analyses show that responses take longer and interruptions are more likely in the smart speaker interview. These differences explain part of the variation in general satisfaction and also affect self-reports on evaluation items. Longer response times and a higher likelihood of interruptions are negatively associated with perceived flow, ease of use and user experience. Multilevel analyses suggest that the lower satisfaction with the smart speaker interview can largely be explained by these mediating factors.

ADDED VALUE

At this point smart speaker interviews are still in its infancies. This study provides initial insights into respondents' perceptions of smart speaker interviews and identifies key aspects that require improvement to advance the development and successful implementation of smart speaker interviews in the future.

DO RESPONDENTS SHOW HIGHER ACTIVITY AND ENGAGEMENT IN APP-BASED DIARIES COMPARED TO WEB-BASED DIARIES? A CASE STUDY USING STATISTICS NETHERLANDS' HOUSEHOLD BUDGET DIARY.

DANIELLE REMMERSWAAL1, BELLA STRUMINSKAYA1, BARRY SCHOUTEN2
1Utrecht University, The Netherlands; 2Statistics Netherlands; d.m.remmerswaal@uu.nl

RELEVANCE & RESEARCH QUESTION

Smartphones offer opportunities for official statistics, promising improved user experience, reduced response burden, and higher data quality. We investigate whether respondents show higher activity and engagement in app-based diaries compared to traditional web-based diaries.

METHODS & DATA

We use Statistics Netherlands' Household Budget Survey (HBS) as a case study. The HBS is a diary survey conducted every five years to capture household expenditure on goods and services. In 2020, Statistics Netherlands conducted a 4-week web-based survey (N = ff3,000). In 2021, they conducted a 2-week app-based survey on a smaller sample (N =ff700). We compare participation and response behavior of respondents in the two modes. Among the indicators are the amount and spread of the reporting.

RESULTS

First results show that initial dropout is higher in the web diary. During the first two weeks, dropout is gradual (ff1% per day) and very similar across modes. % of registered respondents who submit at least one purchase and/or validate at least one day is higher for the app respondents. More results on objective burden (time spent in study) and reporting patterns will follow.

ADDED VALUE

App-based surveys are currently transitioning from the pilot phase to full implementation in panel surveys and official statistics. Our goal is to evaluate the expectation that app-based data collection enhances activity and user engagement in diary studies.

10.2: DATA DONATION

12:00 - 01:00PM RH, SEMINAR 02

DATA DONATIONS IN ONLINE PANELS: FACTORS INFLUENCING DONATION PROBABILITY

VANESSA LUX1, JESSICA DAIKELER1, LAURA BOESCHOTEN2, BELLA STRUMINSKAYA2

1GESIS - Leibniz Intitute for the Social Sciences, Germany; 2Utrecht University; vanessa.lux@gesis.org

RELEVANCE & RESEARCH QUESTION

Digital data donations in online panels have become an additional avenue for researchers to collect health and fitness tracking data, social media data, and web search and viewing histories from panel participants. Initial studies investigating the general willingness to donate data reported encouraging hypothetical consent rates, underscoring the potential of this method. Nevertheless, actual participation rates in donation studies generally remain low and systematically biased across demographic groups, raising concerns about representativeness and generalizability. In contrast, data donations within online panels have demonstrated consistent participation rates, indicating that online panels might offer favorable conditions for successful data donations. In our study, we explored the conditions and participant characteristics that influence data donation success in online panels compared to other data donation studies conducted within survey contexts.

METHODS & DATA

We conducted a systematic review of studies that implemented digital data donation protocols in the context of surveys using comprehensive searches across major academic databases (see preregistration: https://osf.io/nad2y). Studies meeting eligibility criteria were included in a quantitative meta-analysis assessing key indicators for donation success. For this exploratory part of the study presented here, we assessed whether indicators for donation success differed between donation studies in online panels and those in cross-sectional recruitments.

RESULTS

In general, donation probabilities varied largely across studies and were strongly influenced by study design and contextual factors, such as donation method, the requested data type, and the recruitment strategy. Studies with a lower mean age generally showed higher donation probabilities. However, online panels displayed more consistent donation probabilities compared to cross-sectional recruitments. Potential moderating factors, such as participants' familiarity with online data collections and trust in the institution receiving the donation, warrant further consideration.

ADDED VALUE

Our results offer meta-analytic insights into key indicators of data donation probabilities for studies conducted in online panels compared to cross-sectional recruitments.

MOTIVATIONS, PRIVACY, AND DATA TYPES: WHAT DRIVES WHATSAPP CHAT DATA DONATION IN A PROBABILITY SAMPLE?

JESSICA DAIKELER1, BARBARA FELDERER1, JULIAN KOHNE1, CARINA CORNESSE1, HENNING SILBER3, FLORIAN KEUSCH2

1GESIS, Germany; 2University of Mannheim, Germany; 3University of Michigan, USA; jessica.daikeler@gesis.org

RELEVANCE & RESEARCH QUESTION

Data donations are increasingly discussed as a valuable source for social science research. However, little probability-based evidence exists on what drives individuals' hypothetical willingness to donate personal communication data. We examine three components that could influence consent behavior: motivational framing (societal benefit, personal benefit, no benefit information), privacy (consent from chat partners required vs. consent not required), and the requested data type (aggregated metadata vs. full chat content).

METHODS & DATA

The study was fielded in the May 2025 wave of the German Internet Panel (GIP), a probability-based online panel of the German adult population. Respondents were randomly assigned to one of 18 conditions in a $3 \times 2 \times 2$ experimental design. The dependent variable was a binary measure of hypothetical willingness to donate WhatsApp chats, and we controlled for characteristics such as WhatsApp usage intensity and perceived data sensitivity.

RESULTS

Overall, 11% of respondents (approx. 350 individuals) reported willingness to donate their chat data. Willingness was higher when societal benefit was emphasized (13%) compared with personal benefit or no benefit information (each 10%). Requiring consent from chat partners showed no significant effect (12% vs. 10%). Participants were more willing to donate when full chat content was requested (15%) rather than aggregated metadata (8%). Multivariate analyses indicated that willingness increased among heavy WhatsApp users and decreased with higher perceived data sensitivity. No significant interaction effects across experimental factors were found.

ADDED VALUE

This study provides one of the first probability-based examinations of willingness to donate WhatsApp chat data—an especially sensitive and understudied data type. The results indicate that, in this hypothetical context, variation in content sensitivity influenced stated willingness less than expected.

These findings offer empirically grounded guidance for implementing future data donation infrastructures and highlight which informational cues and design choices may reduce barriers, increase trust, and support responsible integration of chat-based digital trace data into social research. For example, trying to limit the content sensitivity of a data donation request may not be as promising to increase data donation rates as simply emphasizing the research's societal benefit.

MOTIVATE AND PERSUADE: TESTING STRATEGIES TO INCREASE PARTICIPATION IN DATA DONATION STUDIES

FLORIAN KEUSCH1, FRIEDER RODEWALD1,2, VALERIE HASE3, SEBASTIAN PRECHSL2,4, FRAUKE KREUTER4, MARK TRAPPMANN2
1University of Mannheim, Germany; 2Institute for Employment Research, Germany; 3University of Klagenfurt, Austria; 4LMU Munich, Germany; f.keusch@uni-mannheim.de

RELEVANCE & RESEARCH QUESTION

Data protection regulations in the EU, Brasil, and California give users the right to access the information online platforms hold about them. Data donation studies capitalize on this requirement by asking web survey respondents to request and donate their data at the end of the survey. However, these studies often suffer from modest donation rates. In this study, we experimentally test whether using different motivational appeals in the data donation request and persuading respondents who initially decline the request increase participation.

METHODS & DATA

In summer of 2025, we conducted an experiment involving over 2,000 participants from a German online access panel. Panel members were asked at the end of a web survey to donate data from LinkedIn, Instagram, or YouTube. Here, we randomly varied the motivation appeal, either (1) emphasizing respondents' ability to quantify their online platform behavior (personal appeal), (2) emphasizing respondents' ability to learn what online platforms know about them (prosocial appeal), and (3) no appeal. Those who declined were asked for their reason for not donating, followed by a persuasion message that either provided (1) an argument tailored to their reason for declining or (2) a random argument not tailored to their reason for declining.

RESULTS

We found that the motivational appeals did not significantly affect respondents' willingness to donate data or their actual donation rates. However, the inclusion of motivational appeals had some effects on donation bias, for example, the prosocial appeal led to an overrepresentation of respondents with high trust in science among the donors. While providing a persuasion message led about 7% (60 out of 876) of decliners to reconsider their initial decision, only 3 additional respondents actually donated their data after receiving the persuasion message. The type of persuasion argument (tailored vs. random) did not affect participation.

ADDED VALUE

Our study shows that basic motivational appeals and persuasion messages have little effect on the participation behavior in data donation studies, but can influence donation bias. As our persuasion message was rather static, future studies investigating such adaptive survey designs could adopt a more tailored approach based on conversational agents.

10.3: SOCIAL MEDIA RECRUITMENT

12:00 - 01:00PM RH, SEMINAR 03

STATIC OR ANIMATED? HOW AD DESIGN SHAPES SURVEY RECRUITMENT

ANNA HEBEL
GESIS, Germany; Anna.Hebel@gesis.org

RELEVANCE & RESEARCH QUESTION

Social networking sites have become popular tools for recruiting survey respondents through targeted advertisements. Ad design is crucial, as it must capture users' attention within seconds.

While previous studies highlight the relevance of ad design in recruitment performance, sample composition, and data quality, they have almost exclusively focused on static images. However, static images represent only one possible design format, and the potential effects of animated visuals remain underexplored. This study extends prior research by systematically comparing the effects of static and animated ad images on two key aspects of survey recruitment: sample composition and response quality. It addresses the following research questions:

How are different visual elements (static vs. animated) related to sample composition? How are different visual elements (static vs. animated) related to response quality?

METHODS & DATA

Data stem from the recruitment campaign for the new online panel GP.dbd, which combines survey data with digital behavior data (e.g., web tracking and app data). The target population comprised adults living in Germany, and recruitment was conducted in 2023 via Facebook and Instagram. Four static images and their animated counterparts were tested. Differences between the two ad formats were examined using descriptive and comparative analyses focusing on respondent demographics and data quality indicators.

RESULTS

Findings show that the visual format of an ad image influences both who participates and how attentively respondents engage with the survey. Static images tend to attract women and highly educated individuals,

whereas animated ads appeal more to men, those with lower or middle education, and older respondents. Analyses of extreme response times, item nonresponse, and break-offs yielded mixed findings, suggesting that animation influences attentiveness in complex and context-dependent ways.

ADDED VALUE

These results highlight that ad design choices can subtly shape both the composition and engagement of recruited samples. Rather than favoring one format over the other, the findings suggest that static and animated ads serve different purposes in recruitment. A practical implication is to combine both formats strategically. Together, these insights provide nuanced, evidence-based guidance for researchers and practitioners seeking to optimize recruitment on social media platforms.

IS A VIDEO WORTH A THOUSAND PICTURES? THE EFFECT OF ADVERTISEMENT DESIGN ON SURVEY RECRUITMENT WITH SOCIAL MEDIA

ALEXANDER WENZ1, NICOLE SCHWITTER1,2
1University of Mannheim, Germany; 2University of Warwick, UK; a.wenz@uni-mannheim.de

RELEVANCE & RESEARCH QUESTION

Social media platforms, such as Facebook and Instagram, are increasingly used for survey recruitment, particularly for targeting hard-to-reach populations. Previous research has shown that the visual design of advertisements plays a key role in the effectiveness and costs of the recruitment and the data quality of the resulting samples (e.g., Donzowa et al., 2025; Höhne et al., 2025). Pictures are typically used in the advertisements to attract the social media users' attention and motivate them to click on the survey invitation. However, there is yet a limited understanding of how other visual formats, in particular videos, influence survey recruitment and data quality. In this study, we examine the effectiveness of pictures vs. short videos for the survey recruitment of young adults with social media. We hypothesize that videos are more engaging than pictures, resulting in a larger number of completed surveys at a lower cost, but do not expect differences in sample composition and response quality.

METHODS & DATA

We will conduct an online survey in December 2025 among young people aged 18-25 from Germany who traveled with Interrail in the last year. The survey includes questions about their travel behavior, attitudes towards the European Union, and European identity. Survey respondents are recruited through Meta (Facebook and Instagram). Using snowball sampling, respondents are also asked to forward the survey invitation to other people who fit the target criteria. In the ad campaign, we use pictures showing different contents, such as a person within a train, trains in front of different landscapes, and railway stations. We create corresponding videos with the AI software Midjourney by setting the respective pictures as the starting frame.

RESULTS

We compare the effects of the two advertisement formats on survey recruitment regarding their effectiveness, measured by the number of completed surveys and referrals through snowballing sampling, and their cost efficiency. Furthermore, we evaluate differences in sample balance across sociodemographics, in particular age and gender, and response quality, measured by completion time and item-nonresponse.

ADDED VALUE

The findings will be highly relevant for survey practitioners who plan to recruit respondents through social media.

SOCIAL MEDIA SAMPLING TO REACH MIGRANT POPULATIONS FOR MARKET AND OPINION RESEARCH

MARKUS WEISS, CLEMENS RATHE, ORKAN DOLAY
Bilendi; c.rathe@bilendi.com

RELEVANCE & RESEARCH QUESTION

Traditional survey methods consistently face critical challenges in achieving adequate coverage and response rates among migrant populations, leading to significant sampling bias in market and opinion research. Understanding these diverse groups is vital for both commercial and public sector decision-making. Research Question: Can targeted, non-probability sampling methods utilizing social media platforms (SMs) effectively recruit demographically diverse and representative samples of specific migrant populations in European countries, and how do the resulting data quality and efficiency metrics compare to surveys via online access panels?

METHODS & DATA

We've run 5 online surveys between January and October 2025 focusing on first- and second-generation migrants from Turkish and Arabic origin countries residing in France, Germany and Belgium. We used stratified recruitment campaigns across Meta platforms (Facebook/Instagram), utilizing the advertising API for targeting based on age, gender and language. The surveys were run in local languages + Arabic + Turkish.
The collected data was compared with online access panel data regarding metadata on response behaviour (survey speed, devices used, drop-outs€) as well as survey results including a deeper analysis and comparison of language impacts on survey results.

RESULTS

The Social Media Sampling (SMS) method demonstrated clear advantages in efficiency and reach. Crucially, SMs proved highly effective at accessing younger and lower-assimilation migrant cohorts who are severely underrepresented in standard frames.

The results showed, for example, that participants had significantly better knowledge of Turkish and Arabic compared to panel members. In addition, the proportion of Muslims was on average 20 percentage points higher than in the online access panels. Furthermore, we observed that many of the participants via social media were 1st generation migrants and an overrepresentation of people who recently moved to the respective country.

ADDED VALUE

This research provides an essential, validated framework for survey practitioners, demonstrating that social media can be leveraged as a rapid and reasonably cost-effective primary recruitment tool for hard-

to-reach, mobile populations. It offers a robust, tested procedure for mitigating sampling biases compared to online panels. Ultimately, this study promotes greater inclusivity and accuracy in survey results by ensuring the reliable representation of migrant voices.

11.1: SAMPLING AND WEIGHTING

02:00 - 03:00PM RH, SEMINAR 01

ENHANCING DATA ACCURACY IN KNOWLEDGEPANEL EUROPE: LEVERAGING DIFFERENT WEIGHTING TECHNIQUES AND ADJUSTMENT VARIABLES FOR OPTIMAL OUTCOMES

FEMKE DE KEULENAER, CRISTINA TUDOSE
Ipsos; femke.dekeulenaer@ipsos.com

RELEVANCE & RESEARCH QUESTION

Although online probability-based panels aim for accuracy, they can exhibit a left-leaning bias in public opinion research due to the overrepresentation of politically and civically engaged individuals. Researchers employ weighting techniques to correct sample imbalances relative to the population. This study aims to assess the extent to which diverse adjustment variables and weighting techniques can mitigate this left-leaning bias and enhance the accuracy of estimates from probability-based panels.

METHODS & DATA

This research examines samples from KnowledgePanel Europe (KP Europe), Ipsos' random probability online panel, which has been operational in the US since 1999 and the UK since 2020, with expansion across Europe since 2022. The evaluation focuses on the effectiveness of different weighting techniques, including raking, propensity score adjustments and generalized regression estimation (GRE). These methods are implemented using two sets of adjustment variables: basic demographics (age, sex, education and geographic region) and a more comprehensive set that includes demographics and variables linked to political attitudes and engagement. To gauge the relative advantages of various adjustment procedures and variables, each was evaluated based on its success in reducing bias for different benchmarks from high-quality, "gold-standard" surveys. These benchmarks cover a range of topics, like civic engagement, living situation and technology use. Besides biases, the variance or precision of estimates is crucial. The "margin of error" (MOE) describes the expected variance in survey estimates if repeated multiple times under identical circumstances. The MOE is calculated for estimates from all benchmark variables to see how different weighting procedures and variables affect variability.

RESULTS

Initial findings reveal variability in left-leaning bias in KP Europe samples. While various weighting methods effectively reduce bias and align results with population distributions, the choice of adjustment variables significantly affects the accuracy of the estimates. Additionally, incorporating political variables alongside basic demographics has a different impact on the MOE across KP Europe's countries.

ADDED VALUE

This study highlights the critical role of adjustment variables in improving the accuracy of estimates and provides valuable insights into the effectiveness of weighting techniques for reducing bias in political and public opinion research across diverse European contexts.

EXPLORING THE REPRESENTATIVENESS OF WEB-ONLY SURVEYS OF THE GENERAL POPULATION

PABLO CABRERA ÁLVAREZ1, ANNETTE JÄCKLE1, JAMIE C MOORE1, GABRIELE DURRANT2, JONATHAN BURTON1, PETER W F SMITH2
1Institute for Social and Economic Research, University of Essex; 2Department of Social Statistics and Demography, University of Southampton; pcabre@essex.ac.uk

RELEVANCE & RESEARCH QUESTION

The expansion of internet access in many countries raises the question of whether it is now feasible to conduct web-only surveys of the general population without compromising representativeness. This presentation will examine the representativeness of web-only surveys of the general population in the United Kingdom. The analysis addresses the following three research questions:

- RQ1: How have internet exclusion and intensity of internet use changed over time?
- RQ2: What are the characteristics of different types of internet users and non-users? How representative are these groups? How has this changed over time?
- RQ3: How does the representativeness of web respondents compare to the representativeness of different groups of internet users? How has this changed over time?

METHODS & DATA

We use data from the United Kingdom Household Longitudinal Study (UKHLS). The main survey (2009-2022), a probability-based sample of the UK household population, is used to explore the evolution of internet exclusion over time (RQ1 and RQ2). Moreover, we benefit from a mixed-mode experiment embedded in the Innovation Panel (2012-2023), a probability-based sample of the Great Britain household population. In the experiment, a random sample of households was assigned to a web-first and CAPI sequential design. This enables us to evaluate the representativeness of web respondents and compare it with that of internet users (RQ3).

We use coefficients of variation of the response propensities to estimate the representativeness of internet users and web respondents with regard to a set of auxiliary variables.

RESULTS

The results show a significant decrease in internet exclusion. Internet users are increasingly representative of the general population, although gaps still remain among older adults, those with lower education levels, and other disadvantaged groups. Web survey respondents have also become more representative of the general population over the past decade, but they remain less representative than internet users.

ADDED VALUE

The results offer valuable empirical evidence about the quality of web-only surveys in the past and present, which will assist survey practitioners in understanding the opportunities and risks of conducting web-only surveys now and in the near future.

11.2: ENSURING PARTICIPATION

02:00 - 03:00PM RH, SEMINAR 02

THE EFFECTS OF PUSH-TO-COMPLETE REMINDERS

KLARA PERSSON, SEBASTIAN LUNDMARK
The SOM Institute, University of Gothenburg, Sweden; klara.persson@gu.se

RELEVANCE & RESEARCH QUESTION

Survey researchers often allow respondents to fill out questionnaires both online and by paper-and-pencil forms in a mixed mode fashion. However, respondents who choose to complete questionnaires on paper tend to submit their questionnaires with fewer unanswered questions than respondents who choose to complete questionnaires online. Furthermore, many respondents who choose to fill out a questionnaire online do so without ever submitting their questionnaire. The aim of the present study is to evaluate whether sending digital push-to-complete reminders to respondents who have chosen to fill out a questionnaire online but have not yet submitted it are more likely to submit their questionnaires and submit them with fewer unanswered questions than similar respondents who do not get a digital push-to-complete reminder.

METHODS & DATA

The assessment will be made on a self-administered push-to-web mixed-mode survey (web and sequential paper-and-pencil questionnaire) distributed to a random sample of 9,000 individuals residing in Gothenburg, Sweden. In the experiment, potential respondents were randomly assigned to one of two groups:

One group was sent digital push-to-complete reminders a few days after they had started but not submitted the questionnaire online, whereas the other group did not receive such a reminder.

RESULTS

Data collection began in August 2025 and was completed in early January 2026. Preliminary results indicate that 14 percent of those who received a push-to-complete reminder clicked on the link within the reminder. Despite this, those who received a push-to-complete reminder were not more likely to complete the questionnaire than those not sent a push-to-complete reminder. Final results will be presented at the conference.

ADDED VALUE

The present study contributes to existing research on survey reminders by examining whether digital push-to-complete reminders can increase response rates and data quality in push-to-web mixed-mode questionnaires specifically, and online questionnaires generally.

THE EFFECT OF SURVEY BURDEN AND INTERVAL BETWEEN SURVEY WAVES ON PANEL PARTICIPATION: EXPERIMENTAL EVIDENCE FROM THE GLEN PANEL

MANUELA SCHMIDT1, CLAUDIA SCHMIEDEBERG2, CHRISTIANE BOZOYAN2, HENNING BEST1
1RPTU Kaiserslautern-Landau, Germany; 2LMU Munich; manuela.schmidt@rptu.de

RELEVANCE & RESEARCH QUESTION

Panel attrition remains a challenge in longitudinal survey research. Methodologists discuss several mechanisms affecting respondents' motivation to continuously participate in panel surveys. We focus on two of them: On the one hand, we test the idea that more frequent contact may increase respondents' engagement with the panel and hence reduce attrition. On the other hand, we address the question of how respondent burden in terms of questionnaire topic and complexity may impact future panel participation.

METHODS & DATA

We use data from the German Longitudinal Environmental Study (GLEN), a large-scale, nationwide randomly sampled panel on environmental topics launched in 2024. The experiments were implemented in an inter-wave survey in September 2025, followed by a panel wave in November 2025, used to measure participation effects.

EXPERIMENT 1

The first experiment investigates the effect of time interval between survey waves by not inviting a random 10% (N = 1,727) of the eligible sample (N = 16,772) to the inter-wave survey. We expect longer intervals between panel waves to reduce participation rates, as a higher survey frequency creates habituation and increases familiarity and engagement with the panel.

EXPERIMENT 2

In the second experiment, we examine the effect of complexity and thematic content of the questionnaire. In a first random split, 25% of participants received a long item battery on climate change skepticism, expected to increase burden due to its repetitive nature, while the rest answered a more diverse module on internet use. In a second random split, 90% were assigned a complex factorial survey experiment on CO2 pricing policies, expected to increase burden through topic complexity, while the rest answered questions on cultural participation. The assumed completion time was kept constant across groups, allowing us to isolate the effect of the questions on subsequent participation.

RESULTS

For both experiments, we analyze the effect on participation in the next panel wave. Data collection will be completed by the end of 2025. We will present results at the conference.

ADDED VALUE

We contribute to the literature on panel nonresponse and survey experience by providing experimental evidence from a nationwide randomly sampled panel.

ARE INTERVIEWER ADMINISTERED FOLLOW-UPS OF WEB NON-RESPONDENTS STILL NEEDED TO MAXIMISE DATA QUALITY? EVIDENCE FROM UNDERSTANDING SOCIETY: THE UK HOUSEHOLD LONGITUDINAL STUDY

GABRIELE DURRANT1, JAMIE C. MOORE2, PABLO CABRERA ÁLVAREZ2, ANNETTE JÄCKLE2, PETER, W.F. SMITH1, JONATHAN BURTON2

1University of Southampton, United Kingdom; 2University of Essex, United Kingdom; g.durrant@southampton.ac.uk

RELEVANCE & RESEARCH QUESTION

Many surveys have transitioned to online data collection. To minimize the risk of nonresponse bias, surveys often adopt a web-first mode with follow-up of nonrespondents via face-to-face or telephone interviewing. Evidence suggests such designs may reduce costs and may produce datasets of higher quality than web only designs.

However, with proportions of populations using the internet increasing markedly and people becoming less willing to welcome interviewers, in recent years the contributions of web with face-to-face or telephone modes to minimizing non-response biases that justify such a design may have changed.

This paper addresses this issue. The main research questions are: Do we still need to follow up web-non-respondents in a second mode to
RQ1: maximise response rates?
RQ2: maximise dataset representativeness?
RQ3: maximise response by under-represented hard-to-reach population subgroups?
RQ4: minimise non-response biases remaining after non-response weighting? and how has this changed over time?

METHODS & DATA

This study uses data from Understanding Society (UK Household Longitudinal Study, UKHLS). We focus on the Innovation Panel component of the study, in which a subset of sample members has been offered web interviews with face-to-face or telephone follow-ups of non-respondents.

For each survey wave, we use Coefficients of Variation of response propensities to quantify the representativeness of web only and web plus face-to-face or telephone respondents. In addition, we use the UKHLS main survey, which enables investigation of hard-to-reach population groups.

RESULTS

Key findings are: 1) follow-ups are still required to maximise response rates and dataset sizes, though impacts have declined; 2) the impact of follow-ups on representativeness has declined, with web and web plus face-to-face datasets not differing; 3) impacts of follow-ups on the under-representation of hard-to-reach population subgroups have become negligible; and 4) impacts of follow-ups on non-response biases remaining after non-response weighting, have similarly declined and are now negligible.

ADDED VALUE

We discuss the implications for survey practice. This paper is the first to investigate if follow-ups are still needed in web surveys in the UK context. If follow-ups are not needed any more, this could potentially have large cost-saving implications for survey agencies.

11.3: INFERENTIAL LEAP: FROM DIGITAL TRACE DATA TO MEASURING CONCEPTS

02:00 - 03:00PM RH, SEMINAR 03

EXPLORING TYPES OF MASCULINITY IN THE DISCOURSE OF FRINGE ONLINE COMMUNITIES

MO CHEN, TAIMOOR KHAN, DIMITAR DIMITROV, STEFAN DIETZE, ALEXIA KATSANIDOU
GESIS Leibniz Institute for the Social Science, Germany; mo.chen@gesis.org

RELEVANCE & RESEARCH QUESTION

4chan is an anonymous online forum known for ephemeral content and internet subcultures. This research examines how masculinities are constructed and contested in such spaces, where gender norms are negotiated through irony, confrontation, and subcultural slang. Combining theories of masculinity with computational text analysis, it bridges qualitative research and data-driven modeling of online discourse. It asks how different types of masculinities (hegemonic, hybrid, negotiating, caring) are constructed and circulated online and how theory-driven annotation and lexicon development can support computational identification.

METHODS & DATA

This study employs a mixed-method design integrating computational text analysis and qualitative annotation. It uses a comprehensive 2.5-year dataset of over 328M 4chan posts, encompassing textual data and metadata across all boards collected via a 4chan text collection tool. The platform's unfiltered nature allows gender and politics to frequently intersect across diverse contexts. The empirical pipeline includes preprocessing (cleansing, fixing misspelled, broken, or repeated words), identifying masculinity-related discourse through SentenceBert and extracting candidate terms for a theory-driven annotation schema grounded in gender studies and discourse theory. The schema operationalizes masculinity types through parameters such as dominance, emotionality, caregiving, and norm stance. Human annotators use a web-based interface (Doccano) supported by automated term suggestions and inter-annotator agreement metrics. The resulting annotations form the basis for a masculinity lexicon and subsequent computational modeling and clustering to explore discursive variations across online communities.

RESULTS

This study generates a dataset and lexicon capturing linguistic construction of different masculinity types. These resources support qualitative interpretation and large-scale computational modeling, revealing patterns in identity performance and negotiation of gender norms. Methodologically, it demonstrates how theory-driven annotation combined with embedding-based term extraction and automated quality control, providing a scalable framework for future studies of digital gender expression.

ADDED VALUE

Conceptually, this study illuminates how different types of masculinities are performed and negotiated in online communities, linking these patterns to broader social and political discourses. Methodologically, it demonstrates how theory-driven annotation and embedding-based lexicon development can create scalable tools for analyzing gendered language. The resulting dataset, annotation framework, and lexicon provide reusable resources for future research across platforms and contexts.

SOCIAL MEDIA AS A DATA COLLECTION TOOL AND ITS IMPACT ON BODY IMAGE PERCEPTION

MARGHERITA SILAN, SOFIA SPINELLI, GIULIA OSELIN
University of Padova, Italy; margherita.silan@unipd.it

RELEVANCE & RESEARCH QUESTION

This study aimed to evaluate the effectiveness of social media platforms, specifically Meta and TikTok, as data collection tools and to explore their influence on body image perception, with a focus on gender differences. The goal was to assess these platforms' efficacy in gathering data on sensitive topics like body image.

METHODS & DATA

Advertising campaigns were conducted on Meta and TikTok to recruit participants for an online survey about body image perception. Throughout the campaign, various images were used for the advertisements, comparing and testing the performance. Additionally, an experiment was conducted to assess the impact of an incentive as a motivation for engaging with the questionnaire. Data analysis focused on campaign performance metrics and survey responses, employing logistic regression and cluster analysis to examine the relationship between social media use and body satisfaction.

RESULTS

The campaigns demonstrated social media's potential to reach a large audience quickly, with TikTok generating high visibility and Meta platforms, especially Facebook, ensuring acceptable participation

rates. However, self-selection bias and sample representativeness emerged as concerns. The survey results revealed a clear link between social media use and body perception. Logistic regression showed that perceiving social media content as realistic and influential significantly increased the likelihood of body dissatisfaction, up to four times higher. Age emerged as a protective factor, while gender differences were less pronounced than expected. Cluster analysis identified three distinct user profiles based on perceived social media influence and content realism.

ADDED VALUE

This study provides insights into the effectiveness of social media as a data collection tool for sensitive topics and contributes to understanding the complex relationship between social media use and body image perception. It highlights the need for cautious interpretation of data collected through these platforms due to potential biases.

The findings suggest that perceived realism of social media content plays a crucial role in body dissatisfaction, offering valuable implications for promoting more conscious social media use and personal well-being. Future research directions are proposed to address limitations and expand on these findings.

12.1: SURVEY RECRUITMENT

03:15 - 04:15PM RH, SEMINAR 01

HOW RECRUITMENT CHANNELS SHAPE DATA QUALITY: EVIDENCE FROM A MULTI-SOURCE PANEL

FABIENNE KRAEMER, JESSICA DAIKELER, BARBARA FELDERER, BARBARA BINDER
GESIS Leibniz Institute for the Social Sciences, Germany; fabienne.kraemer@gesis.org

RELEVANCE & RESEARCH QUESTION

Declining response rates in traditional probability-based surveys have prompted researchers and survey practitioners to increasingly explore alternative recruitment strategies, such as via Social Networking Sites (SNS) and piggybacking (i.e., re-using respondents) from established surveys.

While these strategies offer faster and more cost-efficient access to respondents, they also raise questions about the quality of the resulting data.

SNS may threaten response quality through different motivational motives and an increased risk of satisficing. On the other hand, piggybacked samples may benefit from respondents' experience and commitment but could suffer from conditioning effects. This research provides a comparative assessment of response quality across different recruitment strategies.

METHODS & DATA

We examine response quality across different recruitment strategies using data from the newly established multi-source panel study GESIS Panel.dbd in Germany. The GESIS Panel.dbd combines several recruitment approaches, sampling respondents through a) SNS from Meta, b) piggybacking from the German General Social Survey (ALLBUS), the German Longitudinal Election Study (GLES), and the GESIS Panel. pop, and c) traditional probabilistic sampling from population registers.

To assess response quality, we analyze a range of indicators, such as item nonresponse, break-off rates, straightlining in item batteries, and response times. Our analyses control for sociodemographic composition across the different recruitment groups to disentangle effects of recruitment mode from compositional differences.

RESULTS

Results will be presented at the conference in February. Preliminary analyses reveal notable sociodemographic differences (e.g., in age and education) across recruitment groups, pointing to potential disparities in response behaviors and data quality.

ADDED VALUE

Altogether, this study provides one of the first comparative assessments of response quality across recruitment strategies increasingly used in survey practice. By controlling for differences in sample composition, we disentangle compositional from recruitment-driven effects, offering insights into how integrated recruitment designs affect data quality.

LOOKS GREAT, RESPONDS POORLY: LESSONS FROM TEN YEARS OF INVITATION LETTER EXPERIMENTS

JELMER DE GROOT, RYANNE FRANCOT
Statistics Netherlands, The Netherlands; jcdegroot@cbs.nl

RELEVANCE & RESEARCH QUESTION

What if the success of your survey depended on a single piece of paper? As survey researchers, we are now competing harder than ever for people's attention and time. With the rise of online, self-administered online surveys, interviewers play a smaller role in motivating participation. For these online studies in the Netherlands, households can only be invited by traditional mail – making the invitation letter our sole opportunity to connect. Its wording, tone, access options and layout can decide whether someone visits the link or ignores the request. Over the past decade, Statistics Netherlands has continuously refined its invitation strategy in search of the perfect letter for a diverse population.

METHODS & DATA

Our research combines qualitative pre-testing with large-scale field experiments. After many rounds of testing, a standard letter was designed that performs consistently well – until three new experiments challenged our assumptions. We examined (1) the effect of adding a QR code for easier access, (2) a shorter version of the letter, and (3)

the response to a refreshed, more visually appealing layout. Each experiment used a fresh, representative sample and a corresponding control group.

RESULTS

Adding a QR code had no significant effect (no code: 35.1% vs. QR code: 34.8%, n.s.).

However, the redesigned "fancy" letter, praised in qualitative pre-testing, led to a significant drop in response (35.4% 32.2%, p = 0.0003). The shortened letter, by contrast, increased participation (23,7% 26,8%, p = 0,000) – but also changed the way respondents answered the questions, resulting in fewer short trips being reported in the travel survey.

ADDED VALUE

Our findings reveal how intuitive design choices can have unintended consequences. While respondents claim to prefer modern, appealing letters, with QR-codes, actual behavior tells a different story and may subtly affect measurement. This presentation offers ten years of lessons – and a few humbling surprises – from the ongoing search for the holy grail of survey invitations.

12.2: PUSH TO WEB AND MIXED MODE SURVEYS

03:15 - 04:15PM RH, SEMINAR 02

INTRODUCING WEB IN A TELEPHONE EMPLOYEE SURVEY AND ITS IMPACTS ON SELECTION BIAS AND COSTS

JOSEPH W. SAKSHAUG1,2, JAN MACKEBEN1
1IAB; 2LMU-Munich; joesaks@umich.edu

RELEVANCE & RESEARCH QUESTION

Telephone surveys have historically been a popular form of data collection in labor market research and continue to be used to this day. Yet, telephone surveys are confronted with many challenges, including imperfect coverage of the target population, low response rates, risk of nonresponse bias, and rising data collection costs.

To address these challenges, many telephone surveys have shifted to online and mixed-mode data collection to reduce costs and minimize the risk of coverage and nonresponse biases. However, empirical evaluations of the intended effects of introducing online and mixed-mode data collection in ongoing telephone surveys are lacking.

METHODS & DATA

We address this research gap by analyzing a telephone employee survey in Germany, the Linked Personnel Panel (LPP), which experimentally introduced a sequential web-to-telephone mixed-mode design in the refreshment samples of the 4th and 5th waves of the panel. By utilizing administrative data available for the sampled individuals with and without known telephone numbers, we estimate the before-and-after effects of introducing the web mode on coverage and nonresponse rates and biases.

RESULTS

We show that the LPP was affected by known telephone number coverage bias for various employee subgroups prior to introducing the web mode, though many of these biases were partially offset by nonresponse bias. Introducing the web-to-telephone design improved the response rate but increased total selection bias, on average, compared to the standard telephone single-mode design. This result was driven by larger nonresponse bias in the web-to-telephone design and partial offsetting of coverage and nonresponse biases in the telephone single-mode design. Significant cost savings (up to 50% per respondent) were evident in the web-to-telephone design.

ADDED VALUE

Using a unique experimental design we showed the potential for known telephone number coverage bias in telephone surveys. However, while introducing the web mode eliminates this coverage error, there is potential for other error trade-offs. For practitioners, this underscores the importance of carefully weighing the potential trade-offs between costs and multiple sources of error when designing a specific study.

EXAMINING THE INFLUENCE OF RESPONDENTS' INTERNET-RELATED CHARACTERISTICS ON MODE CHOICE (PAPER VS. WEB MODE) IN A PROBABILISTIC MIXED-MODE PANEL WITH PUSH-TO-WEB DESIGN

LENA REMBSER
GESIS – Leibniz Institute for the Social Sciences, Germany; lena.rembser@gesis.org

RELEVANCE & RESEARCH QUESTION

The web survey mode offers a high degree of flexibility while also being highly cost-efficient. However, in the context of web surveys, potential coverage issues are repeatedly raised as a problem, as target respondents without internet access cannot participate in an internet-based survey. In addition, there are respondents with internet access for whom participating online would be possible from a formal perspective, but who do not want to participate via the internet. I want to know: Are the reasons for abstaining from participating online-based also influenced by internet-related characteristics, apart from simply having or not having internet access?

METHODS & DATA

I use data from the GESIS Panel.pop Population Sample, a probability-based self-administered mixed-mode panel of the German general population, surveyed via web and paper mode (push-to-web design with paper mode as alternative mode). I perform logistic regressions with mode selection as the dependent variable and various internet-related characteristics as independent variables (frequency of internet use, internet skills, variety of internet use, and number of internet-enabled devices). I put particular emphasis on the key challenge of identifying causal effect directions between internet-related characteristics to construct my models with appropriate control variables. For all regressors, I use not only the quasi-metric scale level frequently used in the literature, but I also examine different threshold levels through varying thresholding.

RESULTS

Data analysis is ongoing. I am going to have results in February 2026.

ADDED VALUE

With my analyses, I want to shed light on (internet-related characteristics of) those individuals who are not reached by online-only. Knowledge about those not reached by online-only also contributes to evaluating a target group-oriented use of paper questionnaires when conducting surveys.

DATA QUALITY IN PUSH-TO-WEB LONGITUDINAL SURVEYS: EVIDENCE FROM ELSA'S TRANSITION TO SEQUENTIAL MIXED-MODE DESIGN

LINA LLOYD, REBECCA LIGHT, MARIA TSANTANI
National Centre for Social Research, United Kingdom; rebecca.light@natcen.ac.uk

RELEVANCE & RESEARCH QUESTION

Longitudinal studies are transitioning from face-to-face to push-to-web designs for cost-efficiency and timeliness. However, for ageing populations, concerns persist about mode effects on data quality and measurement equivalence. The English Longitudinal Study of Ageing (ELSA) Wave 12 dress rehearsal provides empirical evidence addressing: Does push-to-web compromise data quality in aging surveys? What mode effects emerge comparing web-first to traditional face-to-face approaches?

METHODS & DATA

Wave 12 implemented sequential web-first then CAPI design for 375 individuals (June-August 2025). Web-first participants (n=260 households) received online invitations with CAPI follow-up for non-responders after 5 weeks. We assessed mode effects across multiple dimensions: item non-response, break-off patterns, scale reliability, occupation/industry coding quality (CASCOT confidence scores), and measurement consistency versus Wave 11 single-mode CAPI. Respondent feedback (n=31) and interviewer reports provided qualitative insights.

RESULTS

Push-to-web showed mixed outcomes. Individual web completion reached 33.6%, but household completion was only 21.5%, requiring 78% of web-first households to receive CAPI follow-up. Web median duration (87.6 minutes) exceeded estimates (60 minutes), with break-offs concentrated at household grid and financial modules. However, web mode demonstrated improved data quality: item non-response was lower compared to CAPI Wave 11 CAPI; transformed web questions improved the data quality for job and accupation responses; scale reliabilities remained excellent (CASP-19 =0.90). Mode preference split: 50.7% respondents preferred web for flexibility and self-pacing; 18.8% valued face-to-face interaction. Telephone-administered cognitive assessments proved challenging, with hearing difficulties affecting completion quality.

ADDED VALUE

Evidence from ELSA study provides critical evidence for longitudinal surveys implementing push-to-web among older populations. Key contributions:

(1) household-level completion metrics are essential—individual response rates mask significant household non-completion, which has big implications on fielwork costs; (2) structured web design reduces measurement error in complex variables despite completion burdens; (3) sequential mixed-mode designs must balance mode-specific strengths – web excels at structured data while face-to-face remains valuable for cognitively demanding tasks and panel engagement; (4) realistic duration expectations and improved re-entry protocols are crucial. Findings demonstrate push-to-web viability for ageing populations when supported by tailored strategies and hybrid approaches for mode-sensitive content.

12.3: METHODS, TOOLS, AND FRAMEWORKS – A BIRD'S VIEW ON DATA COLLECTION

03:15 - 04:15PM RH, SEMINAR 03

THE METHODS HUB: INTEGRATING TOOLS, TUTORIALS, AND ENVIRONMENTS FOR TRANSPARENT ONLINE RESEARCH

ARNIM BLEIER, JOHANNES KIESEL, CHRISTINA VIEHMANN, CHUNG-HONG CHAN, PO-CHUN CHANG, RANIERE GAIA COSTA DA SILVA, MUHAMMAD TAIMOOR KHAN, STEPHAN LINZBACH, FAKHRI MOMENI, FELIX VICTOR MÜNCH, RAN YU, CLAUDIA WAGNER, STEFAN DIETZE
GESIS - Leibniz Institute for the Social Sciences, Germany; arnim.bleier@gesis.org

RELEVANCE & RESEARCH QUESTION
As digital communication increasingly unfolds on online platforms, behavioral data have become central to understanding media exposure, polarization, and social interaction. Yet computational approaches necessary to analyze such data often remain inaccessible to many communication and social science researchers who lack extensive programming expertise or institutional resources. As a result, many research-driven tools remain scattered across personal repositories, supplementary materials, or project websites, reducing their visibility, reusability, and long-term sustainability. This presentation addresses the question of how a community-driven infrastructure can lower entry barriers for computational methods and support transparent, reproducible online research.

METHODS & DATA
The Methods Hub is designed as an open platform that curates computational resources relevant to social science research. It integrates three core components: (1) open-source tools ranging from lightweight scripts to fully developed software packages, (2) tutorials explaining both general principles of reproducible computational workflows and concrete methodological applications, and (3) containerized interactive coding environments that can be executed directly in the browser without local installation.

All contributions follow open licensing and reproducibility standards and are reviewed accordingly. The platform architecture supports interoperability with complementary infrastructures (e.g., KODAQS) to facilitate cross-linking between datasets, tools, and training materials. The development process combines community submissions, expert curation, and iterative user testing to ensure methodological relevance and usability.

RESULTS
Preliminary implementation demonstrates that the platform successfully bridges gaps between computational tooling and social science workflows. Initial contributions include tools for digital trace data collection, automated preprocessing pipelines, validation and reliability routines, and visualization templates. Tutorials and browser-based execution environments have proven effective in enabling researchers to test methods without configuring complex software environments. User feedback from pilot workshops indicates substantial reductions in setup time, increased willingness to experiment with computational approaches, and improved understanding of reproducible research practices.

ADDED VALUE
The platform lowers entry barriers to behavioral data analysis, strengthens methodological knowledge transfer, and promotes long-term visibility and reuse of tools otherwise confined to fragmented project repositories. Through openness, interoperability, and executable documentation, the Methods Hub contributes to building a robust ecosystem for computational communication science.

LET'S TALK ABOUT LIMITATIONS: DATA QUALITY REPORTING PRACTICES IN QUANTITATIVE SOCIAL SCIENCE RESEARCH

FIONA DRAXLER1, JESSICA DAIKELER2
1University of Mannheim, Germany; 2GESIS – Leibniz Institute for the Social Sciences; fiona.draxler@uni-mannheim.de

RELEVANCE & RESEARCH QUESTION
Clearly communicating data quality limitations is essential for transparent research. Data quality frameworks and reporting guidelines support researchers in identifying and documenting potential data quality concerns, but it is unclear how well this translates to reporting practices. In this project, we analyze reports of data quality limitations in substantive social science publications. Thus, we provide insights into typical limitations that reoccur but also highlight underrepresented areas where researchers might require additional guidance.

METHODS & DATA
We analyze the "Limitations" sections and limitation-related paragraphs in "Discussion" sections of substantive survey-based research published in the journals including American Sociological Review and Public Opinion Quarterly. We use a large language model to extract data-

quality-related aspects of these sections and paragraphs and assign them to the measurement and representation sides as defined in Total Data Quality error frameworks. We then cluster the excerpts into themes and compare the themes to components of the error frameworks. Based on this, we discuss which data quality dimensions are commonly and rarely mentioned, and what possible reasons for these differences may be. Through comparisons with reporting guidelines (e.g., AAPOR transparency initiative, datasheets for datasets), we highlight areas where researchers might require additional support. We also analyze areas where current guidelines might be adapted to better represent researchers' needs in reporting.

RESULTS

Initial findings show a prevalence of discussions on measurement validity and on coverage of the target population in contrast to only few mentions of limitations related to data processing. We also find that limitations are often communicated implicitly, adding a challenge for readers from other disciplines. For example, briefly mentioning the concrete implications of using an "online non-probability sample" would increase interdisciplinary validity assessments.

ADDED VALUE

We contribute to the transparent and well-structured communication of data quality as a crucial step for validating research by providing an overview of current reporting practices and directions for improved reporting.

QUALITATIVE RESEARCH IN DIGITAL CONTEXTS: A SYSTEMATIC REVIEW OF ONLINE DATA COLLECTION PRACTICES

MARLENE SCHUSTER, MELANIE GRUBER
FH Wiener Neustadt GmbH City Campus, Austria; melanie.gruber@fhwn.ac.at

RELEVANCE & RESEARCH QUESTION

Synchronous online qualitative data collection as a research method, such as interviews or focus groups via digital video conferencing platforms, has been experiencing a major upswing. Consequently, researchers are challenged to consider if and how the traditional qualitative research process from preparation to follow-up, could be adapted for online contexts, especially concerning quality, distance versus proximity and privacy.

Existing research, however, is usually limited to advantages and disadvantages. We thus ask which strategies and quality practices qualitative researchers use to collect qualitative research in an online context and how these can be interpreted through existing methodological and quality frameworks, situating our project within the topic "survey innovations".

METHODS & DATA

We conducted a systematic literature review (Tranfield et al. 2003) of academic journal articles reporting experiences and reflections of qualitative online data collection from 2000 to 2024. A literature search was carried out in the databases Springer Link, Science Direct and Emerald Insight, using among others, the following terms: "digital OR virtual OR online data collection" AND "qualitative research OR method". Following a three-stage selection process, 44 articles were selected and systematically coded in MaxQDA, combining a deductive framework (according to "before, during and after the survey") with an inductive coding process.

RESULTS

The findings indicate that successful qualitative online research relies on well-coordinated teams with clearly distributed roles, careful platform selection, and transparent communication with participants. Core strategies include developing digital and interpersonal competencies, fostering trust and interactional dynamics, and ensuring secure and ethical data handling. Rather than adhering to fixed standards, researchers maintain quality through adaptive, context-sensitive, and collaboratively negotiated practices.

ADDED VALUE

The online context will continue to play an increasingly important role in qualitative social research. We fulfil the need for practical guidance on conducting qualitative research projects online while maintaining quality standards. Furthermore, we relate research practices to existing debates on research quality. In doing so, it not only offers practical guidance but also theoretical connections for developing a reflexive methodology of digital social research.

THANKS

SESSION CHAIRS

THE GOR ORGANIZERS WOULD LIKE TO EXPRESS THEIR SINCERE GRATITUDE TO ALL SESSION CHAIRS WHO CONTRIBUTED TO AND ENSURED THE SMOOTH RUNNING OF THE SESSIONS.

GEORG AHNERT
THIJS CARRIERE
FIONA DRAXLER
LISA DUST
BARBARA FELDERER
ANNA HEBEL
OTTO HELLWIG
JOHANNA HOELZL
FRANK HEUBLEIN
FLORIAN KEUSCH
ARMIN KÜCHLER
PETER LUGTIG
OLGA MASLOVSKAYA
CHARLOTTE MUELLER
SOPHIA PIESCH
DAVID RANFTLER
TOBIAS RETTIG
DANIELLE REMMERSWAAL
YANNICK RIEDER
FRIEDER RODEWALD RODEWALD
CAMILLA SALVATORE
MANUELA SCHMIDT
BELLA STRUMINSKAYA
OLIVER TABINO
WAI TAK TUNG
ALEX WENZ
OLAF WENZEL
GEORG WITTENBERG
CAGLA YILDIZ
ZAZA ZINDEL