Listing of Sessions and Presentations by SOEP/DIW Berlin Staff Members

11th ESRA Conference
July 14th-18th, 2025
Utrecht University
Utrecht, The Netherlands

SOEP The Socio-Economic Panel

100 YEARS DIW BERLIN

Leibniz Leibniz-Gemeinschaft

TUESDAY, July 15th

Interviewers across the Survey Life Cycle

Session Organisers Dr Mariel Leonard (DIW-Berlin)
Dr Zachary Smith (National Center for Health Statistics (NCHS))

Time and Room 9:00 - 10:30, Ruppert 011

Interviewers are central to survey operations. From qualitative question design and evaluation – cognitive interviewers, focus group moderators, and even expert reviewers – to quantitative survey administration in the field. A large body of literature has identified various ways the identity, behavior, and disposition of interviewers influence the quality of data collected. And, a growing consensus is developing that in both qualitative and quantitative aspects, interviewers should be understood not merely as mindless, faceless data collection machines, but as researchers that contribute to the research process. Indeed, the consequences of ignoring interviewers' humanity and research capabilities may be particularly important for data quality, as research on interviewer effects has already shown.

This panel invites contributions addressing either qualitative pre-fielding or quantitative survey administration that consider:
1. Whether and how best interviewers can be incorporated into the research team;
2. How interviewers affect the quality of data collected (positively or negatively);
3. How interviewers navigate complex issues, for example, sensitive survey topics or respondents unaccustomed to the survey process;
4. Considerations of the "interviewer" in a self-administered context.

Keywords: interviewers, interviewer effects, cognitive interviewing, data quality, sensitive topics

TUESDAY, July 15th

Optimizing Probability-Based Web Panel Performance

Session Organisers Professor Vasja Vehovar (University of Ljubljana, Faculty of Social Sciences)
Dr Gregor Čehovin (University of Ljubljana, Faculty of Social Sciences)
Ms Andreja Praček (University of Ljubljana, Faculty of Social Sciences)

Time and Room 9:00 - 10:30, Ruppert 0.33

The Rolling Cross-Section Panel Design: Causal Inference for Expected and Unexpected Events

Mr Cristóbal Moya (DIW Berlin / Bielefeld University) - Presenting Author
Dr Monica Gerber (Universidad Diego Portales)

This article proposes an innovative approach for panel surveys by developing a rolling cross-section panel design (RCSP). The RCSP design randomly assigns participants to complete surveys at different time points, enabling to draw causal inferences on the effect of expected and unexpected events that occur during panel waves. It also contributes to identifying the potential effect of events between waves.

While rolling cross-section designs have been primarily applied in cross-sectional studies, we show how the design can be extended to panel study designs. The approach relies on the randomization of participants within waves, which creates equivalent groups between any time points within waves. Moreover, balancing participants with different profiles optimizes statistical power for potential effects from events within and between waves.

Our article describes the RCPS by explaining its conceptual foundations, illustrating it with a study case on police legitimacy in Chile, and showing its properties with simulations based on different scenarios of sample size, attrition, and auxiliary information.

We conclude that the RCSP design contributes a promising tool for panel surveys that is especially suitable for studies in online modes. It can enhance studies with outcomes potentially interacting with unexpected events and provide a sound method to asses how expected events may influence the study outcomes. We also discuss the implementation challenges of this study design.

TUESDAY, July 15th

Collection of Genetic Information in Panel Surveys

Session Organisers Professor David Richter (SHARE Berlin Institute)

Time and Room 11:00 - 12:00, Ruppert 011

Genetic Data in the German Socio-Economic Panel

Professor David Richter (SHARE Berlin Institute) - Presenting Author
Dr Jan Goebel (DIW Berlin)

The German Socio-Economic Panel (SOEP) serves a global research community by providing representative annual longitudinal data of respondents living in private households in Germany. The dataset offers a valuable life course panorama, encompassing living conditions, socioeconomic status, familial connections, personality traits, values, preferences, health, and well-being. To amplify research opportunities further, we have

extended the SOEP Innovation Sample (SOEP-IS) by collecting genetic data from 2,598 participants, yielding the first genotyped dataset for Germany based on a representative population sample (SOEP-G). Consent rates for genetic sampling were 58% for adults but lower for children (26%), reflecting parental hesitance.
Leveraging the results from well-powered genome-wide association studies, we created a repository comprising 66 polygenic indices (PGIs) in the SOEP-G sample. We show that the PGIs for height, BMI, and educational attainment capture 22∼24%, 12∼13%, and 9% of the variance in the respective phenotypes. Using the PGIs for height and BMI, we demonstrate that the considerable increase in average height and the decrease in average BMI in more recent birth cohorts cannot be attributed to genetic shifts within the German population or to age effects alone. These findings suggest an important role of improved environmental conditions in driving these changes. Furthermore, we show that higher values in the PGIs for educational attainment and the highest math class are associated with better self-rated health, illustrating complex relationships between genetics, cognition, behavior, socioeconomic status, and health.
In summary, the SOEP-G data and the PGI repository we created provide a valuable resource for studying individual differences, inequalities, lifecourse development, health, and interactions between genetic predispositions and the environment.

TUESDAY, July 15th

Developing Innovative and Model-Based Interventions to Combat Survey Satisficing

Session Organisers Ms Julia Witton (German Institute for Economic Research)
Dr Carina Cornesse (Free University of Berlin and German Institute for Economic Research)

Time and Room 11:00 - 12:15, Ruppert D – 0.24

The concept of satisficing was introduced to the survey context by Krosnick in 1991. Since then, it has kept researchers busy and remains one of the most important contributors to total survey error. Respondents exhibit problematic response behaviors such as midpoint or endpoint selection, nondifferentiation, or item nonresponse, aiming to reduce the cognitive burden that the survey response process entails. These response strategies evidently impair the quality of the generated data, which in turn affects the validity and reliability of the findings that researchers derive from this data. In times of increasing financial costs and decreasing response rates, it is particularly necessary to maximize quality when collecting survey data. Numerous theoretical insights were gained in the past decades, and there are some general recommendations to prevent participants from satisficing. While in several domains of survey data collection, innovative interventions to maximize the collected data's quality (e.g., targeted data collection mode assignment) have become increasingly relevant, innovation on measures that address the prevention of satisficing is still in its infancy. There are, in fact, some promising results with targeted interventions such as prompting participants when speeding is detected or when they try to skip a question before answering it.
In this session, we want to shed light on current developments in the field that help to

innovate the survey landscape's handling of satisficing. We aim to put a spotlight on the practical aspects and welcome contributions that address existing challenges of satisficing in longitudinal as well as cross-sectional data collection settings. We are particularly interested in the following contributions:
- Interventions based on statistical models (e.g. latent class analyses or prediction models)
- Experimental assessments of satisficing interventions (e.g. split-ballots)
- Longitudinal approaches to measuring and combating satisficing

Keywords: satisficing, survey data collection, measurement error, interventions

Identifying Optimizers, Extremists, and Indifferents: Latent Satisficing Patterns in Panel Surveys

Ms Julia Witton (German Institute for Economic Research (DIW Berlin)) - Presenting Author
Dr Carina Cornesse (GESIS - Leibniz Institute for the Social Sciences)

This study investigates the potential of targeted interventions to prevent satisficing in survey research. The presentation focuses on the robustness and predictiveness of survey satisficing in self-administered mixed-mode panels. The key research interests include identifying distinct patterns of satisficing behavior using latent class analysis (LCA), replicating these patterns across survey waves and modes, determining respondent characteristics that correlate with identified satisficing patterns, and predicting future satisficing behavior based on previous survey waves. The analysis was based on the first three waves of the German Social Cohesion Panel (SCP, jointly conducted by the German Institiute for Economic Research, DIW Berlin, and the Research Institute Social Cohesion, RISC), a mixed-mode panel survey with participants self-selecting into either paper-and-pencil (PAPI) or web (CAWI) mode. The study identified three latent classes that replicate over all waves in CAWI and over the first two waves in PAPI mode: 1) Optimizers, who exhibit the least and unspecific satisficing behavior, 2) ExtreMists, who skip questions and select the endpoint of response scales, 3) Indifferents, who select the scales' midpoints and are at risk of speeding through the survey. Using multinomial regression analyses, we find individual-level characteristics such as education and income to be associated with the latent satisficing pattern with limited predictive power. Binomial logistic regression analyses reveal that the estimated posterior probabilities of belonging to a satisficing class in a previous wave were consistently significant predictors of belonging to that same class in the future. In conclusion, the results emphasize the need for innovative, targeted interventions considering situational factors. Given the robustness of identified patterns, tailored intervention strategies should be developed to effectively combat satisficing. Our research suggests that mode-specific approaches for such interventions may be necessary.

TUESDAY, July 15th

Interviewers across the Survey Life Cycle 2

Session Organisers Dr Mariel Leonard (DIW-Berlin)
Dr Zachary Smith (National Center for Health Statistics (NCHS))

Time and Room 15:30-17:00, Ruppert 011

Interviewers are central to survey operations. From qualitative question design and evaluation – cognitive interviewers, focus group moderators, and even expert reviewers – to quantitative survey administration in the field. A large body of literature has identified various ways the identity, behavior, and disposition of interviewers influence the quality of data collected. And, a growing consensus is developing that in both qualitative and quantitative aspects, interviewers should be understood not merely as mindless, faceless data collection machines, but as researchers that contribute to the research process. Indeed, the consequences of ignoring interviewers' humanity and research capabilities may be particularly important for data quality, as research on interviewer effects has already shown.

This panel invites contributions addressing either qualitative pre-fielding or quantitative survey administration that consider:
1. Whether and how best interviewers can be incorporated into the research team;
2. How interviewers affect the quality of data collected (positively or negatively);
3. How interviewers navigate complex issues, for example, sensitive survey topics or respondents unaccustomed to the survey process;
4. Considerations of the "interviewer" in a self-administered context.

Keywords: interviewers, interviewer effects, cognitive interviewing, data quality, sensitive topics

TUESDAY, July 15th

Split Questionnaire Designs in Social Surveys: Challenges and Solutions

Session Organisers Dr Julian Axenfeld (German Institute for Economic Research (DIW Berlin))
Dr Christian Bruch (GESIS Leibniz Institute for the Social Sciences)
Professor Christof Wolf (GESIS Leibniz Institute for the Social Sciences)

Time and Room 15:30-17:00, Ruppert 0.33

In recent years, split questionnaire designs have gained significant importance, primarily due to the need to reduce survey costs and questionnaire length. In these designs, respondents do not receive the complete questionnaire. Instead, specific parts of the questionnaire, known as modules, are randomly assigned to different respondents, and the resulting design-based missing values are subsequently imputed.

Applying split questionnaire designs to social surveys raises four crucial questions:
(1) How should questionnaire splits be designed to yield good response quality and appropriate respondent burden while also ensuring imputations of adequate quality that allow valid analyses? This includes issues such as constructing modules and determining which and how many items (if any) should be presented to all respondents, i.e. the size of the core module.
(2) What are the effects of split questionnaire designs on respondent behavior, such as response rates and respondent burden?
(3) How should imputation be conducted under the conditions of social surveys? This involves, for instance, selecting imputation methods and models suitable for small sample sizes, low correlations between items, many nominally scaled variables, and a large number of variables in a survey dataset that need to be imputed.
(4) How should the (imputed) data that was collected by a split questionnaire design be provided to data users so that they can conduct their analyses?
We invite papers addressing the challenges of split questionnaire designs in survey practice, particularly with regard but not limited to the aspects described above. Contributions from fields beyond social science that tackle the challenges of split questionnaire designs are also welcome.

Keywords: Split questionnaire designs, multiple imputation, missing values

Composition of Core Modules and Item Allocation in Split Questionnaire Designs: Impact on Estimates From Imputed Data

Dr Julian B. Axenfeld (DIW Berlin) - Presenting Author
Dr Christian Bruch (GESIS Leibniz Institute for the Social Sciences)
Professor Christof Wolf (GESIS Leibniz Institute for the Social Sciences)

An increasing number of social science surveys use split questionnaire designs to reduce questionnaire length, presenting only a subset of several questionnaire modules to each respondent while leaving out others. This approach results in large amounts of planned missing data that necessitates imputation. Research shows that imputation is most effective when each module covers various topics. Yet, single-topic modules may sometimes be preferable from a questionnaire-design perspective. A potential alternative from survey practice is using single-topic modules with an extended core module presented to all respondents that includes key items from all topics. This study investigates whether this strategy yields outcomes comparable to mixed-topic modules. Using Monte-Carlo simulations based on the German Internet Panel, we simulate split questionnaire designs, impute the missing data, and calculate estimates based on these data. Findings suggest that while an extended core module improves single-topic module outcomes, it is inferior to randomly allocated mixed-topic modules.

TUESDAY, July 15th

Video interviewing for survey data collection: beyond the pandemic 2

Session Organisers Mr Tim Hanson (ESS HQ (City St Georges, University of London))
Mr Matt Brown (Centre for Longitudinal Studies, UCL)
Professor Gabriele Durrant (NCRM, University of Southampton)

Time and Room 15:30-17:00, Ruppert 002

Video-Interviewing as Part of a Multi-Mode Design in Panel Studies: Insights From the Field

Ms Julia Witton (German Institute for Economic Research (DIW Berlin)) - Presenting Author
Dr Carina Cornesse (GESIS - Leibniz Institute for the Social Sciences)
Dr Markus Grabka (German Institute for Economic Research (DIW Berlin))
Professor Sabine Zinn (German Institute for Economic Research (DIW Berlin), Humboldt University of Berlin (HUB))

Based on previous research from other countries, computer-assisted live video interviewing (CALVI) can be expected to be a useful addition to existing mixed-mode survey designs.
To assess CALVI's feasibility in a German household panel survey, we included a hypothetical inquiry on respondents' willingness to participate in video interviews during the 2022 data collection wave of the German Socio-Economic Panel (SOEP). Of 22,549 respondents with valid answers, 39% indicated a willingness to participate in CALVI.
Based on these findings, we pretested CALVI in a separate survey setting. In the pretest, 73 target persons scheduled a video appointment, and 46 completed an interview. All participants consented to the recording of the interview, and 44 of them enabled their cameras. On average, the interviews were 106 minutes long. We find high levels of satisfaction among both respondents (85% positive) and interviewers (94% positive) and little item nonreponse or survey break-off.
Building on the pretest, we implemented CALVI in the ongoing wave of the SOEP Innovation Sample (SOEP-IS), targeting a mix of households interviewed in the previous wave using either CAWI or computer-assisted personal interviewing (CAPI). Our study employs a stratified randomized experimental design, with 50% of the panels' CAPI households as well as 25% of CAWI households encouraged to switch to CALVI. This design allows us to evaluate which households transition to CALVI based on their data collection mode in the previous wave.
In our presentation, we will share findings from the hypothetical willingness-to-participate inquiry, pretest outcomes, and experimental fieldwork results, including response rates, sample composition, and data completeness. Additionally, we will discuss practical lessons learned and the implications of CALVI for improving the efficiency and quality of panel data collection.

WEDNESDAY, July 16th

Current developments in improving digital accessibility of web surveys

Session Organisers Dr Marika de Bruijne (Centerdata)
Ms Mara Verheijen (Centerdata)

Time and Room 09:00 - 10:30, Ruppert 005

The SHARE Self-completion Questionnaire experiment: Developing a web questionnaire for Europeans aged 50 and older

Dr Barbara Thumann (SHARE Berlin Institute) - Presenting Author
Dr Marika de Bruijne (Centerdata)
Ms Theresa Fabel (SHARE Berlin Institute)
Dr Daniel Horn (SHARE Berlin Institute)
Dr Yuri Pettinicchi (SHARE Berlin Institute)
Dr Elena Sommer (German Institute for Economic Research (DIW Berlin))
Dr Michael Bergmann (SHARE Berlin Institute)

The Survey of Health, Ageing and Retirement in Europe (SHARE) will administer a self-completion web questionnaire as part of an experimental study that utilizes a concurrent mixed-mode design (web questionnaire and paper & pencil questionnaire) in 4–6 European countries in autumn 2025. This experiment aims to provide valuable insights into the feasibility of conducting web surveys among Europeans aged 50 and older, many of whom face disabilities such as impaired eyesight.

In this context, the Web Content Accessibility Guidelines (WCAG) play a crucial role. We will evaluate the applicability of these guidelines to our web survey and detail the specific recommendations that have been implemented. For example, the web questionnaire will allow respondents to pause and resume completion without losing previously entered information (guideline 2.2).

We will also address the guidelines that could not be fully implemented, such as ensuring all functionality is accessible via a keyboard (guideline 2.1), and provide an explanation for these limitations. Lastly, we will outline additional measures taken to improve accessibility for our target population. Examples include designing the survey to be compatible with various devices and providing access details to the survey via both email and postal letters.

WEDNESDAY, July 16th

Exploring Organizational Structures in Survey Infrastructures 2

Session Organisers Dr Olga Grunwald (NIDI)
Dr Roman Auriga (LIfBi)
Mr Niccolo Ghirelli (ESS HQ (City St George's, University of London))
Ms Victoria Salinero-Bevins (ESS HQ (City St George's, University of London))

Time and Room 09:00 - 10:30, Ruppert 002

Panel Survey Life Spirals: A Framework for Longitudinal Survey Research

Mr Florian Griese (German Institute for Economic Research, Berlin) - Presenting Author
Professor Sabine Zinn (German Institute for Economic Research, Berlin)
Dr Christian Hunkler (Humboldt University Berlin)

The design and implementation of panel studies follow recurring processes often visualized as a Survey Life Cycle, which organizes survey tasks into distinct, sequential steps. While effective for single waves or cross-sectional surveys, this framework falls short when managing overlapping waves in longitudinal studies. To address this, we propose the Panel Survey Life Spirals, a framework capturing the dynamic and interconnected nature of successive survey waves.

The German Socio-Economic Panel (SOEP), one of the world's longest-running household surveys, highlights the challenges of overlapping waves. With extended fieldwork and a complex dataset, the SOEP team simultaneously works on multiple waves at different stages. For instance, in winter, the team prepares the wave for two years ahead, plans next year's data collection, releases the current year's data, and delivers training on the latest released wave. This overlapping workflow creates significant interdependencies between waves.

The Panel Survey Life Spirals framework enhances the traditional life cycle by addressing the overlapping, iterative, and interdependent nature of longitudinal surveys. It emphasizes:

- Interwave Dependencies: Critical stages, like sample definition, depend on timely completion of earlier waves.
- Continuous Learning: Lessons from prior waves improve quality and efficiency.
- Dynamic Chronology: The spiral visualizes multiple waves at various stages.

This paper outlines SOEP's processes within the spiral framework, emphasizing their chronological positioning and role in ensuring data quality. By adopting the Panel Survey Life Spirals, researchers can address unique challenges in longitudinal studies and adapt their workflows.

As societal change accelerates and demand for reliable data grows, the Panel Survey Life Spirals offer a flexible yet standardized guideline for managing complex survey processes and producing high-quality data to meet evolving needs.

WEDNESDAY, July 16th

Survey Data Integration: Nonprobability Surveys, Administrative and Digital Trace Data 3

Session Organisers Dr Camilla Salvatore (Utrecht University)
Dr Angelo Moretti (Utrecht University)

Time and Room 09:00 - 10:30, Ruppert Wit – 0.52

Linking survey data on migrants and refugees to administrative register data: Consenter selectivity and cross-validation in the SOEP-CMI-ADIAB sample

Mr Manfred Antoni (IAB - Institute for Employment Research)
Mr Mattis Beckmannshagen (The German Institute for Economic Research - DIW)
Mr Markus Grabka (The German Institute for Economic Research - DIW)
Mr Sekou Keita (IAB - Institute for Employment Research) - Presenting Author
Ms Parvati Trübswetter (IAB - Institute for Employment Research)

Abstract
The SOEP-CMI-ADIAB record linkage project linked the Socio-Economic Panel (SOEP) core, migration, and innovation (CMI) samples to administrative data available at the IAB (ADIAB) to generate a new dataset available for research and policy advice. This study examines the quality of the successfully linked data in terms of selectivity. In addition, it identifies steps that can be taken as best practices in handling the data to minimize potential problems resulting from measurement inaccuracies or selectivity of the sample.

1. Selectivity analyses
Selectivity and thus lack of representativeness of the SOEP-CMI-ADIAB sample can arise at three levels.
a) At the level of consent: Were respondents with certain characteristics more likely to give consent to data linkage than others?
b) At the level of data linkage: Does the probability of success of data linkage vary based on certain characteristics of the respondents?
Logistic regressions within each relevant population are used to test whether certain characteristics had significant influence at any of the three levels. Subsequently, hints are given on how to compensate for potential selectivity by creating weighting factors that take all levels into account.

2 Cross-validation of gross compensation and part-time/full-time variable
Because of the linked data set, two sources of information exist for many respondents on the same (or at least very similar) questions. For example, the administrative data contain information on employees' gross daily pay, while the SOEP asks about gross monthly pay. At the same time, the administrative data contain information on whether employees work full or part time, while the SOEP survey data contain actual as well as contractually agreed weekly working hours.
On this basis, cross-validations can be performed. Does the SOEP information agree (approximately) with the administrative data?

WEDNESDAY, July 16th

Interviewers across the Survey Life Cycle 3

Session Organisers Dr Mariel Leonard (DIW-Berlin)
Dr Zachary Smith (National Center for Health Statistics (NCHS))

Time and Room 09:00 - 10:30, Ruppert 0.33

Interviewers are central to survey operations. From qualitative question design and evaluation – cognitive interviewers, focus group moderators, and even expert reviewers – to quantitative survey administration in the field. A large body of literature has identified various ways the identity, behavior, and disposition of interviewers influence the quality of data collected. And, a growing consensus is developing that in both qualitative and quantitative aspects, interviewers should be understood not merely as mindless, faceless data collection machines, but as researchers that contribute to the research process. Indeed, the consequences of ignoring interviewers' humanity and research capabilities may be particularly important for data quality, as research on interviewer effects has already shown.

This panel invites contributions addressing either qualitative pre-fielding or quantitative survey administration that consider:
1. Whether and how best interviewers can be incorporated into the research team;
2. How interviewers affect the quality of data collected (positively or negatively);
3. How interviewers navigate complex issues, for example, sensitive survey topics or respondents unaccustomed to the survey process;
4. Considerations of the "interviewer" in a self-administered context.

Keywords: interviewers, interviewer effects, cognitive interviewing, data quality, sensitive topics

Interviewer Understandings of Rapport

Dr Mariel Leonard (DIW-Berlin) - Presenting Author

Rapport is considered essential to the success of interviews, particularly those that are long or involve sensitive questions. Literature on survey methods – from Standard Survey Interviewing to in-house training manuals – emphasize the absolute necessity of establishing

rapport with survey respondents. Yet, despite its frequent appearance in the literature, "rapport" is a term that is often left undefined, resulting in unclear standards for training and field monitoring.

In this paper, I consider how interviewers themselves define rapport within their work. I analyze data from in-depth interviews collected as part of the Interviewer Quality of Life Study. I find that interviewers are able to provide cogent definitions of rapport that may provide a strong foundations for improving interviewer training as well as monitoring of rapport-building in the field.

WEDNESDAY, July 16th

Questionnaire translation in a changing world: challenges and opportunities

Session Organisers Dr Alisú Schoua-Glusberg (Research Support Services)
Dr Brita Dr. Dorer (GESIS-Leibniz Institute for the Social Sciences)
Dr Dorothée Behr (GESIS-Leibniz Institute for the Social Sciences)

Time and Room 11:00 - 12:30, Ruppert paars - 0.44

Questionnaire design of the revised Human Values Scale for cross-cultural surveys

Mr Tim Hanson (European Social Survey HQ (City St Georges, University of London)) - Presenting Author
Dr Elena Sommer (German Institute for Economic Research (DIW Berlin))
Dr Brita Dorer (GESIS)
Ms Ulrike Efu Nkong (GESIS)
Professor Shalom Schwartz (Hebrew University of Jerusalem)

Since its first round, the European Social Survey (ESS) has included a 21-item measure of ten basic values shared across cultures. This instrument developed by Shalom Schwartz and known as the Human Values Scale (HVS) has been widely used in various disciplines. Following recent recommendations from Schwartz, the ESS will use the new revised 20-item HVS starting from its 12th round (2025-26). The revised scale offers greater reliability and uses shorter and simpler items than the initial scale. It also introduces a single gender-neutral version for all respondents, replacing the previous separate male and female versions. This change is particularly important given the ESS's upcoming transition to self-completion mode and use of a paper questionnaire. To ensure that the new scale measures the same concepts in all participating countries and languages, the ESS conducted an "advance translation" review producing a list of translation annotations to clarify the precise meaning of source items.

The new HVS will also be fielded for the first time in the German Socio-economic Panel (SOEP), including the refugee sample, in 2025. While both surveys use the same source text, the SOEP additionally employs images to visualise the statements, aiming to enhance

respondents' engagement in the self-completion module on social distance at the end of a long interview. A multilingual cognitive pretest was conducted to assess whether the images match the statements and whether the statements were interpreted consistently by respondents with different migration background.

As part of a collaborative effort, findings from the ESS' advance translation and the SOEP's cognitive pretest were used to refine the final version of the new HVS. Our presentation provides an overview of the new HVS highlighting key findings from the testing procedures and the challenges of cross-cultural adaptation.

WEDNESDAY, July 16th

The Use of Machine Learning Techniques When Dealing with Missing Data

Session Organisers Dr Barbara Felderer (GESIS)
Dr Christian Bruch (GESIS)

Time and Room 11:00 - 12:30, Ruppert D - 0.24

Comparison of machine learning procedures to impute missing values in survey data

Mr Christian Bruch (GESIS Leibniz Institute for the Social Sciences) - Presenting Author
Mr Julian Axenfeld (German Institute for Economic Research (DIW Berlin)

In recent years, machine learning methods such as random forests, neural networks or k-means clustering are often discussed as means to improve the imputation of missing values in social surveys over the established imputation methods.
However, machine learning techniques are often based on complex algorithms and many components that have to be estimated. Furthermore, many of these algorithms were originally developed for "big data" problems in which tremendous amounts of observations and variables are used, for instance, to predict a target variable as accurately as possible. Survey data imputation is different though: There are often only up to a few thousand observations available, and the main goal of imputation typically is preserving the relationships between all the different variables, rather than accurately predicting variables using a (data-driven) selection of predictor variables. In addition, in survey data many relationships of interest tend to be relatively weak, which makes it more difficult for imputation techniques to detect and reproduce them accurately even with the established imputation methods. Therefore, the question is if (and to what extent) machine learning methods are capable of reproducing relationships in the data in practice.
To evaluate the procedures, we will use Monte Carlo Simulation studies. For ensuring realistic conditions in the simulation studies, we will use real survey data in which we will simulate item nonresponse. In this presentation, we will show first findings on how different procedures affect correlation estimates after imputation.

WEDNESDAY, July 16th

Evaluating Mixed-Mode in Panel Surveys: Where do we stand and where do we go from here 2?

Session Organisers Dr Patricia Hadler (GESIS - Leibniz Institute for the Social Sciences)
Dr Steffen Pötzschke (GESIS - Leibniz Institute for the Social Sciences)

Time and Room 11:00 - 12:30, Ruppert 042

Introducing Mixed-Mode in a Refugee Household Panel – An Initial Evaluation

Ms Theresa Müller (infas Institut für angewandte Sozialwissenschaft GmbH (Institute for Applied Social Sciences)) - Presenting Author
Mr Michael Ruland (infas Institut für angewandte Sozialwissenschaft GmbH (Institute for Applied Social Sciences))
Dr Elena Sommer (Socio-Economic Panel at DIW)
Professor Sabine Zinn (DIW-SOEP / Humboldt University Berlin)

The IAB-BAMF-SOEP Survey of Refugees in Germany is an annual household panel conducted as part of the Socio-Economic Panel (SOEP) since 2016. Its primary data collection method has always been CAPI, at both the household and individual level. Households are approached by interviewers, and until recently, at least the household interview was required to be conducted face-to-face but there are options to switch to a self-administered mode (CAWI or CASI) at the individual level.

In the 2024 survey wave, households were, for the first time, invited to complete the entire survey online as part of a CAWI follow-up. This sequential mixed-mode design at the household level was introduced during the final phase of fieldwork to boost response among households, that were hard to reach or motivate.

One the one hand, this paper evaluates which households could be reached through the CAWI follow-up and whether panel attrition could be reduced by offering this additional mode. On the other hand, we compare data quality indicators between households surveyed via CAPI and CAWI. Furthermore, we analyze which households should be offered the option to switch to CAWI at the household level and at which point in the field period this switch offer could be most effective.

WEDNESDAY, July 16th

Item Nonresponse and Unit Nonresponse in Panel Studies

Session Organisers Dr Uta Landrock (LIfBi – Leibniz Institute for Educational Trajectories)
Dr Ariane Würbach (LIfBi – Leibniz Institute for Educational Trajectories)
Mr Michael Bergrab (LIfBi – Leibniz Institute for Educational Trajectories)

Time and Room 13:30 – 15:00, Ruppert rood – 0.51

Unveiling Hidden Patterns: A Flexible Approach to Understanding Survey Nonresponse

Mr Felix Süttmann (German Institute for Economic Research) - Presenting Author
Professor Sabine Zinn (German Institute for Economic Research)

Recent developments in mixed survey mode designs and the growing availability of auxiliary and administrative data have opened new pathways for improving models that address unit nonresponse. While machine learning methods, like random forests, have gained attention for their strong predictive performance, they often lack interpretability, limiting their utility for researchers seeking to understand the underlying selection mechanisms in their data. On the other hand, traditional parametric Generalized Linear Models (GLMs) with main effects, commonly used for this purpose, impose rigid assumptions about variable relationships and fail to capture or explain interactions between covariates, reducing their descriptive power.

To overcome these limitations, we propose a logit model with grouped LASSO to detect interpretable interaction effects. This innovative approach accommodates larger sets of variables while ensuring that any identified interactions always include their corresponding main effects. Through a structured penalization technique, our method relaxes the restrictive functional form of traditional models, enhancing flexibility without compromising interpretability. Moreover, it leverages the LASSO's favorable bias-variance tradeoff, offering a robust balance between predictive accuracy and explanatory clarity.

Using data from the German Socio-Economic Panel (SOEP), we construct a comprehensive dataset integrating past-wave information, regional data, and household fieldwork metadata. We then implement and benchmark our model against a random forest and traditional logit models to assess performance. Finally, we identify the most relevant interaction effects and discuss their implications for understanding nonresponse mechanisms.

WEDNESDAY, July 16th

Good, fast, cheap: Pick two – Optimizing Sampling Strategies for Modern Survey Research

Session Organisers Professor Sabine Zinn (DIW-SOEP / Humboldt University Berlin)
Dr Hans Walter Steinhauer (Socio-Economic Panel at DIW)

Time and Room 13:45 – 15:00, Ruppert A – 0.21

Survey research is increasingly adapting to the demands of fast-paced environments where timely, reliable data is crucial, often within limited budgets. To meet these demands, researchers frequently use non-random sampling and online data collection, which provide quick results but may lack reliability. Traditional methods that ensure accuracy are slower and more costly, yet essential for scientific research and policymaking.

This session invites contributions on the practical use of sampling frames for generating random samples, such as population registers or geo-referenced databases. We are also interested in research on non-probability sampling methods, including data from social media, affordable access panels like Prolific, and respondent-driven sampling schemes. Our goal is to examine the pros and cons of these sampling strategies, focusing on coverage, bias, generalizability, cost, and speed.

We seek discussions on optimal sampling approaches tailored to specific study needs, where researchers must balance the urgency of obtaining rapid results with the need for high-quality studies that can inform policy recommendations.

We invite submissions on:
- Innovative sampling frames for social science surveys
- Combining different sampling frames to enhance data quality and timeliness
- Methods for improving accuracy and quick data access
- Cost analyses of various sampling strategies
- Experiences using fast-access data from web providers like Prolific and Respondi for social science research

Through these discussions, we aim to guide the development of more effective and efficient approaches to survey research in today's fast-paced data environment.

Keywords: random sampling, non-random sampling, combining sampling frames

WEDNESDAY, July 16th

New developments in research data infrastructures 2

Session Organisers Dr Daniel Fuß (Leibniz Institute for Educational Trajectories (LIfBi))

Time and Room 13:45 – 15:00, Ruppert 114

RDCnet - Connecting secure workplaces

Mr Neill Murray (DIW Berlin) - Presenting Author

The RDCnet aims to facilitate access to sensitive research data by fostering collaborations between Research Data Centers (RDCs) and enabling the mutual use of their secure workstations. Rather than requiring researchers to travel to a specific data-providing RDC to analyze their data on-site, RDCnet offers a decentralized network of secure workstations at various partner locations, providing researchers with greater flexibility. At its core, RDCnet is based on the idea that each participating institution provides a secure workstation and, optionally, remote access to their research data while being interconnected with all other institutions in the network. This approach ensures that data providers maintain full control over who can access their data and where, while enabling researchers to work with sensitive data from any secure workstation within the RDCnet. To ensure the required levels of data protection, all secure workstations must be maintained and configured according to standardized security criteria, developed collaboratively with eight RDCs in Germany. This guarantees that sensitive data is processed exclusively within strictly controlled environments. To realize RDCnet, we provide essential services and support to facilitate multilateral collaboration. These include a unified cooperation agreement with clear organizational and technical guidelines, a shared platform for booking secure workstations, and technical support for implementing secure work environments and remote access systems. By lowering access barriers, RDCnet reduces costs for researchers working with sensitive data while simultaneously improving access possibilities for data providers. This approach not only ensures efficient and secure data use but also enhances the visibility of research data, ultimately increasing the number of potential users.

WEDNESDAY, July 16th

Missing Data, Selection Bias and Informative Censoring in Cross-Sectional and Longitudinal Surveys

Session Organisers Dr Angelina Hammon (SOEP, DIW Berlin)
Mx Char Hilgers (SOEP, DIW Berlin)
Professor Sabine Zinn (SOEP, DIW Berlin)

Time and Room 14:00 – 15:00, Ruppert 0.33

Sample selection bias, item non-response, and dropouts (a form of censoring) are common challenges in large-scale population surveys. In longitudinal surveys, selection bias occurs at the start, item non-response during the survey, and dropout (censoring) at the end. In cross-sectional surveys, selection bias and non-response are the primary sources of missing data. These issues can severely impact the quality of analysis and the validity of inferences if not properly addressed.

A special challenge for analysis occurs when the mechanism driving one of those phenomena depends (additionally) on unobserved information, making the missing data not random and potentially leading to non-ignorable selection bias, informative censoring or non-ignorable missing data. Consequently, it is crucial to assess the robustness of results under different plausible assumptions about the missing-data, selection, or censoring mechanisms, when it seems plausible that standard assumptions may not hold.

In this session, we welcome research on novel and innovative methods to prevent misleading inference under one or several of the described challenges related to incomplete, biasedly selected, or censored survey data. This research might cover:

1. Use cases showing the harm of non-ignorable selection bias, informative censoring, or non-ignorable missing data.
2. Novel approaches for detecting (non-ignorable) selection bias in traditional surveys and non-probability samples.
3. Novel imputation procedures, likelihood-based approaches, machine learning and tree-based methods and Bayesian estimation techniques to address (non-ignorable) missing data and/or informative censoring.
4. Methods for conducting sensitivity analyses in cases where deviations from missing at random mechanisms are realistic.

Keywords: (non-ignorable) missing data, multiple imputation, (non-ignorable) selection bias, informative censoring, panel dropout, missing not at random, sensitivity analysis

Filling in the Blanks: Augmenting Survey Data Imputation with External Data and Rubin's SIR Algorithm

Mx Char Hilgers (DIW Berlin, Socio-Economic Panel) - Presenting Author
Professor Sabine Zinn (DIW Berlin, Socio-Economic Panel)

Multiple imputation of missing values in survey data analysis is a state-of-the-art technique. Typically, methods like multivariate imputation by chained equations (mice, van Buuren 2018) are employed, replacing missing values on a variable-by-variable basis. The information used for imputation usually comes from the survey dataset being analysed. Valid analysis results are achieved when the missing values are either missing completely at random (MCAR) or missing at random (MAR). However, the situation becomes more complex if the values are missing not at random (MNAR).

There are some approaches handle this issue. One approach incorporates sensitivity analyses into the imputation to make it as robust as possible. Alternatively, the data set to be imputed can be enriched with further information, so that an MNAR mechanism becomes MAR, and thus the imputation and analysis of the imputed data can be valid. The advantages of this approach are clear, but often the full range of variables of the data set is already included in the imputation, and still the suspicion of MNAR remains.
We present a new method that integrates external data into the mice imputation process to reduce the risk of MNAR and better justify the assumption of a MAR mechanism.

Specifically, we integrate Rubin's SIR (Sampling/Importance Resampling) algorithm (Rubin 1987) into the mice framework to incorporate external distribution information for the variable of interest. Importance ratios, derived from the differences between the external distribution and the survey data's estimated distribution, guide the selection of replacement values for missing data. We also provide an estimate of uncertainty introduced by the method.
We demonstrate the effectiveness of our new approach with a simulation, involving the imputation of a typical income variable. Additionally, we apply this method to two datasets from the German Socio-Economic Panel Study.

WEDNESDAY, July 16th

Missing Data, Selection Bias and Informative Censoring in Cross-Sectional and Longitudinal Surveys 2

Session Organisers Dr Angelina Hammon (SOEP, DIW Berlin)
Mx Char Hilgers (SOEP, DIW Berlin)
Professor Sabine Zinn (SOEP, DIW Berlin)

Time and Room 16:00 – 17:30, Ruppert 002

Sample selection bias, item non-response, and dropouts (a form of censoring) are common challenges in large-scale population surveys. In longitudinal surveys, selection bias occurs at the start, item non-response during the survey, and dropout (censoring) at the end. In cross-sectional surveys, selection bias and non-response are the primary sources of missing data. These issues can severely impact the quality of analysis and the validity of inferences if not properly addressed.

A special challenge for analysis occurs when the mechanism driving one of those phenomena depends (additionally) on unobserved information, making the missing data not random and potentially leading to non-ignorable selection bias, informative censoring or non-ignorable missing data. Consequently, it is crucial to assess the robustness of results under different plausible assumptions about the missing-data, selection, or censoring mechanisms, when it seems plausible that standard assumptions may not hold.

In this session, we welcome research on novel and innovative methods to prevent misleading inference under one or several of the described challenges related to incomplete, biasedly selected, or censored survey data. This research might cover:

1. Use cases showing the harm of non-ignorable selection bias, informative censoring, or non-ignorable missing data.
2. Novel approaches for detecting (non-ignorable) selection bias in traditional surveys and non-probability samples.
3. Novel imputation procedures, likelihood-based approaches, machine learning and tree-based methods and Bayesian estimation techniques to address (non-ignorable) missing data and/or informative censoring.
4. Methods for conducting sensitivity analyses in cases where deviations from missing at random mechanisms are realistic.

Keywords: (non-ignorable) missing data, multiple imputation, (non-ignorable) selection bias, informative censoring, panel dropout, missing not at random, sensitivity analysis

WEDNESDAY, July 16th

Good, fast, cheap: Pick two – Optimizing Sampling Strategies for Modern Survey Research 2

Session Organisers Professor Sabine Zinn (DIW-SOEP / Humboldt University Berlin)
Dr Hans Walter Steinhauer (Socio-Economic Panel at DIW)

Time and Room 16:00 – 17:30, Ruppert A – 0.21

Survey research is increasingly adapting to the demands of fast-paced environments where timely, reliable data is crucial, often within limited budgets. To meet these demands, researchers frequently use non-random sampling and online data collection, which provide quick results but may lack reliability. Traditional methods that ensure accuracy are slower and more costly, yet essential for scientific research and policymaking.

This session invites contributions on the practical use of sampling frames for generating random samples, such as population registers or geo-referenced databases. We are also interested in research on non-probability sampling methods, including data from social media, affordable access panels like Prolific, and respondent-driven sampling schemes. Our goal is to examine the pros and cons of these sampling strategies, focusing on coverage, bias, generalizability, cost, and speed.

We seek discussions on optimal sampling approaches tailored to specific study needs, where researchers must balance the urgency of obtaining rapid results with the need for high-quality studies that can inform policy recommendations.

WEDNESDAY, July 16th

Number of hours usually worked? Methodological challenges in accurately measuring working time

Session Organisers Miss Carolin Deuflhard (Humboldt-Universität zu Berlin)
Professor Lena Hipp (University of Potsdam/ WZB Berlin Social Science Center

Time and Room 16:00 – 17:30, Ruppert 134

Time is Money

Dr Mattis Beckmannshagen (German Institute for Economic Research (DIW Berlin))
Dr Markus Grabka (German Institute for Economic Research (DIW Berlin))
Dr Ralf Himmelreicher (Federal Institute for Occupational Safety and Health (BAuA))
Ms Juliane Pehla (Federal Institute for Occupational Safety and Health (BAuA)) - Presenting Author
Mr Johannes Seebauer (German Institute for Economic Research (DIW Berlin))
Professor Carsten Schröder (German Institute for Economic Research (DIW Berlin))

In labor economics, hourly wages are critical as proxies for productivity and human capital and as variables in policy interventions. Valid hourly wage calculations require accurate documentation of earnings and working hours. While earnings documentation is strictly regulated, working hours often lack comparable rigor, causing variability in wage estimates based on datasets, definitions of working time, and recording methods.

Three main working time definitions — contractually agreed, employer-recorded, and self-reported actual hours — pose specific measurement challenges. To address these, the German Minimum Wage Commission launched a specialized survey within the German Socio-Economic Panel (GSOEP), a key labor market data source. This survey, conducted from April to December 2024, focuses on minimum wage employees. While GSOEP regularly collects monthly earnings and weekly working hours data, this initiative introduces a postal

drop-off survey for detailed insights into how companies document earnings and working hours. A primary goal is aligning earnings and working hour reference periods to the same calendar month for more valid hourly wage estimates.

By comparing Drop-Off survey data with regular GSOEP data, the study identifies measurement errors from differing reference periods and examines discrepancies between agreed, recorded, and actual working hours. It also investigates whether preparation and follow-up times for jobs are remunerated. Respondents can report hours weekly or monthly, offering insights into reporting preferences across employee groups.

Findings aim to refine future surveys, advancing research on labor markets and minimum wage policies. Enhanced survey designs will enable more accurate analyses of working hours and earnings. Initial results will be shared at the conference, contributing to the broader discussion on innovative methods for measuring working hours.

THURSDAY, July 17th

New Developments in Using, Sharing, and Re-using Metadata

Session Organisers Mr Knut Wenzig (DIW Berlin/SOEP)
Mr Daniel Bela (LIfBi)
Dr Arne Bethmann (SHARE Germany and SHARE Berlin Institute)

Time and Room 09:00 - 10:30, Ruppert 119

Metadata systems have evolved from passive documentation tools into active drivers of data management and utilization. This session explores recent advancements that enhance the use, sharing, and re-use of metadata across the data lifecycle, emphasizing innovative methods that improve data quality, interoperability, and efficiency.

With machine-readable metadata, processes like survey instrument generation, data validation, and preparation are increasingly automated, reducing errors and enhancing data-driven decision-making. Metadata systems are becoming essential components in not just documenting data, but actively shaping and streamlining the entire data lifecycle.

We invite papers that highlight:

- Innovative Uses: Examples of how metadata systems are leveraged for automation and optimization in data collection, processing, and analysis.
- Interoperability: Experiences with implementing metadata standards (e.g., DDI, SDMX) to facilitate sharing and re-use across different systems and institutions.
- Collaborative Platforms: Case studies on platforms that support community-driven creation, sharing, and re-use of metadata.
- FAIR Principles: Approaches that ensure metadata adheres to the FAIR (Findable, Accessible, Interoperable, Reusable) principles.
- Future Directions: Emerging technologies, such as AI and machine learning, that could

revolutionize metadata use.

This session aims to provide a comprehensive overview of current trends and future directions in metadata management. We seek presentations that not only showcase technological advancements but also discuss the practical challenges and lessons learned in implementing these innovations. By bringing together researchers, data managers, and technologists, this session will foster a rich exchange of ideas on how new developments in metadata can lead to more effective and insightful data management practices.

Keywords: Metadata Management, Interoperability, FAIR Principles, Data Automation, Metadata Standards

THURSDAY, July 17th

Good, fast, cheap: Pick two – Optimizing Sampling Strategies for Modern Survey Research 3

Session Organisers Professor Sabine Zinn (DIW-SOEP / Humboldt University Berlin)
Dr Hans Walter Steinhauer (Socio-Economic Panel at DIW)

Time and Room 13:45 – 15:00, Ruppert A – 0.21

Survey research is increasingly adapting to the demands of fast-paced environments where timely, reliable data is crucial, often within limited budgets. To meet these demands, researchers frequently use non-random sampling and online data collection, which provide quick results but may lack reliability. Traditional methods that ensure accuracy are slower and more costly, yet essential for scientific research and policymaking.

This session invites contributions on the practical use of sampling frames for generating random samples, such as population registers or geo-referenced databases. We are also interested in research on non-probability sampling methods, including data from social media, affordable access panels like Prolific, and respondent-driven sampling schemes. Our goal is to examine the pros and cons of these sampling strategies, focusing on coverage, bias, generalizability, cost, and speed.

We seek discussions on optimal sampling approaches tailored to specific study needs, where researchers must balance the urgency of obtaining rapid results with the need for high-quality studies that can inform policy recommendations.

We invite submissions on:
- Innovative sampling frames for social science surveys
- Combining different sampling frames to enhance data quality and timeliness
- Methods for improving accuracy and quick data access
- Cost analyses of various sampling strategies
- Experiences using fast-access data from web providers like Prolific and Respondi for social science research

Through these discussions, we aim to guide the development of more effective and efficient approaches to survey research in today's fast-paced data environment.

Keywords: random sampling, non-random sampling, combining sampling frames

THURSDAY, July 17th

New Developments in Using, Sharing, and Re-using Metadata 2

Session Organisers Mr Knut Wenzig (DIW Berlin/SOEP)
Mr Daniel Bela (LIfBi)
Dr Arne Bethmann (SHARE Germany and SHARE Berlin Institute)
Dr Yuri Pettinicchi (SHARE)

Time and Room 14:00 - 15:00, Ruppert 119

Metadata systems have evolved from passive documentation tools into active drivers of data management and utilization. This session explores recent advancements that enhance the use, sharing, and re-use of metadata across the data lifecycle, emphasizing innovative methods that improve data quality, interoperability, and efficiency.

With machine-readable metadata, processes like survey instrument generation, data validation, and preparation are increasingly automated, reducing errors and enhancing data-driven decision-making. Metadata systems are becoming essential components in not just documenting data, but actively shaping and streamlining the entire data lifecycle.

We invite papers that highlight:

- Innovative Uses: Examples of how metadata systems are leveraged for automation and optimization in data collection, processing, and analysis.
- Interoperability: Experiences with implementing metadata standards (e.g., DDI, SDMX) to facilitate sharing and re-use across different systems and institutions.
- Collaborative Platforms: Case studies on platforms that support community-driven creation, sharing, and re-use of metadata.
- FAIR Principles: Approaches that ensure metadata adheres to the FAIR (Findable, Accessible, Interoperable, Reusable) principles.
- Future Directions: Emerging technologies, such as AI and machine learning, that could revolutionize metadata use.

This session aims to provide a comprehensive overview of current trends and future directions in metadata management. We seek presentations that not only showcase technological advancements but also discuss the practical challenges and lessons learned in implementing these innovations. By bringing together researchers, data managers, and technologists, this session will foster a rich exchange of ideas on how new developments in metadata can lead to more effective and insightful data management practices.

Keywords: Metadata Management, Interoperability, FAIR Principles, Data Automation, Metadata Standards

Metadata-Driven Production of Longitudinal Datasets: Introduction and Potential of the Metadata Attribute

Mrs Claudia Saalbach (DIW Berlin/SOEP)
Mrs Jana Nebelin (DIW Berlin/SOEP) - Presenting Author

The production of longitudinal survey datasets poses significant technical and substantive challenges for data producers. In particular, the integration, analysis, and documentation of large volumes of data require sophisticated processes and methodological expertise. From questionnaire development to dataset definition and the final generation of longitudinal datasets, all steps are increasingly driven by metadata. At the same time, it is essential to ensure that the extensive data offerings are user-friendly and efficiently accessible.
At the SOEP (Socio-Economic Panel), approximately 23 survey instruments are deployed annually, yielding around 41 raw survey datasets and roughly 15 longitudinal datasets, covering nearly 15,000 longitudinal variables from 1984 to the present. The SOEP's metadata system plays a pivotal role in this process, offering various levels of granularity – from study, questionnaire, dataset, and variables to topics and concepts. A new metadata attribute, the "module," extends this system by positioning itself conceptually between topics and concepts and structurally between datasets and variables.
The introduction of the "module" attribute allows for greater flexibility and efficiency in the generation of longitudinal data products. Instead of processing large datasets globally, targeted groups of variables – thematically cohesive and methodologically grounded – can be handled separately. This approach not only enhances the efficiency and speed of data production but also contributes to improving data quality.
In our presentation, we provide insights into the practical implementation of metadata-driven production at SOEP and demonstrate the potential of the "module" metadata attribute for both data producers and users. Against the backdrop of related concepts, such as "topic" or "concept" (e.g., at GESIS), we also propose an initial definition of the term "module" and discuss its added value in the context of longitudinal data production.

THURSDAY, July 17th

Methods of including and boosting under-represented population subgroups

Session Organisers Dr Olga Maslovskaya (University of Southampton)
Dr Carina Cornesse (DIW Berlin)
Mr Curtis Jessop (NatCen)
Dr Mariel Leonard (DIW)

Time and Room 15:30-17:00, Ruppert 002

Some population subgroups are consistently under-represented or excluded from survey samples or may be present in too small frequencies for robust subgroup analysis. While approaches to targeting these population subgroups exist, they may not provide sufficient level of quality for social surveys (e.g., use of convenience samples) or may be prohibitively expensive or unfeasible in a self-completion mode (e.g., over-sampling for the Postcode Address File (PAF) and/or screening). As survey research globally aims to become more inclusive, it is crucial to explore effective methods of including and boosting under-represented population subgroups.

In this session we would like to investigate different effective methods of including and boosting under-represented population subgroups including respondent-driven sampling approach.

We encourage papers from researchers with a variety of backgrounds and across different sectors, including academia, national statistics, and research agencies. We particularly welcome contributions that use experimental designs, and/or other designs that can inform future strategies for including and boosting under-represented subgroups in large scale high-quality probability-based surveys.

The session is proposed by Research Strand 1 of the Survey Futures project, “Enhanced Sampling Frames and Procedures”. Survey Futures is a UKRI-ESRC funded research programme focused on ensuring large-scale social surveys in the UK can innovate and adapt in a changing environment. https://surveyfutures.net/

Keywords: under-represented subgroups, respondent-driven sampling, high quality surveys

Implementing respondent-driven sampling to increase the diversity of a general population sample

Dr Mariel Leonard (DIW-Berlin) - Presenting Author
Ms Julia Witton (DIW-Berlin)
Dr Carina Cornesse (GESIS)
Dr Julian Axenfeld (DIW-Berlin)
Dr Jean-Yves Gerlitz (University of Bremen)

Dr Olaf Groh-Samberg (University of Bremen)
Dr Sabine Zinn (DIW-Berlin)

Respondent-driven sampling (RDS) is a network sampling technique for surveying complex populations in the absence of sampling frames. The idea is simple: identify some people ("seeds") who belong or have access to the target population, encourage them to start a survey invitation chain-referral process in their community, ensure that every respondent can be traced back along the referral chain. Due to the reliance on respondent referral, RDS is frequently implemented with hidden or rare target populations, where members are assumed to know each other and thus have a higher degree of access than researchers.

We conducted a pilot study in 2023 where we invited 5,000 panel study members to a general population multi-topic online survey. During the survey, we asked respondents whether they would be willing to recruit up to three of their network members. Willing respondents then received personalized links with which to recruit their network members.

We found that younger individuals, individuals with higher incomes, and individuals with migration backgrounds were all (1) more likely to recruit, and (2) they typically recruited individuals similar to themselves. We additionally found that those recruits also had a higher propensity to participate, thereby potentially increasing the overall diversity of the survey. In this paper, we present a detailed overview of our results, along with relevant methodological findings such as method of recruitment. Additionally, we discuss 2025 fielding of RDS in the Social Cohesion Panel which builds upon the findings from our pilot study.

THURSDAY, July 17th

Automatic coding with deep learning: feature extraction from unstructured survey data

Session Organisers Dr Arne Bethmann (SHARE Germany and SHARE Berlin Institute)
Ms Marina Aoki (SHARE Germany and SHARE Berlin Institute)

Time and Room 15:30 - 17:00, Ruppert A - 0.21

Automatic classification of multilingual occupations using transfer learning

Dr Arne Bethmann (SHARE Germany, SHARE Berlin Institute) - Presenting Author
Mr Ming Yuan Lee (SHARE Germany, SHARE Berlin Institute)
Mr Knut Wenzig (SOEP, DIW Berlin)

Socio-economic research often requires detailed information on individual occupations in order to study aspects such as occupational prestige, hazards, or gender disparities. Traditionally, this data collection has involved manual coding of respondents' job descriptions into specific classifications. However, for reasons of efficiency and quality, this process is increasingly being automated using machine learning. While these automated

approaches have shown success with large, high-quality datasets in single languages, multilingual occupation coding remains a challenge, especially for low-resource languages with limited training data.

This paper investigates the potential of transfer learning to improve multilingual occupation coding. We explore how models can leverage the predictive power of other languages within large multilingual datasets. Using extensive training data from the German Socio-Economic Panel (SOEP) and the Survey of Health Aging and Retirement in Europe (SHARE), we fine-tune several pre-trained language models (DistilBERT) to predict one- and four-digit ISCO08 codes, representing simple and complex classification tasks, respectively.

We compare the prediction accuracy of models trained exclusively on country-specific data against those trained on the full multilingual dataset. In addition, we examine the impact of boosting a specific language, in this case German using SOEP data, on the accuracy for this and other countries. Our results aim to demonstrate the effectiveness of transfer learning in improving multilingual occupation classification. Specifically, for cross-country studies such as SHARE, we provide insights into whether and how we can mitigate the challenge of coding occupations with high accuracy in low-resource languages.

FRIDAY, July 18th

New Data Spaces for the Social Sciences - An Interdisciplinary Program for Survey Innovation in Germany

Session Organisers Professor Cordula Artelt (Leibniz Institute for Educational Trajectories)
Dr Anika Schenck-Fontaine (Leibniz Institute for Educational Trajectories)
Professor Corinna Kleinert (Leibniz Institute for Educational Trajectories)

Time and Room 9:00 – 10:30, Ruppert 042

New Data Spaces Cross-National Synergies at the Example of Respondent-Driven Sampling

Dr Carina Cornesse (GESIS - Leibniz Institute for the Social Sciences) - Presenting Author
Dr Jean-Yves Gerlitz (University of Bremen)
Professor Olaf Groh-Samberg (University of Bremen)
Mr Curtis Jessop (The National Centre for Social Research (NatCen))
Professor Olga Maslovskaya (University of Southampton)
Professor Sabine Zinn (German Institute for Economic Research)

The German Infrastructure Priority Program "New Data Spaces for the Social Sciences," funded by the German Research Foundation, was established to facilitate collaboration among research infrastructures, leverage synergies, and drive innovation in the field of data

collection. This initiative shares a similar vision with the UK's "Survey Futures," a program funded by the Economic and Social Research Council to advance survey data collection methods. To propel the development of data collection methodologies, it is essential to expand the perspective on infrastructure innovation from a national to an international level, fostering synergies between countries and infrastructure programs.

This presentation contributes to this dialogue by focusing on two projects that explore an innovative data collection methodology—respondent-driven sampling (RDS)—within the German and UK programs. It will outline the two projects, including the distinct study designs planned for 2025 in both countries, discuss how the projects mutually benefit each other, and explore strategies for strengthening and expanding international collaboration between infrastructure programs. Additionally, the presentation will highlight key similarities and differences between the German and UK infrastructure programs, providing insights from the perspective of researchers funded under these initiatives.

FRIDAY, July 18th

Surveying Ukrainian Refugees in Europe: Implementation, Methods, Experiments, and Challenges 2

Session Organisers Dr Jean Philippe Décieux (University of Bonn and Federal Institute of Population Research (BiB))
Ms Silvia Schwanhäuser (Institute for Employment Research)

Time and Room 11:00 – 12:30, Ruppert 002

(Non-)Probability-Based Sampling of Refugees: A Comparative Analysis of Ukrainian Refugee Samples in Germany

Dr Steffen Pötzschke (GESIS - Leibniz Institute for the Social Sciences) - Presenting Author
Dr Bernd Weiß (GESIS - Leibniz Institute for the Social Sciences)
Dr Andreas Ette (Federal Institute for Population Research)
Dr Manuel Siegert (Federal Office for Migration and Refugees)
Professor Sabine Zinn (German Institute for Economic Research)
Ms Silvia Schwanhäuser (Institute for Employment Research)

Recently arrived refugees are notoriously hard-to-sample and survey. Yet, political realities indicate that the need for research on and with refugees might rather increase than decrease in the foreseeable future. Hence, the question of how survey sampling design might impact sample composition is highly relevant.
In refugee research, non-probability sampling has long been used, for example, due to the absence of reliable sampling frames, budget constraints, the need for quick data availability, or a combination of all these reasons. Despite the inherent analytical limitations of non-probability-based samples such data can provide crucial insights. However, it is important to better understand how such samples might differ from probability-based ones.

Against this background, we focus on comparing the sample composition of a probability-based (n ≈ 11,000) and non-probability-based (n ≈ 9,000) survey of Ukrainian refugees in Germany. While the former stems from a project carried out by a consortium of German research institutes that used register-based sampling (Brücker et al., 2022), the latter comes from a survey whose respondents were recruited through social media advertisements. Both surveys were conducted within a time frame of 10 months. Since the second survey largely replicated the questionnaire of the first, our data sources show a high level of measurement equivalence. We investigate how the samples differ from each other and from the distribution of the target population's key sociodemographic information in administrative data at the respective time of data collection. Furthermore, we will highlight the specific advantages and shortcomings of both sampling approaches.

Effects of Socio-Demographic Composition on Attrition between the First and Second Waves in a Refugee Panel

Mr Johannes Schuett (DIW Berlin e. V. / SOEP) - Presenting Author
Mrs Elena Sommer (DIW Berlin e. V. / SOEP)

Previous research has highlighted attrition as one of the key challenges for probability panels of migrants and refugees (e.g. Jacobsen & Siegert 2023; Lynn et al. 2018; Kalter 2006). Among other factors, attrition can be attributed to residential mobility, language barriers, and the specific socio-demographic composition of migrant and refugee samples. The IAB-BAMF-SOEP Refugee Survey, which has been conducted annually in Germany since 2016, regularly draws refresher samples to account for panel attrition and changes in refugee migration patterns. In 2023, a refresher sample M9 was added to the IAB-BAMF-SOEP Survey of Refugees, consisting of two subsamples – refugees from Ukraine and refugees from other countries. In the first wave, participation was similar in both subsamples among households that had completed a short self-administered contact form before being invited for a face-to-face interview. However, in the second wave in 2024, participation was much higher in the Ukrainian refugee subsample than in the non-Ukrainian refugee subsample, despite the use of the same questionnaires, contact methods and incentive strategy. We expect that differences in the socio-demographic composition of these two subsamples – such as differences in gender and age composition, legal status and labour force participation – account for the differences in attrition between the first and second waves. In our study, we analyse the composition effect on participation in the second wave by considering household, respondent, and interviewer characteristics as well as interview aspects (e.g. change of interviewer, language, and mode preferences). Our study makes an additional contribution to the research on attrition in migrant panels by offering a comparative perspective within the broader refugee population in Germany and going beyond a descriptive comparison of survey participation by analysing the factors relevant to panel attrition across different groups of refugees.