PAN-Metrics
Seminar on Novel Measurement Problems in Social Sciences: Response Styles, Careless Responding, Measurement Invariance, Log-Data, and Technology-Based Measurement

May 15-17, 2024
Warsaw, Poland

1 Conference venue

The meeting will be held at:

Institute of Philosophy and Sociology of the Polish Academy of Sciences

Staszic Palace
The Erazm Majewski Hall (3rd floor)

72 Nowy Świat Street
Warsaw, Poland

Staszic Palace is easy accessible by public transport: using buses (bus stop “Uniwersytet”) or subway (“Nowy Świat-Uniwersytet” station of the M2 line).

To get to the Erazm Majewski Hall, please use the main entrance to the building, which is located on the north side, just behind the Nicolaus Copernicus Monument and take the elevator on the right of the main stairways to the 3rd floor.

It is also possible to reach the 3rd floor by stairs, but you will need to use the staircase to the right of the main stairways, as the main stairways only lead to the 1st floor.

2 Timetable

May 15 (Wednesday)

Time	Presenting author & Title
10:00	Opening and welcome address Prof. dr hab. Andrzej Rychard – Director of the Institute of Philosophy and Sociology, Polish Academy of Sciences Prof. IFiS dr hab. Artur Pokropek – Chair of the Organising Committee
10:30	State-of-the art speaker I: Esther Ulitzsch *Confirmatory mixture models for investigating careless and insufficient effort responding in ecological momentary assessments*
11:30	Coffee break
12:00	Session I: *Response processes in assessments*. Chair: Ulf Kroehne Janine Buchholz (Institute for Educational Quality Improvement): *The effect of adaptive testing on engagement in PISA* Jana Welling (Educational Measurement, Leibniz Institute for Educational Trajectories): *How homogenous is the test-taking process? The effect of rereading the text on item difficulties and test performance in a reading comprehension* Augustin Mutak (Freie Universität Berlin): *Empirical validation study of the intra-individual speed-ability relationship (ISAR) mode* Michalis Michaelidis (University of Cyprus): *Rapid guessing behavior through item response times in international large-scale assessments and consequences for country rankings*
14:00	Lunch break
15:00	Session II: *Response styles & log-data*. Chair: Esther Ulitzsch Oliwia Szczupska (SWPS University, Warsaw, Poland): *Rating scale effects in measuring voters' ideological and partisan attitudes* Artur Pokropek (Institute of Philosophy and Sociology of the PAS) *Behind Computer-Based Assessments: How Paradata Unveils Respondent Characteristics* Tomasz Żółtak (Institute of Philosophy and Sociology of the PAS): *Loglime: Workflow for Log-data Collection and Processing* Marek Muszyński (Institute of Philosophy and Sociology of the PAS): *Employing log-data indices to understand response styles in questionnaire data*
17:00	Coffee break
17:30	State-of-the art speaker II: Francesca Borgonovi TBA
18:30	End of first day Evening event

May 16 (Thursday)

Time	Author(s) & Title
10:00	State-of-the-art speaker III: Ulf Kroehne *Design of computer-based assessment and the interpretation of log data*
11:00	Coffee break
11:30	Session III: *Developing new measurement instruments*. Chair: Francesca Borgonovi Katarzyna Chyl (International Studies Unit, Eucational Research Institute, Warsaw): *Making the test: Polish Adult Literacy Assessment* Magdalena Pokropek (Doctoral School of Social Sciences, University of Warsaw): *Building a test of critical thinking in new media environment* Christoph Jindra (Institute for Educational Quality Improvement Humboldt Universität zu Berlin): *Effects of Private Tutoring in Year 7 on Later Math Competencies and Grades – Observational Evidence Based on Targeted Maximum Likelihood Estimation, OLS, and G-computation*
13:00	Lunch break
14:00	State-of-the-art speaker IV: Alexander Robitzsch *Why Full, Partial, or Approximate Measurement Invariance Are Not a Prerequisite for Meaningful and Valid Group Comparisons*
15:00	Discussion Panel
16:00	Coffee break
16:30	Session IV *Machine learning.* Chair: Alexander Robitzsch Andrzej Jarynowski (University of Warsaw; Freie Universität Berlin): *Methodological implications of single-response declarative vs complex annotative language use on the example of Ukrainian refugees in Poland* Kristoph Schumann (Institute for Educational Quality Improvement, Humboldt-Universität zu Berlin, Germany): *Handling Large-Scale-Assessments with Prediction Rule Ensembles. On the Advantages of Machine Learning for Characterizing At-Risk Students in the IQB-Bildungstrend 2021* Karoline A. Sachse (Institute for Educational Quality Improvement, Humboldt-Universität zu Berlin, Germany): *Interpreting Random Forests Using Partial Dependence Plots. An Illustration Using a Teacher-Questionnaire-Dataset from the Pandemic Era* Hubert Plisiecki (Institute of Psychology of the Polish Academy of Sciences): *Modeling Emotion Intensity in Political Texts: A Comparison of Supervised Methods and Annotation with Popular LLMs*
18:30	End of second day Evening event

May 17 (Friday)

Time	Author(s) & Title
10:00	State-of-the-art-speaker V: Oliver Luedtke: TBA
11:00	Coffee break
11:30	Session V: *Survey methodology.* Chair: Oliver Luedtke Sara Bojarczuk (University of Warsaw): *The importance of data collection mode in studying discrimination on the Polish rental housing market* Barbara Jancewicz (University of Warsaw): *Enriching surveys with environmental data: the case of a Thermosurvey* Barbara Kowalczyk (SGH Warsaw School of Economics): *Robustness of models for a sensitive latent variable used in Item Count Techniques* Michał Taracha (SGH Warsaw School of Economics): *Modelling complex associations between personality traits and COVID-19 vaccination decisions* Michał Bojanowski (Universitat Autònoma de Barcelona, Kozminski University): *Analyzing the Efficacy of Modeling Adjustments in Network Scale-Up Method for Degree Estimation*
14:00	End of third day

3 Abstracts of presentations

The effect of adaptive testing on engagement in PISA

Janine Buchholz (Institute for Educational Quality Improvement)

The core feature of adaptive testing consists of an increased match between student performance and test difficulty, leading to gains in measurement precision. Additional benefits for students' test-taking experience are implicitly assumed, but prior research is scarce and failed to produce evidence in favor of such effects, possibly as a result of the design of those studies which are typically based on comparisons between adaptive and linear versions of the test. This study draws on a unique feature of the adaptive design implemented in PISA 2018 which allows for the direct examination of the effect of the performance-difficulty match on engagement by means of intentionally misrouting students to a mismatching testlet. In addition, this study examines the effect for PISA 2022 which implemented an alternative design, administering either an adaptive or linear test, the latter of which can be considered a milder version of misrouting.

Analyses were conducted conditional on students' performance level and show differential effects: while high-performing students benefit from a mismatching (i.e., too easy) test, low-performing students show lower engagement in a mismatching (i.e., too hard) test, both in the current and in a subsequently administered testlet. Jointly, the findings demonstrate the beneficial effects of administering (too) easy items on test engagement for students at both ends of the proficiency distribution, and the potential harm caused by administering too difficult items to low-performing students.

The findings are discussed in the context of large-scale assessments for which student engagement is of particular concern due to their low-stakes nature.

How homogenous is the test-taking process? The effect of rereading the text on item difficulties and test performance in a reading comprehension

Jana Welling (Educational Measurement, Leibniz Institute for Educational Trajectories),
Esther Ulitzsch (University of Oslo),
Gabriel Nagy (Educational Measurement and Data Science, Leibniz Institute for Science and Mathematics Education)

Using the novel potential provided by process data, recent studies have shown that the prevalent assumption of a homogenous test-taking process might not hold, with test-takers varying in their test-taking engagement and strategies. Therefore, the aim of the present study was to investigate the effect of rereading a text in a reading comprehension test on the item and person parameters. We specified three item response mixture models that distinguish on the response level between the three latent classes rapid guessing, solution behavior with text rereads and solution behavior without text rereads. The different models assumed either (1) equal item parameters, (2) equal item discriminations but varying item difficulties, or (3) varying item parameters between the two different solution behavior classes. In a reading comprehension test of the German National Educational Panel Study (N = 1933 students, 14 multiple-choice items), the second model with equal item discriminations but varying item difficulties between the two latent classes fitted the data best. Descriptive analysis revealed that the reread class did not differ extensively from the no reread class in the average item difficulty, but rather exhibited less variation in the item difficulties, with hard items becoming easier and easy items becoming harder. Furthermore, the tendency to reread the text was slightly positively correlated with test performance. The results suggest that the test-taking process might indeed be more heterogenous than often expected, as rereading the text in reading comprehension tests seems to affect both the item difficulties and test performance.

Empirical validation study of the intra-individual speed-ability relationship (ISAR) mode

Augustin Mutak (Freie Universität Berlin),
Sören Much (Martin-Luther Universität Halle-Wittenberg),
Jochen Ranger (Martin-Luther Universität Halle-Wittenberg),
Steffi Pohl (Freie Universität Berlin)

The speed-ability trade-off (SAT) is a phenomenon very well documented in psychological research which describes a drop in the performance in a task due to an increase in speed. Recently, a novel psychometric model, the ISAR model (Mutak et al., in press) has been developed to capture the SAT. The model is an extension of van der Linden's (2007) hierarchical speed-accuracy model and it introduces growth terms for speed and ability, which in turn allows us to examine the correlation between the intra-individual change in speed and ability, which should reflect the speed-ability trade-off. However, the interpretation of this correlation as SAT can be questionable. There are other factors, such as concentration and motivation, which can change over the course of the test and confound the intra-individual relationship of speed and ability. For the purpose of disentangling these different effects, we devised an empirical study. The study included a matrix reasoning test as the main measure and concentration and motivation as confounder measures taken at several timepoints. In data analyses, we statistically control for the changes in concentration and motivation in order to examine how they impact the intra-individual speed-ability relationship and in order to gain a less biased estimate of SAT. Through this, we aim to validate the interpretation of the model and its parameters. This study can also serve as a basis for future research which relies on the non-stationarity of ability and speed in studying SAT.

Rapid guessing behavior through item response times in international large-scale assessments and consequences for country rankings

Michalis Michaelidis (University of Cyprus),
Militsa Ivanova (University of Cyprus),
Demetris Avraam (University of Liverpool)

When test-takers do not invest adequate effort in low-stakes assessments, test scores underestimate the individual's true ability, and ignoring the impact of test-taking effort may harm the validity of test outcomes. Using item response times from digital assessments, this study examined examinees' rapid guessing behavior and accuracy in the Programme for International Student Assessment (PISA) across countries and different item types. The 2015 PISA computerized assessment was administered in 59 jurisdictions. Behavioral measures of students' test-taking effort were constructed for the Science, Mathematics and Reading assessments by applying a fixed and a normative threshold on item response times to identify rapid guessing. The proportion of rapid guessers on each item was found to be small on average, but variable within and between countries. Average performance for rapid guessers was on average much lower than for test-takers engaged in solution behavior for all types of items. Weighted response time effort indicators by country were very high, and positively correlated with country mean PISA score. When filtering out examinees who were identified as rapid guessers, country mean scores improved, however the impact on country rankings was in general minor, if any. Computerized assessment programs may monitor rapid guessing behavior to identify cross-country differences prior to comparisons of performance and for developing interventions to promote engagement with the assessment.

Rating scale effects in measuring voters' ideological and partisan attitudes

Oliwia Szczupska (SWPS University, Warsaw, Poland),
Mikołaj Cześnik (SWPS University, Warsaw, Poland),
Maciej Sychowiec (SWPS University, Warsaw, Poland)

In web surveys, political ideology and partisan attitudes are typically captured using rating scales, which can vary in terms of labels, length, and format. In an experiment conducted alongside the 2023 Polish National Election Study, we investigated how different labeling approaches affect response styles and survey results.

First, we studied party evaluation on like-dislike scales with different numerical values. The results show that [-5, ..., 5] and [0, ..., 10] rating scales are not linearly equivalent. There was a certain bias against assigning negative grades on a 0-midpoint scale and high positive grades on a 5-midpoint scale. It was present at the aggregate level, across all parties (not uniformly), and when accounted for voters' party identification. Both extreme and neutral responses remained largely invariant with respect to labels.

Second, we studied ideological self-assessment on sliders with or without numerical labels. The score distributions did not vary significantly. Respondents presented with [0-Left, ..., 10-Right] scale were more likely to select the default midpoint, while respondents facing a continuous left-right spectrum opted for "Don't know" more often, suggesting a perceived interchangeability between these two responses.

Overall, the distortions observed due to different rating scales affected the accuracy of the quantitative results but did not have a strong impact on their qualitative interpretation. Our study contributes to the discussion on how survey questions' design can evoke heterogeneous attitudes depending on the implemented labeling approach.

Behind Computer-Based Assessments: How Paradata Unveils Respondent Characteristics

Artur Pokropek (Institute of Philosophy and Sociology of the Polish Academy of Sciences)

In the era of computer-based assessments, analyzing process data like sequences of actions or mouse cursor movements reveals itself as a valuable tool for studying respondent behavior, providing information about their engagement and allowing for a deeper understanding of their profile. This presentation will demonstrate how various process data can be used in research applications. We show that process data combined with deep learning techniques such as Gated Recurrent Units (GRU) and Bidirectional Long Short-Term Memory (BiLSTM) could be used to predict a range of respondent characteristics such as their engagement, cognitive abilities, and demographic traits like gender and age with high accuracy. In this study, we will focus on two types of paradata: mouse cursor movements in survey research and sequential action data in studies of student skills, using the PISA assessment as an example.

Loglime: Workflow for Log-data Collection and Processing

Tomasz Żółtak (Institute of Philosophy and Sociology of the Polish Academy of Sciences)

Para-data could give important insight into how respondents answer questions in self-report assessments: surveys and questionnaires. Technical capabilities of data collection in computer-based assessments were noticed fairly quickly. However, none of the early proposals on how computer-based paradata should be collected was embraced by researchers, probably because of a rather difficult implementation, requiring basic knowledge of programming and web interface. In this presentation, a framework that allows for easy implementation of log-data collection within LimeSurvey open web-surveying platform is presented (also, possibilities to adapt it to other web-surveying platforms will be shortly discussed). Moreover, an accompanying R package that enables easy data transformations and calculating a wide group of response process indicators (Goldhammer et al., 2021; Kroehne & Goldhammer, 2018) is also presented. Various response time indicators, cursor/mouse moves/trajectories and velocity, and hovering (Horwitz et al., 2017; 2020) are among process indicators that can be obtained by processing log-data in this package.

Employing log-data indices to understand response styles in questionnaire data

Marek Muszyński (Institute of Philosophy and Sociology of the Polish Academy of Sciences

The presentation aims to broaden knowledge on using log data to understand response styles, which are one of main threats to validity and comparability of self-report results. The log data is understood here as a kind of additional survey information (paradata) possible to collect in computer-based modes.

For example, the presentation intends to verify the inverted-U effect (e.g. Akrami et al., 2007) that assumes that response time decreases when distance between trait level and item difficulty is increasing, in other words – when responses are easy and clear-cut, fast responses are expected. Indeed, such a pattern was often identified in the data with extreme categories noting the fastest response times and the middle categories noting the slowest ones (Casey & Tryon, 2001; Kroehne & Goldhammer, 2018; Kulas & Stachowski, 2009). Such a result could be explained in the frames of a well developed self-knowledge (e.g. self-schemata theory; Markus, 1977) but it can be also a result of response styles and/or careless straightlining as extreme categories may be responded to quickly because at least some of the participants who have endorsed them are straightliners (Leiner, 2019) or subjects responding stylistically (Henninger & Plieninger, 2020). Such analyses often bring surprising patterns, e.g. that high levels of extreme response style are related with longer, not shorter, response times (Henninger & Plieninger, 2020). We will try to replicate the result that responses are faster when the response category matches individual response style (Henninger & Plieninger, 2020).

Apart from response times, I will employ other paradata as well, mostly mouse tracking information. Mouse trajectories will be used as a pseudo-eye-tracker in order to replicate results from the genuine eye-tracking studies (e.g. Kaminska & Foulsham, 2013; Muller et al., 2017) that linked extreme response style with fixations mainly on extreme response options, less organised way of reading items and lower number of total fixations. Here I will try to replace fixations with mouse moves, especially with hovering (periods without movement) and moves such as underlining items' text/question stem (indicating reading) or moves/hovers around specific response categories.

I also relate mouse measures such as: speed (velocity – number of pixels travelled per second, and acceleration), flips (changes in movement direction) or distance (total distance travelled by the cursor) to response styles parameters as estimated by multidimensional IRTree models.

Making the test: Polish Adult Literacy Assessment

Katarzyna Chyl (International Studies Unit, Eucational Research Institute, Warsaw),
Artur Pokropek (International Studies Unit, Eucational Research Institute, Warsaw; Institute of Philosophy and Sociology, Polish Academy of Sciences, Warsaw)

In my presentation, I will discuss preparing, piloting, and normalizing the Polish Adult Literacy Assessment (PALA; working title). PALA measures predominantly functional reading skills, providing real-life reading materials, such as pamphlets and advertisements, and multiple-choice comprehension questions. We aimed to develop a computer-based test that would accurately and reliably assess functional reading in adults, especially at the lower level. Almost one in five Poles have serious problems with reading, but until now, there were no functional literacy tests available in Polish.

In the three-stage pilot study, we conducted cognitive labs, supervised online testing, and unsupervised online testing. In total, 131 participants took part in the quantitative assessment. We also conducted passageless test taking with 32 participants, which ensured us that the selected questions could not be answered with only background knowledge. In the analysis, we utilized classical test theory and IRT modeling. After an initial screening, we were left with a set of 100 questions. Several other questions have borderline values of the IRT discrimination parameter (alpha < 0.6) or a low correlation between the correct answer and the estimated skill level in IRT (rpb.WLE < 0.3). We finally selected 12 texts with the best psychometric properties and had at least 3 questions per booklet (total of 41).

Currently, we almost collected the planned 400 datasets coming from the representative sample (education x age). In May, we will already have test norms and will be able to share our tool with the researchers interested in adult literacy.

Building a test of critical thinking in new media environment

Magdalena Pokropek (Doctoral School of Social Sciences, University of Warsaw)

Critical thinking in the context of new media has become incredibly important in the era of misinformation. Therefore, its fostering is crucial, but in order to develop critical online media literacy, its measurement must first be conducted. Previous measurements of constructs similar to critical thinking in new media environment have primarily relied on questionnaire-based research, which has its inherent limitations.

I have designed a cognitive test specifically tailored to assess critical thinking skills in the context of new media platforms, incorporating original content that students may encounter. The core part of the test consists of 24 tasks measuring knowledge, skills, and attitudes in the following areas: Understanding the sender's intentions (explicit and implicit); Identifying bias; Distinguishing between facts and opinions; Evaluating the strength and quality of evidence; Motivation to verify; Responsible sharing of information. The whole test is supposed to last no longer than 40 minutes. I have prepared two versions of the test (A and B), totaling 44 different tasks, but some of them appear in both version A and version B. The test is to be conducted in schools on the Lime Survey platform. In the presentation I will introduce the test.

Twenty cognitive interviews were conducted, providing the first proof of the validity. Following this, a quantitative pilot study was conducted (sample size 100 students). In the presentation I will share the results of the cognitive interviews and the initial findings from the quantitative validation.

Effects of Private Tutoring in Year 7 on Later Math Competencies and Grades – Observational Evidence Based on Targeted Maximum Likelihood Estimation, OLS, and G-computation

Christoph Jindra (Institute for Educational Quality Improvement Humboldt Universität zu Berlin),
Karoline A. Sachse (Institute for Educational Quality Improvement Humboldt Universität zu Berlin)

State-of-the-art causal inference methods for observational data have emerged as powerful tools, offering the potential to relax assumptions that risk compromising the validity of causal inferences. One such method, Targeted Maximum Likelihood Estimation (TMLE), stands out as a semiparametric, doubly-robust approach capable of yielding unbiased estimates when either the outcome or treatment model is correctly specified. The method distinguishes itself from conventional approaches by mitigating the risk of misspecification bias, allowing the incorporation of (nonparametric) machine learning techniques, including super learning, to estimate relevant segments of the distribution.

In this study, we leverage TMLE to estimate the effects of receiving private tutoring in mathematics during Grade 7 on two distinct outcomes—math competencies and math grades—using observational data obtained from NEPS SC3 (N = 4,167). Comparisons are drawn between TMLE estimates and those derived from Ordinary Least Squares (OLS) and G-computation. Our findings reveal nearly identical estimates across the methods when eval-uating end-of-year grades. However, variations emerge when examining math competencies as the outcome, underscoring that substantive conclusions may be contingent on the chosen analytical approach.

This research underscores the significance of employing advanced causal inference methods, such as TMLE, when navigating the complexities of observational data and highlights the nuanced impact of methodological choices on the interpretation of study outcomes.

Methodological implications of single-response declarative vs complex annotative language use on the example of Ukrainian refugees in Poland

Andrzej Jarynowski (University of Warsaw; Freie Universität Berlin),
Karolina Czopek (University of Warsaw),
Andrea Palmini (Freie Universität Berlin),
Alexander Semenov (University of Florida),
Vitaly Belik (Freie Universität Berlin),
Michał B. Paradowski (University of Warsaw)

We investigate peer learner networks of 251 Ukrainian refugees enrolled in an intensive course of the Polish language from a hospitable city in Northwestern Poland. Apart from the special situational context, together with the close typological similarity between the languages spoken and being acquired, the students present a unique language constellation profiles, with different degrees of dominance in Russian vs Ukrainian and complicated attitudes to the latter.

We apply Social Network Analysis, Predictive Modelling and Time Series Analysis to discover patterns of social network behaviour and language use. 40% of speakers whose dominant language was indicated as Russian tend to declare Ukrainian as their first language. Our findings also reveal that linguistic background, particularly the use of Ukrainian, significantly correlates with learners' position within the peer networks as well as with integration into host society. This behavior is reflected in the network structure, where Ukrainian speakers demonstrate higher centrality, suggesting potential linguistic segregation and marginalization among Russian-dominant speakers.

Using regression and random forest models, the study also identifies key predictors of language learning success among this cohort. The frequency of interaction with Russian-speaking friends emerges as a significant factor decreasing self-perceived progress, particularly in grammar. We subsequently employ segmented regression to detect two distinct phases of language acquisition depending on time elapsed since arrival in the host country.

Our analysis suggests that using direct questions via surveys or interviews in emotionally loaded topics is often of limited utility, and that only data collected via logs, diaries, approaches such as experience sampling method, or objective tracing of activity can provide more reliable results.

Handling Large-Scale-Assessments with Prediction Rule Ensembles. On the Advantages of Machine Learning for Characterizing At-Risk Students in the IQB-Bildungstrend 2021

Kristoph Schumann (Institute for Educational Quality Improvement, Humboldt-Universität zu Berlin, Germany),
Karoline A. Sachse (Institute for Educational Quality Improvement, Humboldt-Universität zu Berlin, Germany)

Educational data mining refers to the use of machine learning methods on educational data sets. Unlike other machine learning methods, the goal is not so much to predict new cases as accurate as possible, but rather the data-driven identification of associations between the predictors. Prediction Rule Ensembles (PRE, Fokkema, 2020) are an innovative tree-based machine learning method aiming at a balance between interpretability and accuracy by combing the rules from a random-forest-like approach with linear predictors in a lasso-regularized (logistic) regression.

Using data from the nationally representative German National Trends in Student Achievement Study 2021 (IQB-Bildungstrend 2021; Stanat et al., 2022), this presentation focuses on the characterization of at-risk students, who fail to meet the minimum standards (in mathematics this concerns about one in five children, Schumann & Sachse, 2022). The focus is on the questions of (a) which combinations of variables are most important for predicting failure to meet the minimum standards and (b) the extent to which PRE provides a better predictive performance than other tree-based methods and a regression analysis.

First results of the characterization of the N = 24,500 cases using PRE highlight the importance of cultural capital operationalized through books at home and motivational characteristics like self-concept in mathematics as well as subject-related anxiety for failing to meet the minimum standards in mathematics. A comparative analysis of predictive accuracies shows a significantly lower error loss for PRE compared to lasso-regularized logistic regression ($p < .001$) and random forests ($p < .001$).

Interpreting Random Forests Using Partial Dependence Plots. An Illustration Using a Teacher-Questionnaire-Dataset from the Pandemic Era

Karoline A. Sachse (Institute for Educational Quality Improvement, Humboldt-Universität zu Berlin, Germany),
Kristoph Schumann (Institute for Educational Quality Improvement, Humboldt-Universität zu Berlin, Germany)

Random Forests, a powerful ensemble learning algorithm, has gained widespread popularity for its exceptional predictive accuracy across diverse domains. Its ability to aggregate predictions from multiple decision trees makes it robust and resilient to overfitting, resulting in reliable models. However, this predictive prowess comes at a cost – the inherent challenge of interpretability. While Random Forests excel in making accurate predictions, understanding the intricate decision-making process within the ensemble can be elusive.

We administered a survey consisting of 47 questions on personal information and distance teaching during the early months of the pandemic era to teachers of all federal states in Germany. Responses were collected from an online survey conducted in June and July 2020. Complete questionnaires were submitted by 857 elementary school teachers and 1,590 secondary school teachers. Due to the novel situation of distance teaching during school closures, we did not use existing questionnaires, but rather collected questions adapted to the new context. In the need for more explorative data analysis strategies, we ran Random Forest analyses.

After achieving good predictive accuracy we aimed for interpretability. Calculating variable importance measures is a common initial step to shed light on how much each predictor variable is contributing to the model's predictive performance. Additional avenues to enhance further the interpretability of Random Forests are partial dependence plots, illustrating the relationship between a selected variable and the model's predicted outcome while keeping other variables constant. Is this enough for understanding data? We critically discuss this analysis strategy as well as alternatives.

Modeling Emotion Intensity in Political Texts: A Comparison of Supervised Methods and Annotation with Popular LLMs

Hubert Plisiecki (Institute of Psychology of the Polish Academy of Sciences)

In this study, we analyzed 10,000 texts from the Polish section of portal X (formerly known as Twitter), concerning political topics. These texts were examined for seven emotional categories: five basic emotions (happiness, sadness, anger, disgust, fear) and two emotional dimensions (positivity and tension), aiming to compare the effectiveness of supervised machine learning models with leading language models (LLM) in the market. A key aspect of our analysis was the assessment of emotion intensity, where twenty annotators rated each of the 10,000 texts using a five-point Likert scale. This methodology allowed us to account for a broader spectrum of emotional fluctuations in the analyzed texts. During the model training process, we utilized two different Polish-language models based on the transformer architecture, conducting extensive searches for optimal parameters. This study includes a performance comparison of the selected model with the “gpt-3.5-turbo-1106" and GPT-4 Turbo (“gpt-4-0125-preview”) models. After determining the most suitable "multiple shot" configuration on the validation set using the “gpt-3.5-turbo-1106" model, we proceeded to compare on the test set with both LLM models. The results indicate that while supervised models still seem to be the most optimal choice for predicting emotion intensity, the difference in prediction quality is small.

The importance of data collection mode in studying discrimination on the Polish rental housing market

Sara Bojarczuk (University of Warsaw),
Barbara Jancewicz (University of Warsaw)

Many rental market discrimination studies take advantage of online advertising websites and the option to contact landlords via email to simplify and standardise data collection. A recent study by Antfolk et al. (2019) conducted in England and Poland follows this trend by sending standardised emails that request viewing and signal potential renters’ ethnicity to landlords. However, in many countries, including Poland, email is not the typical way to contact landlords and real estate agents, limiting the representativeness of such studies to only those subjects that actively use email in their renting. Our advertisement analysis and qualitative interviews with landlords and real estate agents confirmed that this mode of research could impact results, distorting the discrimination level found. Therefore in our research, we used the method of contact preferred by agents and landlords: telephone calls. However, we included a second mode of data collection to test whether we could substitute the highly labour intensive and individualised calling (testers have different voices, speech mannerisms etc.), with a more standardizable and partially automatable one: text messaging. We received mixed results with both methods showing some level of discrimination, but also both facing different challenges and introducing biases. We will showcase and discuss our results in hope of improving the methodology of further studies.

Enriching surveys with environmental data: the case of a Thermosurvey

Barbara Jancewicz (University of Warsaw)

More and more people are present online making internet mediated research methods more and more representative (at least in theory). This process coupled with the already existing advantages of online research such as low cost and standardised interview made them even more successful. In person interviews however still hold the advantage of physicality, being then and there with the respondent. This, while introducing the interviewer’s bias, opens the data gathering process to inclusion of further elements. Traditionally interviewers were asked to for example: evaluate apartment condition, number of books. Now with technological progress and device minimization we can include more measurements into the interview, enabling new insights that were previously inaccessible.

The Thermosurvey, a study of older adults in Warsaw and Madrid, is one example of such enrichment. In the study, computer assisted personal interviews, which already included the GPS tracking, were further expanded by temperature and humidity measurements before and during the interview. While the measurements’ aim was mainly to evaluate respondents' apartments' vulnerability to heat, the temperatures proved to correlate with some heat related answers, showcasing how interview environments can be treated both as a bias and as a result in itself. Since the survey also contained a set of questions relating to thermal comfort and thermal experiences, our results show how adding a relatively simple measurement to a survey, enables not only to focus on respondent’s declared answers, but also on the ways they embody their environment.

Robustness of models for a sensitive latent variable used in Item Count Techniques

Barbara Kowalczyk (SGH Warsaw School of Economics),
Robert Wieczorkowski (Statistics Poland)

Item count techniques (ICTs) are statistical methods of indirect questioning that are broadly used by applied researchers in surveys with sensitive questions, i.e. questions about stigmatizing, socially unaccepted or illegal attributes and behaviors. These techniques require the use of some control variables while the variable under study is not directly observable (it remains latent). For many years in research practice to estimate the unconditional probability of possessing the sensitive attribute moment-based estimators were widely used. However, in the modern statistical methodology of the item count techniques the problem is treated as a problem of incomplete data and therefore maximum likelihood (ML) estimators via expectation-maximization (EM) algorithm are employed to address the latent variables. This parametric approach has many advantages in terms of estimation. However, it also introduces some new problems to item count models regarding theoretical assumptions about distribution of the control variable and implementation of ML estimation via EM algorithm. In the presentation we analyze the problem of robustness of various item count models to different violations in data distribution. We conduct a comprehensive Monte Carlo simulation study and examine the consequences of violations of theoretical assumptions in the modelling of the latent variable in item count models.

Modelling complex associations between personality traits and COVID-19 vaccination decisions

Michał Taracha (SGH Warsaw School of Economics)

2021 was a year of mass distribution of COVID-19 vaccines. Despite widespread availability of vaccines, the level of vaccination of the European population varied widely. This study aims to analyse the association between personality traits and COVID-19 vaccination attitudes among elderly residents of Central and Eastern European countries.

We performed the disaggregation of 11 analysed countries into 41 regions in a unique way: based on 2 modules of the SHARE database (Housing Generated-Variable and Retrospective Accommodation modules), as well as linguistic differences – for Baltic countries. The micro-level data was obtained from SHARE, whereas macro-level data, including excess mortality, was retrieved from Eurostat and Quality of Government datasets. We prepared 6 maps in the descriptive part of the analysis.

In order to account for the nuanced association between personality traits and vaccination attitudes, Bayesian model averaging was performed. Posterior inclusion probabilities indicated the order in which variables were added in the hierarchical regression models explaining COVID-19 vaccination willingness. Additionally, logistic regression models with clustered standard errors were computed – to account for heteroskedasticity across regions.

Vaccination willingness increases with age and education, is higher for respondents living with a partner in the household, in urban areas, and with multimorbidity. Low neuroticism, low conscientiousness, low extraversion and to a lower extent high openness were found to be the most relevant for vaccination willingness. For men, neuroticism seems particularly significant. Meanwhile, for females, it is difficult to find a significant personality trait once environmental variables are included (openness being slightly more significant for females).

Analyzing the Efficacy of Modeling Adjustments in Network Scale-Up Method for Degree Estimation

Michał Bojanowski (Universitat Autònoma de Barcelona, Kozminski University),
Miranda Lubbers (Universitat Autònoma de Barcelona)

The Network Scale-Up Method (NSUM) was developed to estimate sizes of hard-to-reach populations. It relies on respondents’ knowledge of particular sub-populations and their personal networks. The method originally treated the size of the sub-population as unknown, but follow-up iterations have reversed this, focusing instead on estimating the network degree. However, this method faces challenges due to assumptions about respondent accuracy and random mixing, leading to biases in degree estimates. This study assesses the success of modeling adjustments in correcting effects and biases in degree estimation based on Aggregated Relational Data (ARD). We investigate various corrective strategies designed to deal with the impacts of barrier effects, transmission bias, and recall bias. Once the data have been collected a researcher is restricted to modeling strategies in the hope for correcting above mentioned effects and biases. In this study we apply different statistical models to ARD from questionnaire items on first names and on social positions and evaluate their effectiveness. We consider the following methods:

- Generalized linear model by Zheng, Salganik, and Gelman (2006) with barrier effects
- Bayesian model by Maltiel et al. (2015)
- Dichotomous approaches of Baum and Marsden (2023)
- Different re-scaling approaches to correcting recall bias.

These models were applied to survey data collected in Spain in 2021, involving 1,500 adults with questions about 20 first names and 16 social positions. We observe that degree distributions obtained from first names ARD are similar across the methods. However, distributions from social positions ARD varied significantly depending on the applied method. The barrier effects model (BEM) applied to occupations ARD yields results more similar to those from first names ARD than the MLE model did. The MLE method, when applied to social positions ARD, yields degree estimates that are an order of magnitude smaller than those obtained from first names ARD. However, adoption of more complex models for ARD results in significant improvements in the accuracy of degree estimates based on social positions, but substantial discrepancies remain.

PAN
POLISH ACADEMY of SCIENCES

IFiS PAN

PAN-METRICS