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

10th ESRA Conference
July 17-21, 2023
University of Milan-Bicocca
Milan, Italy

SOEP The Socio-Economic Panel

DIW SOEP

Leibniz Leibniz-Gemeinschaft

TUESDAY, July 18th

Recruiting Web Surveys via Postal-Mail: Best-Practice, Experiments, and Innovation 1

Session Organisers Jean Philippe **Décieux** (Federal Insitute for Population Research)
Carina **Cornesse** (Socio-Economic Panel at DIW Berlin)

Time and Room 11:00 - 12:30, U6-01a

Since e-mail addresses are usually unavailable on standard sampling frames of broader population surveys (e.g., population registers), recruiting high-quality web surveys is challenging. When conducting such large-scale and large-scope web surveys, recruitment and surveying is, therefore, typically conducted in two separate steps: First, a (probability-)sample of the study population is drawn and contacted offline, often during a brief face-to-face or telephone recruitment-interview. Second, members of the sample are asked to switch to the online mode for the actual survey.
Compared to interviewer-administered contact and recruitment, postal-mail strategies are becoming increasingly popular and a large number of cross-sectional as well as longitudinal web survey projects are currently being initiated using postal-mail recruitment in combination with online survey methodology. There are several reasons for this. For example, recruiting web surveys via postal-mail is usually both more time- and cost-efficient than the available alternatives. In addition, this strategy avoids undesirable interviewer effects and allows respondents to read through study and recruitment material at their own speed, time, and convenience.
Currently, the methodology for successful postal-mail recruitment of web surveys is advancing fast. Therefore, this session aims to provide a broad exchange forum for researchers and projects working on and with postal-recruited web surveys. In addition to sharing experiences and best-practices, we are particularly interested in experimental approaches that might include, topics such as:

- Strategies for enabling the transition from offline contact to web data collection mode
- Comparing the success of postal-mail recruitment to other web survey recruitment strategies
- Optimizing initial response, panel consent, and panel registration for postal-mail recruited longitudinal studies
- Push-to-web and other mixed-mode recruitment approaches
- Cost-benefit analyses of different incentive and reminder strategies
- Design and layout effects

Keywords: Web Survey; Recruitment; Mixed-Mode; Survey Costs; Postal Recruitment; Experimental survey research

TUESDAY, July 18th

Perceptions of Inequality and Justice 1

Session Organisers Jule **Adriaans** (Bielefeld University)
Sandra **Bohmann** (Socio-Economic Panel Study at DIW Berlin)
Stefan **Liebig** (Freie Universität Berlin)
Matteo **Targa** (Università di Roma Tre)

Time and Room 11:00 - 12:30, U6-20

Reducing inequalities in life chances and outcomes is identified as one of the key societal challenges of today and the economic turmoil experienced as a consequence of the pandemic as well as the ongoing war in Ukraine have exacerbated questions of social inequality and social justice across Europe.

One important contribution that the social sciences have made and continue to make in this debate, is highlighting the importance of subjective evaluations in understanding the persistence of inequalities as well as the proposed far-reaching consequences for well-being and social cohesion. Research continues to show that individuals misperceive inequality, evaluate inequalities in terms of justice, and hold normative beliefs that legitimize inequalities – all of which help to understand why inequalities persist, why inequalities do not necessarily translate into adverse consequences, and how individuals will react to policies aiming to address inequalities.

National and international surveys offer rich data for studying the determinants and consequences of such subjective perspectives on inequality. For example, Round 9 of the European Social Survey (2018/2019) featured a module on “Justice and Fairness in Europe”, the International Social Survey Programme fielded the fifth iteration of its “Social Inequality” module in 2019, and the German Socio-Economic Panel Study included a questionnaire module on social inequality in 2021. All of which extend beyond a narrow scope of income and wealth inequality but also cover issues of social mobility, preferences for distributive principles, life chances, and political procedural justice, allowing for a comprehensive account of perceptions of inequality and justice.

We are inviting contributions that address methodological questions with respect to survey measures and survey-embedded experiments that capture attitudes towards inequality as well as substantive applications of survey research that shed light on the determinants and consequences of subjective perspectives on inequality and justice.

Keywords: Social inequality; Justice; Perceptions; Survey research

TUESDAY, July 18th

State of the Metadata Infrastructure

Session Organisers Knut **Wenzig** (Socio-Economic Panel at DIW Berlin)
Daniel **Bela** (LIfBi – Leibniz-Institut für Bildungsverläufe)
Arne **Bethmann** (SHARE Germany)

Time and Room 11:00 - 12:30, U6-08

Metadata are at the heart of the movement towards FAIR data (Findable, Accessible, Interoperable, Reusable) and are gaining more and more importance. These metadata need to follow certain standards and need to be collected and managed in appropriate tools throughout the entire survey lifecycle.

Ideally, a data infrastructure can be implemented based on this, which fosters the FAIR principles at all points of data exchange between the involved parties (data producers, providers, archives, users, and other stakeholders):

- Findable: e.g. search portals use standardized metadata to harvest information from data producers and data providers
- Accessible: e.g. data consumers have to be able to access information without human interaction, guided by standardized communications protocols.
- Interoperable: e.g. data users have to be able to understand data and treat and analyze them in an appropriate way
- Reusable: e.g. data is documented in standard, domain relevant way allowing proper secondary data analysis

Metadata are not to be prepared ex-post, but are ideally collected whenever they first appear during the survey life-cycle, e.g. information on the funding institutions during the project proposal phase, data collection protocols and instruments in order to understand the data during survey development, or data alterations during curation. Hence there is a need for proper (meta)data management tools right from the start and through all steps of the process.

This session will discuss contributions to the broader topic of metadata infrastructure within any part of the data lifecycle, and offers space to assess the progress made in this endeavor. We welcome and encourage presentations regarding the implementation of metadata systems–ideally fostering FAIR data provisioning and use–regardless of the systems' maturity.

Keywords: Metadata FAIR

Linked Open Research Data for Social Science – a concept registry for granular data documentation

Jana **Nebelin** (Socio-Economic Panel at DIW Berlin) - Presenting Author
Antonia **May** (GESIS Leibniz Institute for the Social Sciences)
Pascal **Siegers** (GESIS Leibniz Institute for the Social Sciences)
Andreas **Daniel** (Deutsche Zentrum für Hochschul- und Wissenschaftsforschung (DZHW))
Jan **Goebel** (Socio-Economic Panel at DIW Berlin)
Dagmar **Kern** (GESIS Leibniz Institute for the Social Sciences)
Benjamin **Zapilko** (GESIS Leibniz Institute for the Social Sciences)
Fakhri **Momeni** (GESIS Leibniz Institute for the Social Sciences)
Knut **Wenzig** (Socio-Economic Panel at DIW Berlin)

The re-use of research data is an integral part of research practice in the social and economic sciences. To find relevant data, researchers need adequate search facilities. However, a comprehensive, thematic search for research data is difficult because of inconsistent or absent indexing at the social science concept level. Either the data is not documented at a granular level, or primary investigators use their ad-hoc terminology to describe their data. From the user's perspective, the lack of theory language in data documentation impedes effective data searches and thus significantly limits the research potential of existing data collections. Because there is currently no semantic model for indexing the data content, the specific challenge for improving data search lies in establishing concept-based indexing of research data. Research infrastructures need technology for the harmonized semantic indexing of their data. The LORD concept registry aims at closing this gap by developing a registry of sociological and economic concepts and, following the FAIR principles, making this concept registry generally available to the scientific community. As a first step, we developed a basic data model for the Concept Registry using United Modeling Language (UML). All links between are created and managed in the form of so-called RDF triples. An annotation application allows for linking questions/variables to concepts. The application also includes the two SKOS-compliant thesauri, "Thesaurus Social Sciences" (TheSoz) and "Standard Thesaurus Economics" (STW) but could be extended to other resources like ELSST.

We illustrate the application of the LORD concept registry with examples from three large-scale survey programmes (German Socio-Economic Panel, German General Social Survey, National Academics Panel Study). The initial focus is on variables and questions with overlapping content in the three survey programmes, as they form a sound basis for cross-linking with concepts.

State of the DDI Cloud

Knut **Wenzig** (Socio-Economic Panel at DIW Berlin) - Presenting Author
Claudia **Saalbach** (Socio-Economic Panel at DIW Berlin)
Xiaoyao **Han** (Socio-Economic Panel at DIW Berlin)

An investigation was conducted to examine the extent to which metadata in different Data Documentation Initiative (DDI) standards is openly available and which elements of these standards are used. DDI is a set of international standards for describing and documenting data used in social, behavioural, economic, and health sciences research.

To identify the online repositories, where DDI metadata is available, re3data.org (a global registry of research data repositories that covers research repositories from different academic disciplines) and an enquiry on the DDI-users mailing list were used. We compare this with findings from 2017.

Then we tried to access and analyse the metadata, e.g. by using a standardised protocol like the Open Archive Initiative-Protocol for Metadata Harvesting (OAI-PMH). This makes it possible to show which elements are more commonly used than others.

The findings have implications for deploying DDI metadata and the further development of the standards. They could also inform users like researchers and data stewards, how the standards are used by the community. Overall, the investigation highlights the value of openly available metadata in supporting research to achieve the goals of the FAIR data movement.

TUESDAY, July 18th

Recruiting Web Surveys via Postal-Mail: Best-Practice, Experiments, and Innovation 2

Session Organisers Jean Philippe **Décieux** (Federal Insitute for Population Research)
Carina **Cornesse** (Socio-Economic Panel at DIW Berlin)

Time and Room 14:00 - 15:30, U6-01a

Since e-mail addresses are usually unavailable on standard sampling frames of broader population surveys (e.g., population registers), recruiting high-quality web surveys is challenging. When conducting such large-scale and large-scope web surveys, recruitment and surveying is, therefore, typically conducted in two separate steps: First, a (probability-)sample of the study population is drawn and contacted offline, often during a brief face-to-face or telephone recruitment-interview. Second, members of the sample are asked to switch to the online mode for the actual survey.
Compared to interviewer-administered contact and recruitment, postal-mail strategies are

becoming increasingly popular and a large number of cross-sectional as well as longitudinal web survey projects are currently being initiated using postal-mail recruitment in combination with online survey methodology. There are several reasons for this. For example, recruiting web surveys via postal-mail is usually both more time- and cost-efficient than the available alternatives. In addition, this strategy avoids undesirable interviewer effects and allows respondents to read through study and recruitment material at their own speed, time, and convenience.
Currently, the methodology for successful postal-mail recruitment of web surveys is advancing fast. Therefore, this session aims to provide a broad exchange forum for researchers and projects working on and with postal-recruited web surveys. In addition to sharing experiences and best-practices, we are particularly interested in experimental approaches that might include, topics such as:

• Strategies for enabling the transition from offline contact to web data collection mode
• Comparing the success of postal-mail recruitment to other web survey recruitment strategies
• Optimizing initial response, panel consent, and panel registration for postal-mail recruited longitudinal studies
• Push-to-web and other mixed-mode recruitment approaches
• Cost-benefit analyses of different incentive and reminder strategies
• Design and layout effects

Keywords: Web Survey; Recruitment; Mixed-Mode; Survey Costs; Postal Recruitment; Experimental survey research

TUESDAY, July 18th

From Concurrent Mixed-Mode to Push-to-Web: Experimental Design Change in a Panel Survey Study with Postal Mail Recruitment

Carina **Cornesse** (Socio-Economic Panel at DIW Berlin) - Presenting Author
Jean-Yves **Gerlitz** (University of Bremen)
Olaf **Groh-Samberg** (University of Bremen)

Self-administered panel survey studies where sample members are contacted via postal mail often apply mixed-mode designs with online and paper survey mode options. Via the online mode, data can be collected fast, at low cost, and with high data quality. Via the paper mode, biases due to population members who are unable or unwilling to provide survey data online can be prevented. Researchers usually hope that the vast majority of respondents will choose the online survey mode. The challenge is to encourage respondents to do so without alienating non-internet users and people reluctant to provide data online.

Most experimental research on how to mix modes in surveys with postal mail contact focuses on cross-sectional surveys or on early panel recruitment stages and suggests that response rates are higher and biases lower if the online and paper mode are offered concurrently in the postal mail invitation. However, once trust between respondents and researchers has been established in a panel study, it may be beneficial to switch to offering only the online mode option in the survey invitation letter and leaving the paper mode option for the reminder letters.

We examine the impact of such a design change in the newly established German Social Cohesion Panel. In our experiment, a random subgroup of the panel sample was switched from concurrent mixed-mode to a sequential "push-to-web" design in the first regular survey wave after panel recruitment while the rest of the panel sample remained on the concurrent design for one more panel survey wave. Preliminary results show that the share of online respondents is much higher in the sequential mode design while survey response rate differences between the experimental groups are marginal, thus indicating that changing the mode sequence offered in the postal mail letters of.

TUESDAY, July 18th

Perceptions of Inequality and Justice 2

Session Organisers Jule **Adriaans** (Bielefeld University)
Sandra **Bohmann** (Socio-Economic Panel Study at DIW Berlin)
Stefan **Liebig** (Freie Universität Berlin)
Matteo **Targa** (Università di Roma Tre)

Time and Room 14:00 - 15:30, U6-08

Reducing inequalities in life chances and outcomes is identified as one of the key societal challenges of today and the economic turmoil experienced as a consequence of the pandemic as well as the ongoing war in Ukraine have exacerbated questions of social inequality and social justice across Europe.

One important contribution that the social sciences have made and continue to make in this debate, is highlighting the importance of subjective evaluations in understanding the persistence of inequalities as well as the proposed far-reaching consequences for well-being and social cohesion. Research continues to show that individuals misperceive inequality, evaluate inequalities in terms of justice, and hold normative beliefs that legitimize inequalities – all of which help to understand why inequalities persist, why inequalities do not necessarily translate into adverse consequences, and how individuals will react to policies aiming to address inequalities.

National and international surveys offer rich data for studying the determinants and consequences of such subjective perspectives on inequality. For example, Round 9 of the European Social Survey (2018/2019) featured a module on “Justice and Fairness in Europe”, the International Social Survey Programme fielded the fifth iteration of its “Social Inequality” module in 2019, and the German Socio-Economic Panel Study included a questionnaire module on social inequality in 2021. All of which extend beyond a narrow scope of income and wealth inequality but also cover issues of social mobility, preferences for distributive principles, life chances, and political procedural justice, allowing for a comprehensive account of perceptions of inequality and justice.

We are inviting contributions that address methodological questions with respect to survey measures and survey-embedded experiments that capture attitudes towards inequality as well as substantive applications of survey research that shed light on the determinants and consequences of subjective perspectives on inequality and justice.

Keywords: Social inequality; Justice; Perceptions; Survey research

TUESDAY, July 18th

Agility and the Survey Life-Cycle - If and what survey practitioners can learn from software development 2

Session Organisers Yuri **Pettinicchi** (SHARE Berlin Institute)
Arne **Bethmann** (TU Munich / SHARE Germany)

Time and Room 14:00 - 15:30, U6-06

GitLab at the Socio-Economic Panel

Knut **Wenzig** (Socio-Economic Panel at DIW Berlin) - Presenting Author

The Socio-Economic Panel (SOEP) relies heavily on a GitLab server for its data management and documentation needs. GitLab is a web-based Git repository manager that provides version control for source code, project management tools, and continuous integration. At the SOEP, the GitLab server has been configured to provide several key features, including:

- Version control: GitLab provides a centralised repository for storing and managing source code, making it easier to track changes and collaborate with other team members. The SOEP uses this feature to manage scripts (Stata and R) and metadata.
- Issue tracking: GitLab's built-in issue tracker allows users to report and track bugs, defects, and other issues. At the SOEP, the issue tracker is used to report problems during data preparation, document issues, and track bug reports from data users. It is also used to

manage teams or projects, complex internal processes with multiple stakeholders, and to organise weekly meetings.
- Service desks: GitLab's service desks feature enables users to create and manage support tickets. This feature can be used to centralise and distribute user requests to the relevant specialists, replacing email battles within the team.
- Pipeline: GitLab's pipeline feature allows the SOEP to automate the testing and deployment of metadata stored in CSV files. This feature has also been used to produce almost publication-ready PDF files.
- Wiki: GitLab's built-in wiki feature allows users to create and edit pages of content within the GitLab interface. The SOEP uses this feature to create internal documentation related to its data management activities.

Overall, the GitLab server at the SOEP provides a range of features that support the organisation's research and data management needs. By using GitLab, the SOEP has been able to improve collaboration and streamline workflows.

WEDNESDAY, July 19th

Item Nonresponse and Unit Nonresponse in Panel Studies 2

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

Time and Room 11:00 - 12:30, U6-21

A New Home in Times of Crisis: Changing the Survey Institute in the Pandemic

Felix **Süttmann** (Socio-Economic Panel at DIW Berlin) - Presenting Author
Sabine **Zinn** (Socio-Economic Panel at DIW Berlin)

After nearly 40 years and 40 waves, in 2021, the German Socio-Economic Panel (SOEP) changed the survey institute from Kantar Public to infas. In addition to this transition, the COVID-19 pandemic hindered the usual survey mode of SOEP, i.e., computer-assisted personal interviewing (CAPI). Both issues meant substantial challenges and drastic changes to the SOEP wave 2021. First and foremost, a considerably postponed field start, mixed-mode designs on the level of household members, new interviewers, and more telephone interviews than intended. All in all, the changes affected SOEP response rates to such an extent that frequent interventions were required during the field period. Nonetheless, nonresponse increased to previously unobserved levels. Our aim is to quantify and explain this increase.

First, we will present the changes due to the new field institute and interventions during the field period. This is followed by models to detect socio-economic groups of survey members that had the highest risk of nonresponse. We will also try to disentangle effects of the pandemic and the changed survey institute. The analysis differentiates between the mostly German samples and those of refugees and migrants. We see that existing factors compound and identify new ones associated with COVID-19 and the field institute change. As lessons learned, we formulate practical suggestions for survey institute changes.

WEDNESDAY, July 19th

Developments in survey methods and analysis about LGBTI+ populations 2

Session Organiser Angelo **Moretti** (Utrecht University)

Time and Room 14:00 - 15:00, U6-23

„Free Expression of One's Own Personality?" The Role of Gender Identity and Gender Expression in Poverty Risks.

David **Kasprowski** (Socio-Economic Panel at DIW Berlin, BGSS) - Presenting Author

Mostly, gender is treated as a control variable or an important interaction without any theorization. In addition to strong economic differences based on gender, such as the persistent gender pay gap to the disadvantage of women, it is still not specified which aspects of gender may contribute to disadvantages. However, recent research suggests that self-identified gender non-conforming individuals experience distinctive economic penalties. Based on a German online-survey, this contribution compares the multidimensional assessment of gender with a large sample of 6,956 LGBTI* individuals, including 1,131 transgender, gender non-conforming and non-binary individuals. Beyond the question of the poverty risks of trans* and non-binary individuals compared to cisgender women and men, I particularly address the role of gender expression and perceived gender on social stratification. However, what happens when gender identity and expression do not match and people are ascribed a different gender identity than they have chosen for themselves? Does gender (non) conforming presentation possibly have stronger negative consequences than the gender identity itself? To what extent do common gender theories need to be adapted accordingly or which aspects need to be examined more clearly if we want to understand gendered differences and ultimately reduce discrimination? First results show that especially self-identified non-binary and gender-non-conforming people have to face the highest poverty risk factors like insecure or inadequate housing conditions. The results will be discussed with regard to the theoretical deduction of gender in quantitative research and which possibly false conclusions are drawn when gender is applied as a fixed entity of two genders for analyses without a critical evaluation.

WEDNESDAY, July 19th

Approximating Probability Samples in the Absence of Sampling Frames 1

Session Organisers Carina **Cornesse** (Socio-Economic Panel at DIW Berlin)
Mariel **McKone Leonard** (DeZIM Institute)

Time and Room 14:00 - 15:00, U6-01f

Research shows that survey samples should be constructed using probability sampling approaches to allow valid inference to the intended target population. However, for many populations of interest high-quality probability sampling frames do not exist. This is particularly true for marginalized and hidden populations, including ethnic, religious, and sexual minorities. In the absence of sampling frames, researchers are faced with the choice to discard their research questions or to try to draw inferences from nonprobability and other less conventional samples.

For the latter, both model-based and design-based solutions have been proposed in recent years. This session focuses on data collection techniques designed to result in samples that approximate probability samples. We also invite proposals on techniques for approximating probability samples using already collected nonprobability sample data as well as by combining probability and nonprobability sample data for drawing inferences. The session scope covers but is not limited to research on hard-to-reach and hard-to-survey populations. We are particularly interested in methodological research on techniques such as

- Respondent-driven sampling (RDS) & other network sampling techniques
- Quasi-experimental research designs
- Weighting approaches for nonprobability data (especially those that make use of probability sample reference survey data)
- Techniques for combining probability and nonprobability samples (e.g. blended calibration)

Keywords: nonprobability sample, respondent-driven sampling, blended calibration, weighting, data integration

WEDNESDAY, July 19th

Falsification detection in times of crisis: Challenges, opportunities, and new directions 1

Session Organisers Markus **Bönisch** (Statistics Austria)
Eduard **Stöger** (Statistics Austria)

Time and Room 14:00 - 15:00, U6-11

Detecting falsified interviews in a longitudinal survey

Andreas **Franken** (Socio-Economic Panel Study at DIW Berlin) - Presenting Author

With Covid-19, regulations and social behavior effects increased the burdens for interviewers. For example, interviewees were less willing to let the interviewer into their home, because they were afraid to become infected. Even more so, for some time it was prohibited to let non-household members in one's apartment. This increased pressure on interviewers resulted in more divergences from the formal survey process, ranging from skipping items or questions to falsifying whole interviews.

The German Socio-Economic Panel (GSOEP) is a longitudinal survey with a repeating questionnaire on household and individual level, which started in 1984. It is mainly based on face-to-face interviews and has routines to detect striking, potentially falsified interviews. These routines are built on classic statistic indicators for detecting falsified interviews, such as analysis of the Benford distribution or time stamps. To cover a broader range of indicators for detecting falsified interviews, several supervised and unsupervised machine learning algorithms were tested. For the training and testing of these algorithms, over 500 interviews detected and validated as falsified were used. To build the features of the algorithm, different kinds of data were prepared, i.a. paradata (like time stamps), data of the response behavior (like usage of filter questions, or the Benford distribution) and interviewer information (like the region where the interviewer is operating).

Preliminary results show that it is possible to predict falsified interviews with machine learning. Machine learning approaches thus can facilitate both validation and identification of falsified interviews in future survey research. Various issues remain for discussion, amongst others the small number of validated falsifications and falsifying interviewers, the difficulty to differentiate between a bad quality interview and a real fraud, or the changing amount of data over the years.

WEDNESDAY, July 19th

Approximating Probability Samples in the Absence of Sampling Frames 2

Session Organisers Carina **Cornesse** (Socio-Economic Panel at DIW Berlin)
Mariel **McKone Leonard** (DeZIM Institute)

Time and Room 16:00 - 17:30, U6-07

Research shows that survey samples should be constructed using probability sampling approaches to allow valid inference to the intended target population. However, for many populations of interest high-quality probability sampling frames do not exist. This is particularly true for marginalized and hidden populations, including ethnic, religious, and sexual minorities. In the absence of sampling frames, researchers are faced with the choice to discard their research questions or to try to draw inferences from nonprobability and other less conventional samples.

For the latter, both model-based and design-based solutions have been proposed in recent years. This session focuses on data collection techniques designed to result in samples that approximate probability samples. We also invite proposals on techniques for approximating probability samples using already collected nonprobability sample data as well as by combining probability and nonprobability sample data for drawing inferences. The session scope covers but is not limited to research on hard-to-reach and hard-to-survey populations. We are particularly interested in methodological research on techniques such as

- Respondent-driven sampling (RDS) & other network sampling techniques
- Quasi-experimental research designs
- Weighting approaches for nonprobability data (especially those that make use of probability sample reference survey data)
- Techniques for combining probability and nonprobability samples (e.g. blended calibration)

Keywords: nonprobability sample, respondent-driven sampling, blended calibration, weighting, data integration

THURSDAY, July 20th

Surveying Ukrainian Refugees in Europe: Implementation, Methods, Challenges, and Exchange of Experiences 1

Session Organisers Jean Philippe **Décieux** (Federal Institute for Population Research)
Silvia **Schwanhäuser** (Institute for Employment Research)

Time and Room 09:00 - 10:30, U6-01a

Probability Sampling for a Study of Ukrainian Refugees in Germany

Hans Walter **Steinhauer** (Socio-Economic Panel at DIW Berlin) - Presenting Author
Jean Philippe **Décieux** (Federal Institute for Population Research)
Andreas **Ette** (Federal Institute for Population Research)
Manuel **Siegert** (Federal Office for Migration and Refugees)
Sabine **Zinn** (Socio-Economic Panel at DIW Berlin)

After Russia's invasion of Ukraine in early 2022 millions of Ukrainians fled their country. By now more than one million of these have found refuge in Germany. Integrating such a large number of refugees is challenging. Although having learned from the refugee crisis in 2015 and after, this influx of Ukrainian refugees differs in many aspects concerning demographics, perceived acceptance, allocation, and migration flows. In order to learn about the recently arriving refugees, their needs, and resources, as well as challenges ahead, a survey allowing for generalization to this population was urgently needed. To ease this need quickly, we developed a novel sampling strategy to create a random sample for the population of Ukrainian refugees using of two different registers; namely the central register for foreigners (AZR) and the local residents register (EMR). Being able to access both registers allows for profiting from each registers' advantages, while compensating their disadvantages. Having access to both registers, we were able to show that both consistently cover the population of Ukrainian refugees. Using information from the AZR allowed us to sample 100 municipalities at the first stage based on the number of Ukrainian refuges. Within each sampled municipality all refugees aged 18 to 70 were listed with their address from the corresponding EMR at the second stage. This resulted in a list consisting of addresses to sample from at the second stage. To minimize the risk of sampling multiple refugees from the same household we sorted the list by address and family name and implemented a systematic random sampling at the second stage. This sampling design results in an initial sample of 48,000 individuals in regions with a high density of Ukrainian refuges while also mapping the characteristics of the population.

THURSDAY, July 20th

Opportunities and Challenges in Dealing with Selection Bias in Cross-sectional and Longitudinal Surveys 1

Session Organiser Sabine **Zinn** (Socio-Economic Panel at DIW Berlin)
Jason M. **Fields** (U.S. Census Bureau)
Hans Walter **Steinhauer** (Socio-Economic Panel at DIW Berlin)

Time and Room 09:00 - 10:30, U6-02

Analysing survey data usually also means coping with selection bias. There are proven and well-established strategies for doing so, such as survey weighting or selection modelling. However, still many data users struggle in understanding how to apply these strategies, especially when confronted with the diversity of the information given by the survey providers. Beyond that, increasingly researchers use machine learning and Bayesian statistics in survey data analysis. This is also true for conducting and controlling surveys. Specifically, adaptive contact or motivational strategies are designed for upcoming survey studies or waves based on response processes observed in previous surveys or survey waves. The estimation of population statistics is improved by including information about the entire selection process in the statistical model, both developing these methods and communicating their use are critical.

In this session, we welcome research on novel approaches and strategies to ease data users understanding of how to handle selection bias in their statistical analysis. This research might cover:

-Methods for easing, and communicating, the appropriate use of weights or other methods for addressing selection biases in published microdata files. These may include, but are not limited to, longitudinal weights, calendar year weights, replicate weights, multiple implicates, and other tools to improve the population representativeness and communication of uncertainty in public data products.
-Novel methods to assess and adjust for sources of bias in cross-sectional and longitudinal surveys, including, but not limited to, machine learning interventions, adaptive design, post-hoc weighting calibrations, informed sampling, etc. How are these communicated to data users? How are they adapted as response and biases change?
-Papers are encouraged that investigate the selection processes, papers that leverage novel modelling strategies for coping with selection bias in statistical analysis, and papers that include examples of modelling non-ignorable selection bias in substantive analysis.

Keywords: Selection bias, weighting, adaptive designs, non-ignorable selection, weighting

THURSDAY, July 20th

Opportunities and Challenges in Dealing with Selection Bias in Cross-sectional and Longitudinal Surveys 2

Session Organisers Sabine **Zinn** (Socio-Economic Panel at DIW Berlin)
Jason M. **Fields** (U.S. Census Bureau)
Hans Walter **Steinhauer** (Socio-Economic Panel at DIW Berlin)

Time and Room 14:00 - 15:30, U6-07

Analysing survey data usually also means coping with selection bias. There are proven and well-established strategies for doing so, such as survey weighting or selection modelling. However, still many data users struggle in understanding how to apply these strategies, especially when confronted with the diversity of the information given by the survey providers. Beyond that, increasingly researchers use machine learning and Bayesian statistics in survey data analysis. This is also true for conducting and controlling surveys. Specifically, adaptive contact or motivational strategies are designed for upcoming survey studies or waves based on response processes observed in previous surveys or survey waves. The estimation of population statistics is improved by including information about the entire selection process in the statistical model, both developing these methods and communicating their use are critical.

In this session, we welcome research on novel approaches and strategies to ease data users understanding of how to handle selection bias in their statistical analysis. This research might cover:

-Methods for easing, and communicating, the appropriate use of weights or other methods for addressing selection biases in published microdata files. These may include, but are not limited to, longitudinal weights, calendar year weights, replicate weights, multiple implicates, and other tools to improve the population representativeness and communication of uncertainty in public data products.
-Novel methods to assess and adjust for sources of bias in cross-sectional and longitudinal surveys, including, but not limited to, machine learning interventions, adaptive design, post-hoc weighting calibrations, informed sampling, etc. How are these communicated to data users? How are they adapted as response and biases change?
-Papers are encouraged that investigate the selection processes, papers that leverage novel modelling strategies for coping with selection bias in statistical analysis, and papers that include examples of modelling non-ignorable selection bias in substantive analysis.

Keywords: Selection bias, weighting, adaptive designs, non-ignorable selection, weighting

Validating an index of selection bias for proportions in non-probability samples

Angelina **Hammon** (Socio-Economic Study at DIW Berlin) - Presenting Author

The increasing use of alternative non-probabilistic data collection strategies in survey research demands methods for assessing the sensitivity of respective population estimates. For this purpose, Andridge et al (2019) propose an index to quantify potential (non-ignorable) selection bias in proportions. We validate this index with an artificial non-probability sample generated from a large empirical data set and additionally applied it to proportions estimated from data on current political attitudes arising from a real non-probability sample selected via River sampling. When the requirements of the index are fulfilled, it shows an overall good performance in detecting and correcting present selection bias in estimated proportions, and thus provides a powerful measure for evaluating the robustness of results obtained from non-probability samples.

THURSDAY, July 20th

Approximating Probability Samples in the Absence of Sampling Frames 3

Session Organisers Carina **Cornesse** (Socio-Economic Panel at DIW Berlin)
Mariel **McKone Leonard** (DeZIM Institute)

Time and Room 14:00 - 15:30, Room U6-22

Research shows that survey samples should be constructed using probability sampling approaches to allow valid inference to the intended target population. However, for many populations of interest high-quality probability sampling frames do not exist. This is particularly true for marginalized and hidden populations, including ethnic, religious, and sexual minorities. In the absence of sampling frames, researchers are faced with the choice to discard their research questions or to try to draw inferences from nonprobability and other less conventional samples.

For the latter, both model-based and design-based solutions have been proposed in recent years. This session focuses on data collection techniques designed to result in samples that approximate probability samples. We also invite proposals on techniques for approximating probability samples using already collected nonprobability sample data as well as by combining probability and nonprobability sample data for drawing inferences. The session scope covers but is not limited to research on hard-to-reach and hard-to-survey populations. We are particularly interested in methodological research on techniques such as

- Respondent-driven sampling (RDS) & other network sampling techniques
- Quasi-experimental research designs
- Weighting approaches for nonprobability data (especially those that make use of probability sample reference survey data)
- Techniques for combining probability and nonprobability samples (e.g. blended calibration)

Keywords: nonprobability sample, respondent-driven sampling, blended calibration, weighting, data integration

The Potential of Respondent-Driven Sampling (RDS) in Survey Practice: Who is Willing to Recruit?

Carina **Cornesse** (Socio-Economic Panel at DIW Berlin) - Presenting Author
Jean-Yves **Gerlitz** (University of Bremen)
Olaf **Groh-Samberg** (University of Bremen)
Sabine **Zinn** (Socio-Economic Panel at DIW Berlin)

RDS is a popular network sampling strategy with many advantages. Its general idea is that researchers can select a set of survey respondents (the so-called “seeds”) and encourage them to recruit some of their social network members to the survey, who then again recruit some of their network members, and so on (the so-called “referral chain”). The desired result is a large and diverse dataset in which each person can be traced back to their initial seed.

In theory, RDS has the capability of generating samples which allow valid inference to a researcher’s population of interest without requiring traditional sampling frames. In practice, however, the methodology often fails to generate large and diverse enough datasets to be able to do so. One common challenge is that many selected seeds do not even start recruiting network members, so that referral chains are not initiated. This has particularly been observed for research contexts in which seeds are selected from probability sample surveys and/or where the goal is to draw inferences to a broad and heterogeneous population.

Our study explores the potential of RDS procedures based on general population probability-based survey samples. Questions we address include: To what extent are survey respondents willing to recruit their social network members? And how do survey respondents who are willing to do so differ from everyone else? To answer these questions, we selected a random subset of respondents from the newly established German Social Cohesion Panel (n≈1,600; stratified by age) and asked them for their hypothetical willingness to recruit 3 members of their social network for a survey. Preliminary analyses suggest that for some general population subgroups (e.g. younger adults) RDS may be quite promising while other subgroups (e.g. older adults) are either undecided or hesitant to engage in RDS.

FRIDAY, July 21st

Tools and program developments for data analysis

Session Organiser Xiaoyao **Han** (Socio-Economic Panel at DIW Berlin)

Time and Room 09:00 - 10:30, U6-07

The session “Tools and Program Developments for Data Analysis” will cover the latest tools and programs required for effective data analysis. Having the right tools and programs is critical to the success of data analysis. By streamlining the analysis process, ensuring accuracy, and making it easier to collaborate, data analysis tools can help organizations and individual users make better use of their data and gain valuable insights. The session will discuss various tools for data analysis, highlighting their benefits in terms of improving efficiency, accuracy, scalability, visualization, and FAIR principles. Several tools will be introduced: including a) Open Data Format that includes enriched metadata and can be used across various software programs, b) Amnesia data anonymization tool that facilitates the trusted sharing of research data while protecting privacy, c) methodological approaches and data visualization tools focusing on producing and disseminating model based early estimates of key health outcomes, d) SurveyHarmonies, a tool that enables the creation of ex-ante harmonized, multi-language surveys using reproducible research tools compliant with the DDI standards in the R statistical environment, e) visualizing survey data using hammock plots. These tools are developed for various domains and aim to analyse datasets, automate tasks, and visualize results more efficiently. The session will provide a valuable opportunity to communicate the latest developments in data analysis tools and programs.

Keywords: FAIR data, data visualization, data analysis tools

A metadata enriched Open Data Format across Statistical Programs

Claudia **Saalbach** (Socio-Economic Panel at DIW Berlin) - Presenting Author
Xiaoyao **Han** (Socio-Economic Panel at DIW Berlin) - Presenting Author
Knut **Wenzig** (Socio-Economic Panel at DIW Berlin)

In the social sciences, a diversity of statistical software is used for data processing and analysis. These software programs have specific data formats and handle metadata in different ways. Proprietary data and a variety of data formats that are only partially compatible present obstacles to data reuse by researchers. In particular, proprietary data formats undermine the requirement for interoperability embodied in the FAIR principles. To address this problem, this paper proposes the concept of an open data format, which is intended to facilitate data dissemination in social sciences and to support data analysis

across software. The open data format features multilingual metadata and external links to data portals that allow direct access to online documented material through the statistical software itself. In addition, with technical support, the data format can be loaded and manipulated in popular statistical software, at the same time the metadata can be fully transferred and used. This paper begins by describing the specification of the open data format including data and metadata, which lays the foundation for the potential subsequent use of metadata and the programs based on it. In addition, we explore several technical implementations of statistical software in which data formats can be imported and manipulated together with metadata.

FRIDAY, July 21st

Linking Survey Data with Geospatial Data: Potentials, Methods, and Challenges 2

Session Organisers Simon **Kühne** (Bielefeld University)
Dorian **Tsolak** (Bielefeld University)

Time and Room 09:00 - 10:30, U6-21

Getting personal – Moving from regional to ego-centered provider density to capture healthcare availability in survey data

Barbara **Stacherl** (Socio-Economic Panel at DIW Berlin) - Presenting Author

Regional provider densities, e.g., physician-to-population ratios in administrative regions, are frequently used to depict healthcare availability. Research using survey data also relies on regional provider densities to account for healthcare infrastructure as part of the regional context. However, regional provider densities are based on administrative areas, ignoring border-crossing for healthcare use. Therefore, individualized spatial approaches for measuring healthcare availability in survey research are needed. To capture the healthcare infrastructure (physicians and hospitals) available to individuals, I generate ego-centered provider densities for households in the German Socio-Economic Panel (SOEP), a representative longitudinal survey. Variation across individuals and over time as well as intraregional variation are analyzed. To outline the information gain achieved through ego-centered measures, they are compared to standard regional provider densities. Geocoded address data for all hospitals (2003-2019) and outpatient physicians (2009-2019) in Germany were linked to SOEP data. Using georeferenced population data (100x100m grid), ego-centered provider-to-population ratios, namely the number of physicians and hospital beds per 100,000 inhabitants within a 10km radius, were computed for each household for each year. For comparison, physician and hospital densities at district level (2015-2019) were

linked to the SOEP. Although relatively stable over time, some external (=among non-movers) within-individual changes in ego-centered physician and hospital density were observed. For all years, higher variability was observed for ego-centered measures compared to regional measures. There was substantial variation in ego-centered provider density within the administrative regions at which standard densities are aggregated – in 2019, 61% and 57% of the variation in the ego-centered physician density the hospital density, respectively, remained unexplained when accounting for district-level grouping. Using individualized spatial measures – such as ego-centered provider densities – entails large potential to inform survey research.