Measurement Error

Frühjahrstagung der DGS-Sektion „Methoden der empirischen Sozialforschung“

28./29. April 2017, Mannheim

Lokale Organisation

gesis Leibniz Institute for the Social Sciences

UNIVERSITÄT MANNHEIM

“Measurement Error” ist ein Kernelement des Total Survey Error Ansatzes. Gemeint ist damit die (mangelnde) Qualität der Schlussfolgerung von empirischen Umfragedaten auf den untersuchten theoretischen Gegenstand oder die untersuchte theoretische Beziehung zwischen Merkmalen. Zusammen mit der Generalisierungsmöglichkeit auf die Zielpopulation beschreibt der „Measurement Error“ bzw. die Messqualität die Güte der Ergebnisse einer empirischen Untersuchung. Während sich die Methodenforschung in den letzten Jahrzehnten vermehrt der Problematik der Sicherung der Güte von Stichproben und Reduzierung von Non-Response widmete und hier internationale Standards entwickelte und etablierte, wurden Fragen der Messqualität eher vernachlässigt. Die Tagung stellt die Qualitätsaspekte der Messung in Umfragen gemäß dem Ansatz des Total Survey Error in den Fokus. Enthalten sind daher Beiträge, die die Messqualität der Daten und Methoden ihrer Bestimmung adressieren. Zentrale Themen sind Validitätsbestimmung und –sicherung sowie Bestimmung und Minimierung systematischer und unsystematischer Messfehler. Die Messinvarianz und Antworteffekte in vergleichenden und Mixed-Mode-Untersuchungen, Interviewereinflüsse auf das Antwortverhalten und Codierfehler (processing error) sind weitere in Bezug auf Measurement Error relevante Themen. Von besonderem Interesse sind Arbeiten, die unterschiedliche Fehlerquellen adressieren und die Messqualität ganzheitlich betrachten.

Anmeldung zu der Tagung:

http://www.gesis.org/unser-angebot/veranstaltungen/gesis-tagungen/fruehjahrstagung-methodensektion-dgs/

Anmeldeschluss: 31.03.2017

Programm

Freitag, den 28. April

10:45 – 11:00 Begrüßung durch lokale Organisation und Sektionsleitung

11:00 – 12:00 **Invited Speaker: Randall K. Thomas**
In Search of the Ideal Response Format: Does Traditional Psychometric Theory Impede Good Measurement in the Smartphone Era?

12:00 – 13:00 Mittagspause

13:00 – 14:25 **Session 1: Ansätze zur Modellierung und Erklärung des Messfehlers**

- Optimierung der Messgenauigkeit und Testlänge beim adaptiven Testen. (*S. Otto, M. Merten)*
- Can We Estimate True Criterion Scale Validity in Longitudinal Studies? (*T. Raykov)*
- Quellen systematischer und „zufälliger" Messfehler – eine Mixed-Methods Untersuchung zur Messqualität von Skalen zur Erfassung von Religiosität. (*U. Kelle, B. Langfeldt, B. Metje)*
- Can Satisficing Explain Measurement Errors? Using Difficulty, Ability, and Motivation to Explain Measurement Errors in Pensions Due to Old Age. (*P. Lazarevic*)

14:25 – 15:00 Pause

15:00 – 16:00 **Session 2: Aufnahmevorträge I**

- Satisficing in probability and nonprobability online panels. (*C. Cornesse)*
- The Dynamics of neither/nor Answers in Attitudinal Questions. (*M. Trübner)*
- Exploring the influence of respondents' IT affinity on nonresponse in an online survey. (*J. Herzing)*

16:00 – 16:30 Pause

16:30 – 17:30 **Session 3: Aufnahmevorträge II**

- Unit Nonresponse and Measurement Error – Evidence from a Probability-Based Panel. (*R. Bauer, K. Weyandt, B. Weiß, M. Bosnjak)*
- Interviewer Effects on Multiple Sources of Survey Error. (*D. Ackermann-Piek)*
- Modellierung der zukünftigen Integrationsentwicklung der dritten Migrantengeneration mithilfe einer dynamischen Mikrosimulation. (*D. Bekalarczyk)*

17:30 – 17:45 Pause

17:45 – 19:00 Mitgliederversammlung

20:00 Gemeinsames Abendessen in der Nähe des Tagungsortes

Samstag, den 29. April

9:00 – 10:45 **Session 4: Survey and Questionnaire Design: Effekte auf den Messfehler**

- Wie beeinflussen die Polarität von Ratingskalen und die Verbalisierung der Skalenmitte die Messqualität? *(N. Menold)*
- Do distractions during web survey completion affect data quality? Findings from a laboratory experiment. *(A. Wenz)*
- Der Einfluss des Survey-Modes auf die Datenqualität bei sensiblen Fragestellung am Beispiel von Sexualviktimisierung. *(H. Leitgöb, N. Leitgöb-Guzy)*
- Neuere Weiterentwicklungen der Item-Count-Technik (ICT) zur Vermeidung von Response Bias bei „heiklen Fragen“ in Surveys. *(F. Wolter)*
- “Justice“ and “Fairness“: A different meaning to different people? *(M. Weinhardt, J. Adriaans)*

10:45 – 11:00 Pause

11:00 – 12:40 **Session 5: Interviewereffekte auf die Messung**

- Interviewer Effects as Measurement Errors in a Face-to-Face Survey on Financial Behaviour. Evidence from Zambia. (*M. Fröhlich, F. Kreuter, P. L. Nguyen)*
- Befragte, die antworten ohne sich Gedanken über ihre Antworten zu machen, und Interviewer, die Teile ihrer Interviews fälschen. *(J. Blasius)*
- What interviewer characteristics get the most out of the respondents? An analysis of interviewer effects on the item nonresponse rate in asset questions based on German SHARE data. *(S. Friedel)*
- Interviewer Effects and the Measurement of Financial Literacy. (T. Schmidt, *T. F. Crossley, P. Tzamourani, J. K. Winter)*
- A potential pitfall in modelling the link between education and social network quality. *(M. Trebbels, P. Winker)*

12:40 Verabschiedung

13:00 Ende der Tagung

Lokales Organisationsteam

Natalja Menold: natalja.menold@gesis.org
Silke Schneider: silke.schneider@gesis.org
Beatrice Rammstedt: beatrice.rammstedt@gesis.org
Cornelia Zuell: cornelia.zuell@gesis.org

Veranstaltungsort

Universität Mannheim
Schloss Mannheim
O 138 Fuchs-Petrolub-Festsaal (Ostflügel)
68159 Mannheim

Anreise: https://www.uni-mannheim.de/1/service/anfahrt_lageplan/

Hotelinformation

Ein Kontingent ist beim Wyndham Mannheim Hotel bis 15.03.2017 verfügbar. Kosten EZ 92,00 Euro; DZ 102,00 Euro. Buchen Sie bitte unter dem Stichwort „Tagung Methodensektion“.

Wyndham Mannheim Hotel
F4, 4-11
68159 Mannheim
Tel.: 0621-150393-0

Abstracts

Invited Speaker

In Search of the Ideal Response Format: Does Traditional Psychometric Theory Impede Good Measurement in the Smartphone Era?

Randall K. Thomas (GfK, Research Methods, Public Communications and Social Sciences; Randall.Thomas@gfk.com)

In this presentation, we review the nature of reliability and validity in measurement from both psychometric and sociometric perspectives. With respondents increasingly taking our surveys online using smartphone devices, respondents are pressuring us to reduce both the number of items and the number of responses that we use to measure our ideas of interest. We will look at some of the implications of slimming down surveys in terms of both traditional and new measures of reliability and validity.

Session 1: Ansätze zur Modellierung und Erklärung des Messfehlers

Optimierung der Messgenauigkeit und Testlänge beim adaptiven Testen

Siegmar Otto (siegmar.otto@ovgu.de)
Martin Merten

Unter Messen wird in der Psychologie heute unter anderem das Ausfüllen eines Fragebogens verstanden. Der Fragebogen dient als Messinstrument, mit dem, wie mit einem Thermometer Körpertemperatur, Eigenschaften wie Intelligenz oder Prosozialität gemessen werden (Eid & Schmidt, 2014). Allerdings gibt es in der Qualität verschiedener Messungen einen Unterschied, der sich daran beurteilen lässt, wie hoch der Anteil der Varianz in der Messung ist, der auf Messfehlern beruht. Und genau hier hat die Psychologie ein Problem, aufgrund dessen sie von Wissenschaftlern anderer Disziplinen gern belächelt wird: Die Messgenauigkeit eines Fragebogens ist noch weit entfernt von der eines Thermometers, da der Messfehleranteil selbst bei den bestkonstruierten Fragebögen noch um ein Vielfaches höher ist als bei einem handelsüblichen Thermometer. Mit diesem Mangel an Messgenauigkeit bei Fragebögen und wie man ihm entgegenwirken kann, beschäftigt sich dieser Beitrag. Die Messgenauigkeit eines Fragebogens kann verbessert werden, indem die Anzahl der Fragen erhöht wird. Aus ökonomischen Gesichtspunkten, zum Beispiel bei der Datenerhebung in Panels, ist allerdings eher eine Reduktion der Itemzahl gewünscht. Deshalb bleibt zur Steigerung der Messgenauigkeit nur die Alternative, Items mit dem bestmöglichen Informationsgehalt – und zwar zugeschnitten auf jede einzelne befragte Person – einzusetzen. Dies kann mit adaptivem Testen erreicht werden.

Die vorliegende Studie beschäftigt sich deshalb mit dem Thema des adaptiven Testens und der Steigerung der Messgenauigkeit. Hierfür wurde per Computersimulation anhand eines Datensatzes mit N = 787 Personen und einem Pool von 98 Items zur Umwelteinstellungen untersucht, wie viele Fragen sich ohne Genauigkeitsverlust einsparen ließen, wenn man die Fragen per computergestütztem adaptiven Test (CAT) auswählen würde. Für die Simulation basierend auf dem existierenden Datensatz wurde ein Maximum-Likelihood-Algorithmus mit vorgegebener Testlänge verwendet, wobei die Testlänge systematisch variiert wurde. Als Maßstäbe für die Verbesserungen wurden die Separationsreliabilität, Fit-Werte und die Korrelation mit dem Gesamttest berechnet.

Bereits mit etwa 15 adaptiv ausgewählten Fragen konnten die Voraussetzungen des Rasch-Modells erfüllt und die Reliabilität gegenüber einer klassischen nicht adaptiven 50-Item Version von r = .74 auf r = .88 gesteigert werden. Das überraschende positive Phänomen einer Steigerung der Reliabilität bei gleichzeitiger Verkürzung der Testlänge wird im Vortrag erläutert und damit verbundene potenzielle Probleme eines kurzen adaptiven Tests werden diskutiert.

Can We Estimate True Criterion Scale Validity in Longitudinal Studies?

Tenko Raykov (Michigan State University; raykov@msu.edu)

Longitudinal studies in sociology offer unique opportunities to identify the specificity variance in the components of a psychometric scale that is administered repeatedly. Continuing the earlier tradition in quantitative sociology, this article discusses a procedure for evaluation of the relationship between true scale scores and criterion variables uncorrelated with measurement errors in longitudinally presented measures that comprise unidimensional

multi-component instruments. The outlined approach provides point and interval estimates of the true scale criterion validity with respect to a criterion that is assessed once or repeatedly, as well as a means for testing temporal stability in this validity. The discussed method is based on an application of the latent variable modeling methodology, is readily applicable with popular software, and is illustrated using empirical data.

Quellen systematischer und „zufälliger“ Messfehler – eine Mixed-Methods-Untersuchung zur Messqualität von Skalen zur Erfassung von Religiosität

Udo Kelle (Arbeitsbereich Methoden der empirischen Sozialforschung und Statistik, Helmut-Schmidt-Universität/Universität der Bundeswehr Hamburg; kelle@hsu-hh.de)
Bettina Langfeldt (Helmut-Schmidt-Universität/Universität der Bundeswehr Hamburg)
Brigitte Metje (Helmut-Schmidt-Universität/Universität der Bundeswehr Hamburg)

Wir möchten in unserem Beitrag zeigen, wie unterschiedliche Quellen des measurement error im Rahmen eines „Mixed-Methods-Designs“, durch eine Kombination von standardisierten und explorativ-interpretativen („qualitativen“) Methoden identifiziert, untersucht und beschrieben werden können.

Hierzu werden wir im ersten Teil des Beitrags die Begriffe „Messfehler“ und „Messqualität“ im Hinblick auf ihre forschungsmethodische Operationalisierbarkeit diskutieren: wie lassen sich die Unterschiede zwischen systematischen und unsystematischen Messfehlern unter Bezug auf sozialwissenschaftliche Untersuchungssituationen konzeptuell fassen? Was bedeuten hier standardmäßig (etwa im Rahmen klassischer testtheoretischer Ansätze) gemachte Verteilungsannahmen inhaltlich?

Im zweiten, empirischen Teil des Beitrags beziehen wir diese Überlegungen auf Ergebnisse statistischer Auswertungen und qualitativer Interviewanalysen zu Items des Themenschwerpunkts „Religion und Weltanschauung“ im ALLBUS 2012. Ausgangspunkt bilden dabei prima facie inkonsistente Befunde, die bei einer Untersuchung religiöser Einstellungen von bestimmter Gruppen (wie etwa „religiös aktive Protestanten“) auftreten, wobei häufig verwendete Skalen (etwa zur Erfassung des Lebenssinns und kosmologischer Überzeugungen) auf-fällige, schwer zu interpretierende Antwortverteilungen zeigen. Diese Befunde bildeten den Ausgangspunkt für qualitative Untersuchungen zur Validität und dabei insbesondere zu Frage- und Interviewereffekten bei diesen Items, wobei wir der Frage nachgingen, wie Angehörige unterschiedlicher religiöser (Teil)populationen Fragen zu ihrer Religiosität und Spiritualität verstehen und ob hinsichtlich solcher Items gruppenspezifische Antworttendenzen existieren. Mit Hilfe eines Mixed-Methods-Designs haben wir dabei einerseits verschiedene Hypothesen zur Entstehung der von uns identifizierten Inkongruenzen geprüft und mit Hilfe der ALLBUS-Daten untersucht, inwieweit sich Interviewereffekte oder Effekte durch die Anwesenheit Dritter finden lassen, bzw. ob Zusammenhänge zwischen Antwortverhalten bei diesen Items und der Einschätzung der allgemeinen Kooperationsbereitschaft der Befragten bestehen. Andererseits haben wir qualitative Interviewverfahren genutzt, wie sie im Rahmen von kognitiven Pretests eingesetzt werden, um zu untersuchen, wie Befragte die betreffenden Items verstehen, auf welche Wissensbestände sie bei einer Beantwortung zurückgreifen und welche kognitiven Prozesse hierbei eine Rolle spielen.

Auf der Grundlage unserer empirischen Befunde werden wir abschließend allgemeine methodologische Überlegungen und Empfehlungen hinsichtlich der Differenzierung von Messfehlerquellen diskutieren.

Can Satisficing Explain Measurement Errors? Using Difficulty, Ability, and Motivation to Explain Measurement Errors in Pensions Due to Old Age

Patrick Lazarevic (TU Dortmund; Patrick.Lazarevic@TU-Dortmund.de)

The theory of satisficing provides a useful framework for examining measurement errors in surveys. It proposes that the ability of a respondent to answer a question, the difficulty of a question, and the respondent's motivation to answer the question all influence the quality of an answer: A higher difficulty should increase errors while a greater ability and motivation should reduce them. A common problem in determining the amount of measurement errors in survey research is that the 'true value' is usually unknown. One possibility to obtain a 'true' reference value to compare the measurement to, is linking self-reports from a survey to process-produced or administrative data that represent the value in question. Therefore, in order to determine the measurement error for self-reported pensions due to old age, administrative data from the German Pension Fund are linked to self-reported data of 917 participants (434 male, 483 female) from the Survey of Health, Ageing and Retirement in Europe (SHARE). The discrepancies between these data-sources relative to the administrative value are used as a dependent variable in structural equation models. Ability is measured by variables representing the respondent's cognitive functioning like a performance test of the respondents's memory, difficulty is operationalized via the share of pensions due to old age on the total household income, and motivation comprises data on the interviewer's self-rated willingness to react to the respondent's needs. As suggested by the theory, paths from the latent variables ability, difficulty, and motivation are specified. Preliminary results show insignificant path coefficients for ability and motivation on errors while a lower difficulty significantly reduces errors. All results are consistent for both genders separately and for the whole sample. This suggests that the ability and motivation of a respondent are not the most influential factors when it comes to report single contributions to the income but rather its complexity and the relative importance of the inquired contribution.

Session 2: Aufnahmevorträge I

Satisficing in probability and nonprobability online panels

Carina Cornesse (SFB 884 "Politische Ökonomie der Reformen", Universität Mannheim und GESIS – Leibniz Institut für Sozialwissenschaften; carina.cornesse@uni-mannheim.de)

The ongoing debate about the quality of nonprobability online panels predominantly discusses whether or not these panels have representative sets of respondents. While the number of publications on nonprobability panel representativeness is increasing, less attention has so far been paid to potential measurement errors in nonprobability as compared to probability panels. In our paper, we investigate whether there are differences in satisficing across probability and nonprobability online panels using three indicators to operationalize survey satisficing (item nonresponse and non-substantive answers, straight-lining in grids, and mid-point selection in a visual design experiment). These indicators are included in a questionnaire module that was implemented across nine online panels in Germany: one academic probability online panel that includes the offline population, one commercial probability online panel, and seven nonprobability online panels, all differing with respect to their sampling and recruitment methods. Our analyses show significantly less straight-lining in probability than in nonprobability online panels, but no significant differences regarding mid-point selection. With respect to non-substantive answers, we find that significantly more respondents in the probability than in the nonprobability panels say that they don't know who they voted for in the last general election or refuse to report their height and body weight.

The Dynamics of neither/nor Answers in Attitudinal Questions

Miriam Trübner (Soziologie, Rheinische Friedrich-Wilhelms-Universität Bonn; truebner@uni-bonn.de)

The neither agree nor disagree category in Likert scales is usually treated as a substantive answer reflecting ambiguity or neutrality. However, a lack of opinion and a reluctance to answer truthfully can also lead to the use of the midpoint reflecting unsubstantive answers. Factors inducing this divergent use of the middle category may vary over time. In order to make causal conclusions about the mechanisms of response behavior, it is therefore important to analyze differences between as well as within respondents. By using the British Household Panel Survey (BHPS) for the years 1991 until 2007, the frequency of neither agree nor disagree answers in multiple Likert scaled item blocks is analyzed applying hybrid regression models. Thereby, it is possible to decompose response behavior into age, cohort, and period effect, and thus to disclose the dynamics underlying the use of the middle category. It will be demonstrated that the ageing effect is due to an increase in ambivalence and neutrality. Unrelated to this ageing process, older cohorts display different response behavior which is only partly related to their lower level of education. Moreover, in recent years the tendency to use the midpoint has increased. This can be explained by a lack of motivation when answering attitudinal questions over several waves, but also by the unobserved heterogeneity of long-term panel participants who are more prone to use the midpoint compared to short-term participants.

Exploring the influence of respondents' IT affinity on nonresponse in an online survey

Jessica Herzing (Political Economy of Reforms, Universität Mannheim; jessica.herzing@uni-mannheim.de)

Researchers have expressed concerns about the generalizability to the general population of estimations based on online surveys. While much of this discussion circles around the suitability of nonprobability sampling methods, the lack of coverage of persons without computers and/or internet has also received attention. Probability-based online panels only account for this potential source of error, if they specifically cover the offline population, for example by equipping them with devices and internet connection. However, even when covered, offliners tend to be underrepresented in the final sample because of nonresponse. Research in this area has thus far considered selectivity of sample units to be a binary phenomenon: sample units were either offline or online. In this paper, we extend this binary characteristic into the multi-dimensional characteristic of IT affinity. We use IT affinity to predict nonresponse in the German Internet Panel at registration and across waves. We find that respondents who belong to different classes of IT affinity have systematically different socio-demographic characteristics and show different voting behavior. In addition, we find that response propensities vary by classes of IT affinity. The combination of these two findings mean that IT affinity is an indicator of bias in socio-demographic characteristics and voting behavior due to differential nonresponse among the different classes of IT affinity. Our findings emphasize the importance of looking beyond a binary classification of offliners versus onliners when monitoring recruitment into probability-based online panels.

Session 3: Aufnahmevorträge II

Unit Nonresponse and Measurement Error – Evidence from a Probability-Based Panel

Robert Bauer (Survey Design and Methodology, GESIS Leibniz-Institut für Sozialwissenschaften; robert.bauer@gesis.org)
Kai Weyandt (GESIS Leibniz-Institut für Sozialwissenschaften)
Bernd Weiß (Universität Duisburg-Essen)
Michael Bosnjak (GESIS Leibniz-Institut für Sozialwissenschaften)

Systematic unit-nonresponse is a serious threat to the measurement quality of survey items. Therefore, most longitudinal studies put much effort in counteracting downward trends in response rates. On the other hand, it is argued that increasing the level of effort to drawing in reluctant respondents may actually increase the measurement error. So, the overall aim of this study is to contribute to the understanding of the relationship between unit nonresponse and measurement error. In a first step, we identify groups of panelists with similar nonresponse patterns, which in a second step will be analyzed for group specific differences in measurement error. Data were taken from the GESIS Panel online sample and include unit nonresponse time series across 15 waves and over 3,000 panelists. The GESIS panel is a probability-based mixed-mode access panel representative of the German population. Responses are categorized as dropout, unit nonresponse, partial response, and complete response. First, time series of unit nonresponse were clustered to identify groups of panelists with similar nonresponse patterns. For clustering, we applied optimal matching which is typically used for sequential data. Five robust groups are identified. The first group consists of panelists with very stable responses and includes about 80 percent of all panelists. A second group encompasses panelist showing alternate responses and non-responses (2.5 percent). The three remaining groups show substantial panel dropouts and are mainly distinguishable by time of dropout. Second, group specific differences in measurement error will be analyzed by means of several indicators of measurement error (e.g. straightlining, primacy effect (choice of left-aligned options), presence of item non-response and paradata). Especially, we will analyze the differences between the stable responses group and the early dropout group in regard to measurement error. Results and further implications for the understanding of the relationship between unit nonresponse and measurement error in panel studies will be discussed.

Interviewer Effects on Multiple Sources of Survey Error

Daniela Ackermann-Piek (SFB 884 , Universität Mannheim; ackermann-piek@uni-mannheim.de)

Concerns about interviewer effects in interviewer-mediated surveys have accompanied generations of survey researchers. As early as the late 1920s, Rice (1929) found that interviewers introduce measurement bias. However, interviewers' influence is not limited to measurement error, but affects nearly all aspects of survey errors, including sampling (Eckman, 2013; Eckman & Kreuter, 2011; Tourangeau, Kreuter, & Eckman, 2012), nonresponse (e.g., Blom, de Leeuw, & Hox, 2011; Durrant, D'Addio, & Steele, 2013; Jäckle, Lynn, Sinibaldi, & Tipping, 2012) and, to a lesser extent, coding and editing of survey responses (e.g., Campanelli, Thompson, Moon, & Staples, 1997). Following the Total Survey Error (TSE) framework, this paper examines interviewer effects in multiple areas of a survey: interviewer effects on measurement and interviewer effects on unit nonresponse. The aim is

to address whether the same interviewer characteristics are associated with interviewer effects across these multiple error sources. Researchers typically address interviewer effects on a single source of error. Part of the reason for this is that sufficient data on multiple error sources are seldom available in a single survey. In addition, analyses into one source of error are usually sufficiently complex to warrant publication. Using PIAAC data, I am in the fortunate situation of being able to combine the results of analyses into individual error sources described in the TSE. The analyses draw on various interlinked data sources: direct measures of interview quality (interview recordings), fieldwork processes (call-record data) as well as interviewer characteristics (interviewer survey) collected during the first wave of the German implementation of the Programme for the International Assessment of Adult Competencies (PIAAC). The analyses into measurement error and nonresponse are supplemented with paradata from the BBSR (2015) describing the sample composition to disentangle interviewer from sample composition effects.

Modellierung der zukünftigen Integrationsentwicklung der dritten Migrantengeneration mithilfe einer dynamischen Mikrosimulation

Dawid Bekalarczyk (Institut für Soziologie, Universität Duisburg-Essen; dawid.bekalarczyk@uni-due.de)

Integrationsprobleme treten nicht nur temporär unmittelbar nach der Einreise von Migranten auf, sondern sind Realisationen komplexer Prozesse, die nur unter Berücksichtigung sozialer, politischer und demographischer Dynamiken adäquat zu erfassen sind. Trotz gewisser Angleichungsprozesse reproduziert sich ethnische Ungleichheit teilweise auch in der zweiten Generation. Für die aus solchen Befunden resultierenden Frage nach der langfristigen Integrationsentwicklung wird die Arbeitsmarktperformance der dritten Generation sehr aufschlussreich sein. Eine entsprechende empirische Analyse ist gegenwärtig noch nicht möglich, da die Mitglieder dieser Generation im Schnitt noch sehr jung sind. Mit einer auf empirischen Daten basierenden Zukunftsprojektion existiert allerdings eine Möglichkeit, dennoch Aussagen über diese Entwicklung zu machen. Ziel des hier vorzustellenden DFG-Projekts ist daher die Modellierung der zukünftigen Integrationsentwicklung in der dritten Migrantengeneration als das Resultat oben genannter Dynamiken. Die Skepsis gegenüber Zukunftsprojektionen innerhalb der empirischen Sozialforschung liegt in der falschen Annahme, dass eine wirkliche Vorhersage der Zukunft angestrebt wird. Dabei lassen sich Vorhersagemethoden ebenso als quasi-experimentelle Designs verstehen, welche das Ausmaß der Reaktion auf einen Stimulus (Szenario) messen. Der Erkenntnisgewinn liegt darin, zu verstehen, wie eine Entwicklung beschaffen ist, wenn die diese Entwicklung bedingenden Faktoren ein derart komplexes Zusammenspiel aufweisen, dass ihre verlässliche Vorhersage nicht durch einfaches logisches Denken/Rechnen zu erreichen ist. Die Integrationsentwicklung in der dritten Generation wird dadurch zu komplex für einfache Modellierungsansätze, dass zusätzlich zu kausalen Einflüssen auf Individualebene demographisch bedingte Veränderungen der ethnischen und sozialstrukturellen Bevölkerungszusammensetzung Kompositionseffekte auslösen können. Daher muss das Prognosemodell modular organisiert werden, anstatt alle Mechanismen, wie in gängigen statistischen Verfahren, im Vorfeld durch ein zusammenhängendes Set mathematischer Funktionen zu spezifizieren. Eine hierfür geeignete Methode ist die dynamische Mikrosimulation, mit der sich im Vergleich zu anderen Simulationsansätzen detaillierte empirische Informationen implementieren lassen – vor allem auf Individualebene. Die Schätzung von Fortschreibungsparametern für die relevanten Mechanismen im Prognosemodell erfolgt durch eine umfassende panelregressionsanalytische Auswertung. Datenbasis hierfür ist ein selbst verzeigerter Datensatz, der das volle Spektrum des Sozio-oekonomischen Panels einbezieht. Als Startdatensatz für die Simulation dient ein Mikrozensusdatensatz, welcher die Bevölkerungsstruktur, insbesondere Kombinationen aus

Herkunftsland und Generation, ausreichend präzise abbilden kann. Vorgestellt werden Ergebnisse der Simulation eines empirischen Basisszenarios und eines aus der Migrationstheorie abgeleiteten Assimilationsszenarios im Vergleich.

Session 4: Survey and Questionnaire Design: Effekte auf den Messfehler

Wie beeinflussen die Polarität von Ratingskalen und die Verbalisierung der Skalenmitte die Messqualität?

Natalja Menold (GESIS Mannheim; natalja.menold@gesis.org)

Ratingskalen sind als unipolar oder als bipolar realisierbar. Unipolare Ratingskalen enthalten eine Bewertungsdimension, wie z. B. Zustimmung in der Ratingskala mit den Polen „stimme überhaupt nicht zu“ und „stimme voll und ganz zu“. Bipolare Ratingskalen hingegen enthalten zwei gegensätzliche Bewertungsdimensionen, wie z. B. Ablehnung und Zustimmung in der Ratingskala mit den Polen „lehne völlig ab“ und „stimme völlig zu. Die Skalenmitte hat unterschiedliche Bedeutung in uni- und bipolaren Ratingskalen. Während die Skalenmitte in einer bipolaren Ratingskala eine neutrale Position oder den Nullpunkt zwischen dem positiven und dem negativen Bereichen ausdrückt, bedeutet die Skalenmitte in einer unipolaren Ratingskala eine mittlere Ausprägung der einzustufenden Eigenschaft. Dementsprechend sind die Bezeichnungen der Skalenmitte „teils-teils“ und „weder-noch“ für bipolare Ratingskalen charakteristisch; für unipolaren Ratingskalen kommen Bezeichnungen „mittel“, „mittelmäßig“ oder „einigermaßen“ in Frage. In der Umfragepraxis wird auf die unterschiedliche Bedeutung der Skalenmitte in unipolaren und bipolaren Ratingskalen wenig geachtet und unterschiedliche Verbalisierungen inkonsistent verwendet. In Bezug auf die Polarität stellt sich zum einem die Frage, ob die Realisierung einer Ratingskala als uni- oder bipolar die Messqualität beeinflusst. Eine weitere wichtige Frage betrifft die Verbalisierung der Skalenmitte. Diese Forschungsfragen wurden in einer experimentellen Studie bearbeitet. Untersucht wurde eine heterogene Erwachsenenstichprobe eines kommerziellen Online-Panels. Erste Ergebnisse zeigen, dass die inkonsistente Verwendung der Skalenmitte die Reliabilität mindern kann. Außerdem wurden starke mindernde Effekte solcher Inkonsistenz auf die Validität beobachtet. In der Diskussion werden die Konsequenzen für die Fragebogenkonstruktion aufgezeigt und offene Forschungsfragen formuliert.

Do distractions during web survey completion affect data quality? Findings from a laboratory experiment

Alexander Wenz (Institute for Social and Economic Research, University of Essex; awenz@essex.ac.uk)

Web surveys are increasingly considered as cost-effective mode of data collection for large-scale social surveys. In contrast to interviewer-administered surveys, survey researchers lose control over the environment in which respondents complete the survey. Web respondents can decide when and where to fill in the questionnaire, and might be exposed to various sources of distractions or might choose to get involved in other activities while completing the survey. In particular, respondents who use a mobile device might be in distracting environments, where other people are present. Distractions and multi-tasking during web survey completion are potential threats to data quality as they might keep respondents from giving their full attention to the questionnaire. Distracted respondents might lack the cognitive capacity to accurately carry out the survey response process, which potentially introduces measurement error. This paper reports on results from a laboratory experiment that examines how distractions during web survey completion influence data quality, and aims to identify if the physical environment of survey completion is a potential source of measurement error. Subjects (N = 261) were randomly assigned to experimental groups using a 3 (form of distraction) x 2 (device type) factorial design. They were asked to complete a web questionnaire on either a PC or a tablet, and were allocated to one of three

distraction conditions: music versus conversation between other people in the room versus no distraction. The distraction treatments were chosen to represent two sources of distractions that are likely to occur in web survey settings. The study will examine the effects of distraction and device type on various data quality measures, including item-nonresponse, straight-lining, extreme response styles, and response consistency. This paper adds to research on how the environment in which respondents fill in web questionnaires affects response quality.

Der Einfluss des Survey-Modes auf die Datenqualität bei sensiblen Fragestellung am Beispiel von Sexualviktimisierung

Heinz Leitgöb (Universität Eichstätt-Ingolstadt; heinz.leitgoeb@ku.de)
Nathalie Leitgöb-Guzy (Bundeskriminalamt)

Mittlerweile gilt als empirisch gesicherte Erkenntnis der Umfrageforschung, dass sensible Fragen bei selbst-administrierten Erhebungsverfahren unter erhöhten Item-Nonresponseraten leiden. Im Gegensatz dazu liegt bislang kaum empirische Evidenz darüber vor, dass das geringere Niveau fehlender Angaben bei interviewer-administrierten Erhebungsverfahren insbesondere bei sensitiven Fragen mit Einbußen der Messqualität in Form von Falschangaben einhergehen kann. Basierend auf dem kognitiven Modell des Antwortprozesses sowie dem Rational Choice-Ansatz und Elementen der kausalanalytischen Graphentheorie werden diese Prozesse theoretisch eingebettet und so einer empirischen Überprüfung zugeführt. Konkret wird auf Basis von Daten des International Crime Victim Surveys (ICVS-2) geprüft, ob bei der Abfrage von Sexualviktimisierungen (i) aufgrund von höheren Messfehlern im Telefon-Mode geringere Zusammenhänge zwischen dem Selbstbericht und bekannten Korrelaten für Sexualviktimisierung auftreten als im Online-Mode und (ii) der erwartete Opferanteil unter den Item-Nonrespondenten im Online-Mode aufgrund von Falschangaben höher ist als im Telefon-Mode und als im Online-Sample mit gültigen Antworten. Als Methoden kommen hierbei Regressions- und Propensity Score-Verfahren zum Einsatz.

Neuere Weiterentwicklungen der Item-Count-Technik (ICT) zur Vermeidung von Response Bias bei „heiklen Fragen“ in Surveys

Felix Wolter (Institut für Soziologie, Johannes Gutenberg-Universität Mainz; felix.wolter@uni-mainz.de)

Im Forschungsfeld zu sog. heiklen Fragen in Surveys – etwa Fragen zu selbstberichteter Delinquenz, Sexualität oder ausländerfeindlichen Einstellungen – ist in den letzten Jahren die Item-Count-Technik (ICT) in den Vordergrund gerückt, um spezifischen Problemen solcher Fragen, wie Misreporting (Untertreiben von sozial unerwünschten Verhaltensweisen oder Einstellungen, Übertreiben von erwünschten) oder Nonresponse (inbes. bei Einkommensfragen), zu begegnen. Wie bei anderen Spezialtechniken (z. B. Randomized-Response-Technik) ist die Idee der ICT, die Antwortsituation durch eine Verschlüsselung der Befragtenantwort vollständig zu anonymisieren. Bei der ICT geschieht dies durch Fragelisten, in denen die Respondenten mehrere binäre Items (ja/nein) gebündelt beantworten und so die Antwort auf das interessierende heikle Item unbekannt bleibt. Der Beitrag präsentiert empirische Analysen zu aktuellen Weiterentwicklungen der Basis-ICT. Die Item-Sum-Technik (IST) ist eine Variante der ICT für quantitative Variablen (Trappmann et al. 2014). Die Person-Count-Technik (PCT) (für binäre Items) verwendet keine Fragelisten aus Füller-Items, sondern arbeitet mit Personenlisten, in denen der Befragte die heikle Frage

für sich und (mehrere) andere Personen beantwortet (Grant, Moon, Gleason 2012; 2014). Die Analysen gehen der Frage nach, wie gut IST und PCT im Vergleich zu konventionellen, direkten Fragen abschneiden (DQ für direct questioning). Argumentiert wird dabei, was die Prävalenzen negativ konnotierter Verhaltensweisen oder Einstellungen angeht, mit Hilfe der „more-is-better"-Annahme, bei der davon ausgegangen wird, dass höhere Prävalenzen heikler, unerwünschter Verhaltensweisen valider sind. Geprüft wird außerdem, ob sich mit Hilfe der IST Antwortverweigerungen auf die Einkommensfrage reduzieren lassen. Die präsentierten Befunde stammen aus drei experimentellen Surveys: einer CATI-Erhebung (N=499), einem Online-Survey (N=525) sowie einer postalischen Befragung (N=580). In der CATI-Erhebung zeigt sich bei Einsatz der IST u. a. eine deutliche Reduktion von Item-Nonresponse bei der Einkommensfrage im Vergleich zu DQ. Die Anwendung der PCT im Online-Survey (insgesamt zehn heikle Fragen zu Verhaltensweisen und Einstellungen aus verschiedenen Bereichen) zeigt keine Verbesserung im Vergleich zu DQ. In der postalischen Befragung (Einstellungen zu geflüchteten Menschen) sind die PCT-Schätzer wie erwartet höher als die DQ-Schätzer, allerdings nicht bei allen Items signifikant unterschiedlich. Trotz der nicht uneingeschränkt „pro-IST/PCT" ausfallenden Befunde sprechen die Studien dafür, das Potential der Techniken künftig weiter auszuloten.

Justice" and "Fairness": A different meaning to different people?

Michael Weinhardt (Fakultät für Soziologie Universität Bielefeld; michael.weinhardt@uni-bielefeld.de)
Jule Adriaans (Universität Bielefeld)

This study sets out to investigate the concept and measurement equivalence of social justice attitudes across different social groups. It is part of the ongoing development of the questionnaire module "Social Justice and Fairness in Europe" for the European Social Survey round 9. The sociology of justice attempts to explain differences in justice evaluations for example by linking them to different social backgrounds. Empirical work has used a variety of terms such as "just", "fair", and "appropriate" to capture respondents' justice evaluations. However, if people differ in their understanding of these words, survey findings are likely to be biased. We therefore ask: To what extent are the concepts of social justice and fairness understood equally in social groups based on socio-economic position, age and gender? What differences do exist and how large are they? First, a quantitative experiment in a German-wide, representative web survey, the GESIS panel, was conducted. In this panel, respondents were asked about their justice perception of their own income. The wording of the items was varied randomly using the terms "just" ("gerecht"), "fair" ("fair") and "appropriate" ("angemessen"). This experiment will allow us to determine the extent to which different wordings lead to differences in the distribution of justice perceptions overall. Second, online probing (OP) is used which combines the advantages of cognitive interviewing for the assessment of survey questions with the features of an online survey, achieving a greater sample size and broader coverage of the issue. An online convenience sample is asked to answer questions on social justice attitudes followed by probes on how they understood key terms in the question. Using this technique, we can see to what extent differences in justice evaluations can be explained by differing meanings people attach to different question wordings. The discussion of the results will focus on how problems of lacking congruence of concepts may be tackled by developing item wordings which are equivalent across important subgroups of society.

Session 5: Interviewereffekte auf die Messung

Interviewer Effects as Measurement Errors in a Face-to-Face Survey on Financial Behaviour. Evidence from Zambia

Markus Fröhlich (Universität Mannheim; froelich@uni-mannheim.de)
Frauke Kreuter (University of Mannheim/University of Maryland/IAB)
P. Linh Nguyen (University of Essex/University of Mannheim)

Interviewer effects are one of the crucial elements of the total survey error and thus, are a key element to evaluate the quality of survey data. In particular, previous studies show that interviewer characteristics (gender, age, education or occupation), interviewer attitudes, and expectations of respondent reactions influence the answers given by respondents. In order to study interviewer effects as measurement error, this paper draws on survey data of members of village savings groups. Following an interpenetrated design regarding random interviewer assignments, the main objective of the survey was to record numerous sensitive questions on financial behavior, including savings and credit. True values on individual savings and credit amounts were compiled through registry books of the savings groups. The study design further includes self-administered or interviewer-administered survey capturing the interviewer characteristics and attitudes, as well as personal interviews administered to the interviewers of at least the questionnaire part on mobile money.

Befragte, die antworten ohne sich Gedanken über ihre Antworten zu machen, und Interviewer, die Teile ihrer Interviews fälschen

Jörg Blasius (Universität Bonn; jblasius@uni-bonn.de)

Die Qualität von Umfrage-Daten ist abhängig von drei Akteuren, die in jeder face-to-face und in jeder telefonischen Befragung beteiligt sind: Befragte, Interviewer und Mitarbeiter der Erhebungsinstitute. Während vereinfachte Antworten auf der Ebene der Interviewten in der Literatur bereits ausführlich diskutiert wurden, hier insbesondere im Zusammenhang mit den Arbeiten von Krosnick zum „Satisficing“, gibt es nur relativ wenig Forschung zum Thema Fälschungen durch die Interviewer und durch Mitarbeiter von Erhebungsinstituten. In dem Vortrag stehen zwei der drei Akteure im Vordergrund des Interesses: die Interviewer und die Befragten und die Frage, wer für „stark vereinfachte“ bzw. stereotype Antwortmuster verantwortlich ist. Am Beispiel des European Social Survey 2010 (ESS 2010) werde ich diskutieren, wann davon auszugehen ist, dass die Befragten nur irgendwelche Antworten gegeben haben ohne über den Inhalt der Fragen nachzudenken („strong satisficing“), und wann die Interviewer Antworten in den Fragebogen eingetragen haben, wobei sie weder die Fragen den Zielpersonen gestellt noch den genauen Inhalt der Fragen beachtet haben. Kam es im größeren Stil zu Fälschungen durch die Interviewer, so war dies nur dann möglich, wenn Mitarbeiter der Erhebungsinstitute zumindest indirekt an den Fälschungen beteiligt waren, z.B. in dem sie die Kontrolle der Interviewer zwar protokolliert, aber nicht durchgeführt haben. Zum Ab-schluss des Vortrages wird mit Hilfe sozialwissenschaftlicher Theorien ein Erklärungsansatz gegeben, unter welchen Bedingungen es zu Fälschungen von Interviewdaten kommen kann. Als Ergebnis sei vorweggenommen, dass es in mehreren Ländern, die am ESS 2010 teilgenommen haben, zu derart umfangreichen Fälschungen kam, dass diese Länder von zukünftigen inhaltlichen Analysen ausgeschlossen werden sollten.

What interviewer characteristics get the most out of the respondents? An analysis of interviewer effects on the item nonresponse rate in asset questions based on German SHARE data

Sabine Friedel (Munich Center for the Economics of Aging (MEA), Max Planck Institute for Social Law and Social Policy (MPISOC); friedel@mea.mpisoc.mpg.de)

This paper deals with item nonresponse on asset questions in a cross-national face-to-face panel survey, the Survey of Health, Ageing and Retirement in Europe (SHARE). The reason for studying these items is their badly affectedness of nonresponse. The average item nonresponse rate of the asset questions in the considered German subsample is 12% - compared to other question tropics representing the highest. When there are underlying reasons which evoke response tendencies in the sense of missing data one must understand the mechanism behind to handle them, since missing data is often imputed or suffers from listwise deletion. To detect the relationship of item nonresponse rates in asset questions and its emergence I focus on features of the interviewers, as they play a crucial role in the data generating process besides the respondent. Furthermore, interviewers are rather under control of the researcher than the survey participant. It is important to note that missing data does not necessarily mean bad data explicitly. If the survey participant does not know the answer it is a true and correct value. However, if the missing answer is related to any other data collection aspect, like interviewer characteristics, one must care about it and interventions concerning the interviewer might help to avoid missing data. My outcome of interest is the item nonresponse rate on asset questions in the fifth SHARE wave in Germany. I use a multilevel approach to differentiate effects of the respondent from the interviewer precisely. Interviewers' attitudes and sociodemographic characteristics on the second level are highlighted. Detailed information about the interviewers is coming from the SHARE Interviewer Survey. Preliminary results show the importance of the interviewer as the asset item nonresponse rates differ by interviewers. Furthermore, some interviewer characteristics have a significant effect such as gender and likelihood of linking own data with the income tax assessment. Respondents interviewed by male interviewers and by those who are more likely to link have a lower item nonresponse rate than those interviewed by female interviewers and by interviewers who are less likely to link.

Interviewer Effects and the Measurement of Financial Literacy

Tobias Schmidt (Forschungszentrum Deutsche Bundesbank; Tobias.Schmidt@bundesbank.de)
Thomas F. Crossley (University of Essex and Institute for Fiscal Studies)
Panagiota Tzamourani (Deutsche Bundesbank)
Joachim K. Winter (University of Munich)

In this paper we investigate interviewer effects and measurement error for a standard set of survey questions on financial literacy, an important economic concept. Much of the current knowledge about the effects of financial literacy on financial decisions is based on survey data. The premise is that individuals should know the answer to three standard questions on interest rates, inflation and portfolio diversification in order to make sound financial decisions. Despite advances in the analysis of financial literacy, measurement error arising from the survey response process is an important concern. In this paper we will focus on interviewer effects, because in the German Panel on Household Finance (PHF), a large survey on household finance that is representative of the German population, interviewer effects explain a substantial fraction of variance in financial literacy questions. Furthermore, a large

literature in survey methodology has shown that interviewers may lead to complications in variance estimation and nonresponse biases.

We first develop a tractable analytic framework for thinking about (i) interviewers both as a source of error in survey responses but also as a moderator of respondent errors, (ii) the consequences of interviewer effects for the kinds of models estimated in the financial literacy literature, and (iii) how information on interviewers or interviewer effects might be used to improve estimates of the effects of financial literacy on financial choices and outcomes. We then use data on financial literacy collected as part of the PHF from 2010/11 to test for independent interviewer effects, moderating effects of interviewers on respondent error; to explore whether interviewer effects in financial literacy questions be related to interviewer characteristics, and to evaluate the strategies for mitigating the consequences of interviewer effects for substantive analysis.

We find that interviewer effects explain a substantial fraction of the variance of the financial literacy score, and that interview effects are related to interviewer characteristics. Furthermore we find a moderating effect of interviewers on respondent errors. Different corrections for interviewer effects in substantive equations suggest that the measurement error structure is very complex and cannot be fully addressed with simple solutions like IV estimation.

A potential pitfall in modelling the link between education and social network quality

Marina Trebbels (Fakultät für Erziehungswissenschaft Universität Hamburg; marina.trebbels@uni-hamburg.de)
Peter Winker (Justus-Liebig-Universität Gießen)

The analysis of social networks is of high relevance in different fields of social sciences. The collection of survey data on the size and quality of social networks requires substantial input, both from interviewers and respondents. A substantial number of studies have reported that interviewer effects might particularly affect the quality of data on network size, while the quality of data on measures of network quality might more strongly depend on the respondents' cognitive abilities. One way to avoid interviewer effects is using self-administered questionnaires. This, however, comes at the cost of a stronger bias due to differences in the respondents' cognitive abilities. We report empirical findings for a complex instrument used in a self-administered questionnaire applied in the National Educational Panel Study (NEPS) to 9th-graders in the classroom, which was designed to measure the social resources young people have at their disposal at the transition from general into vocational education (including the network's educational background and different social background data). The data allows identifying participants and population subgroups who face particularly strong difficulties in using the proposed instrument in a consistent way. This selection is highly correlated with the educational track the participants are attending as well as – for low levels of education – with students' migration background. Given that also measures of social network quality are found to correlate with educational and migration background, ignoring the selection caused by the complex instrument may heavily bias estimates of the link between education or cognitive skills, migration background and social network quality. We compare results for the nexus between education and social network quality obtained using two approaches. For the first approach, all available observations are used employing a naive procedure to correct for inconsistent answers. This might be considered the state of the art when working with the data at hand. For the second approach, the inconsistent cases are treated separately, either by simply excluding them from the analysis or by taking also the selection into this group into account