Migration Research in a Digitized World: Using Innovative Technology to Tackle Methodological Challenges

Pötzschke, Steffen (Ed.); Rinken, Sebastian (Ed.)

IMISCOE Research Series

Steffen Pötzschke
Sebastian Rinken *Editors*

Migration Research in a Digitized World

Using Innovative Technology to Tackle Methodological Challenges

IMISCOE

OPEN ACCESS

Springer

IMISCOE Research Series

Now accepted for Scopus! Content available on the Scopus site in spring 2021.

This series is the official book series of IMISCOE, the largest network of excellence on migration and diversity in the world. It comprises publications which present empirical and theoretical research on different aspects of international migration. The authors are all specialists, and the publications a rich source of information for researchers and others involved in international migration studies. The series is published under the editorial supervision of the IMISCOE Editorial Committee which includes leading scholars from all over Europe. The series, which contains more than eighty titles already, is internationally peer reviewed which ensures that the book published in this series continue to present excellent academic standards and scholarly quality. Most of the books are available open access.

Steffen Pötzschke • Sebastian Rinken
Editors

Migration Research in a Digitized World

Using Innovative Technology to Tackle Methodological Challenges

Springer

Editors
Steffen Pötzschke
GESIS – Leibniz Institute for the Social Sciences
Mannheim, Germany

Sebastian Rinken
Spanish Research Council (CSIC)
Institute for Advanced Social Studies (IESA)
Córdoba, Spain

The publication of this book has been supported by the Leibniz Association's Open Access Publishing Fund.

IMISCOE

ISSN 2364-4087 ISSN 2364-4095 (electronic)
IMISCOE Research Series
ISBN 978-3-031-01318-8 ISBN 978-3-031-01319-5 (eBook)
https://doi.org/10.1007/978-3-031-01319-5

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Acknowledgements

As tends to happen with edited books, ours has been a long haul, which would not have been possible without the support of many people. Back in 2018, Evelien Bakker, Senior Publishing Editor and curator of IMISCOE's Research Series at Springer, expressed interest in the topic of a conference session we had organized. Justyna Salamońska, then at Warsaw University's Center for Migration Research (and now of Kozminski University, also in Warsaw) suggested focusing on the role of innovative ICTs. Anna Triandafyllidou, chairwoman of IMISCOE's publications committee, and the members of the respective team at IMISCOE's network office, supported the project at all stages. Michael Braun (GESIS) provided timely and helpful feedback on the book's first and last chapter. The initial manuscripts' two external evaluators contributed various helpful suggestions. IMISCOE's Standing Committee Methodological Approaches and Tools in Migration Research (Meth@Mig) provided funding for the linguistic review. We thank GESIS-Leibniz Institute for the Social Sciences for providing crucial administrative assistance. Furthermore, we are particularly thankful to the Leibniz Association's Open Access Publishing Fund for graciously financing the publication's open access fee.

Special thanks are warranted to the editors' colleagues in Meth@Mig's steering committee—Justyna Salamońska and Evren Yalaz—who supported this project throughout these past years. We are especially grateful to Justyna for helping us navigate the administrative hurdles related to the collection's language revision, which was realized in a timely way by Wayne Egers.

We also want to thank all the contributing authors for their insightful chapters, patience, and willingness to put up with several rounds of revisions, follow-up questions, and suggestions from us editors. Last but certainly not least, we thank our families for supporting our, at times, nerdy enthusiasm for this project.

Contents

Part II New Data Sources and Their Potential

About the Editors and Contributors

Editors

Steffen Pötzschke is a postdoctoral researcher and deputy team leader of the GESIS Panel at the GESIS – Leibniz-Institute for the Social Sciences in Mannheim (Germany). Furthermore, he is a corresponding member of the Institute for Migration Research and Intercultural Studies (University of Osnabrück, Germany). Steffen holds a master's degree in international migration and intercultural relations and a doctorate from the University of Osnabrück. Steffen participated in several migration research projects, and has profound practical knowledge in designing and implementing cross-cultural surveys. In his recent research, he investigates the possibility of using social networking sites as tools to sample hard-to-reach populations.

Sebastian Rinken (PhD, European University Institute, 1996) is deputy director of the Spanish Research Council's Institute for Advanced Social Studies (IESA-CSIC) in Córdoba. He has published widely on immigrant populations' social integration and natives' attitudes toward immigration and immigrants, addressing issues such as the relation between ideological polarization and anti-immigrant sentiment, as well as the methodological challenge of eluding social desirability bias, among many others. His methodological repertoire includes qualitative approaches, probability-based surveys, non-probability sampling for on-site and online surveys, and survey experiments.

Contributors

Christopher Antoun is an assistant research professor in the College of Information Studies (iSchool) and Joint Program in Survey Methodology (JPSM) at the

University of Maryland. His research focuses on using smartphones to collect population data through text messaging, mobile questionnaires, or apps and sensors. He obtained his PhD in survey methodology from the University of Michigan and was a postdoctoral fellow at the US Census Bureau. Currently, he is an associate editor for the *Journal of Survey Statistics and Methodology* and serves on the editorial board of *Public Opinion Quarterly*.

Jacob Beck is a PhD student in the Chair of Statistics and Data Science in Social Sciences and the Humanities at LMU Munich, Germany. He received a master's degree in sociology from the University of Mannheim, Germany. His research interests are mainly in data science and the integration of big data in sociological research.

Dorothée Behr is team leader of the Cross-Cultural Survey Methods team at the GESIS-Leibniz Institute for the Social Sciences, Mannheim, Germany. She has a diploma in translation studies from the University of Heidelberg and a doctorate in applied translation studies from the University of Mainz. Her research and services focus on questionnaire translation and cross-cultural web probing.

Emanuel Deutschmann is Assistant Professor of Sociological Theory with a focus on conflict research in the European context at the University of Flensburg, and is also an associate at the European University Institute's Migration Policy Centre (MPC). His research interests cover social networks, transnational mobility, regional integration, and globalization. His latest book is *Mapping the Transnational World: How We Move and Communicate across Borders, and Why It Matters* (Princeton University Press, 2021). He holds an MSc in sociology from Oxford University and a PhD in the same field from the Bremen International Graduate School of Social Sciences.

Stephanie Eckman is a fellow in the Survey Research Division at RTI International, specializing in methods to collect high-quality survey data. Her research focuses on measurement error in surveys and the combination of survey and passive data. Previously, she held teaching and research positions at the Institute for Employment Research in Nuremberg, Germany, and at the University of Mannheim. Dr. Eckman received a PhD in survey methodology from the University of Maryland.

Raffaela Gentile studied sociology (BA) at the University of Mannheim. Since 2016, she has been working as scientific support at the Leibniz Institute for Educational Trajectories (LIfBi) in Bamberg. Until 2018, she worked on the ReGES–Refugees in the German Educational System study. Currently, she is working on the National Educational Panel Study (NEPS) on the research projects NEPS-Migration and NEPS-Returns to Education.

Agata Górny is an associate professor at Warsaw University where she is the head of Population Economics and Demography Chair in the Faculty of Economic Sciences and deputy director of the Centre of Migration Research. Her research interests include patterns of migration, economic integration of migrants, interrelations between family situation and migration, as well as survey methodology and mixed methods in migration research.

Michael Haan (PhD, University of Toronto, 2006) is an associate professor at Western University. He also is the academic director of the Western University Research Data Centre and the director of Migration and Ethnic Relations Collaborative Graduate Specialization. His research interests intersect the areas of demography, immigrant settlement, labor market integration, and data development. Dr. Haan is widely consulted by provincial and federal governments for policy advice in the areas of immigration, settlement services, the Canadian labor market, and population aging. Currently, he is the investigator or co-investigator on over six million dollars of research focused on immigrant settlement, developing welcoming communities, and identifying the factors that predict the successful retention of newcomers. He has published over 50 articles and reports on these topics.

Florian Heinritz has worked at the Leibniz Institute for Educational Trajectories (LIfBi) in Bamberg, Germany, as scientific support for the study ReGES–Refugees in the German Educational System from 2017 to 2019. After graduating with degrees in sociology (MA) from the University of Bamberg, Germany, and sociology and social research (Laurea Magistrale) from the University of Trento, Italy, in 2020, he has been working as a research assistant at ReGES. Since 2021, he also has worked at Universität Hamburg as a research assistant. The focus of his research is on survey methodology.

Kristen Himelein is a senior economist/statistician with the Poverty Global Practice at the World Bank and an adjunct professor at Georgetown University. Her research interests center on sampling and survey statistics, and her work has been published in the *Journal of Development Economics*, the *Journal of Official Statistics*, and the *International Journal of Public Opinion Research*, among others. She holds a master's degree from the Harvard Kennedy School and a graduate certificate in survey statistics from the Joint Program in Survey Methodology at the University of Maryland.

Sophie Horneber received a master's degree in sociology from the University of Mannheim, Germany. She works in people analytics in the HR department at Adidas. Her main research interest lies in innovative research tools, such as big data and virtual reality, and their potential for social science research.

Nataliya Ivanova holds an MA in sociology and is a research fellow in the Group for Migration and Ethnicity Research and Center for Regional and Urban Studies at

the Russian Presidential Academy of National Economy and Public Administration (Moscow, Russia). Nataliya has conducted extensive fieldwork in Russia, and is the author of more than 20 articles devoted to a range of topics including, but not limited to, migrant integration, migration legislation, and migrant residential segregation.

Florian Keusch is Professor of Social Data Science and Methodology at the University of Mannheim, Germany, and adjunct assistant professor in the Joint Program in Survey Methodology (JPSM), University of Maryland, USA. He received a PhD in social and economic sciences (Dr.rer.soc.oec.) and an MSc in business (Mag.rer. soc.oec.) from WU, Vienna University of Economics and Business, Austria. His research focuses on nonresponse and measurement error with respect to (mobile) web surveys and digital trace data collection.

Reilly Lorenz is an editorial assistant at Life Science Alliance, an open-access, peer-reviewed journal founded by EMBO Press, Rockefeller University Press, and Cold Spring Harbor Laboratory Press. Before joining Life Science Alliance, she received her master's degree in political science from the University of Mannheim, Germany. Her research focuses on far-right ideologies, migration, and text analysis.

José Luis Ortega holds a PhD in information science from the University Carlos III of Madrid. He works at the Institute for Advanced Social Studies (IESA), which is part of the Spanish Research Council (CSIC). He has published more than 50 research papers about web metrics (link analysis, altmetrics, etc.), information consumption, web usage mining, and scholarly information sources for scholars (Google Scholar, Microsoft Academic Search). Recently, he published the book *Social Network Sites for Scientists: A Quantitative Survey*, in which he analyzes the most relevant academic social networks (ResearchGate, Academia.edu, Mendeley, etc.) using quantitative techniques.

Howard Ramos (PhD, McGill University, 2004) is a professor at Western University. He is a political sociologist who investigates issues of social justice and equity. He has published 5 books and over 50 articles and book chapters on a range of issues including immigration, ethnicity, race, human rights, urban studies, social change, social movements, indigenous mobilization, and environmental advocacy. He also regularly works with municipal, provincial, and federal policy makers as well as non-governmental organizations.

Ettore Recchi is a professor at Sciences Po Paris where he is the director of the MA and PhD programs in sociology. He also is a fellow of the Migration Policy Centre (MPC) of the EUI (Florence) and a fellow of the Institut Convergences Migrations (Paris). A methodologically versatile sociologist, his last book is *Everyday Europe: Social Transnationalism in an Unsettled Continent* (Policy Press, 2019), a co-authored volume on European integration "from below." During the COVID-19 pandemic, Recchi has coordinated a project on the effect of air travel on the

spread of COVID globally, and a longitudinal study on the impact of the COVID-19 pandemic on social life in France.

Anna Rocheva holds a PhD in sociology and is a research fellow in the Group for Migration and Ethnicity Research and Center for Regional and Urban Studies at the Russian Presidential Academy of National Economy and Public Administration (Moscow, Russia). Anna has conducted extensive fieldwork in Russia and internationally, and is the author of more than 30 articles on a range of migration-related topics including, but not limited to, migration in Russia, migrant integration, gender and migration, second-generation migrants, and migrant residential segregation.

Justyna Salamońska is an associate professor in the Department of Management in Networked and Digital Societies, Kozminski University. Justyna holds a PhD in sociology from Trinity College Dublin. Her research and teaching interests include contemporary migrations in Europe, multiple migrations, migrant labor market integration, cross-border mobilities, attitudes towards migration and diversity, and quantitative and qualitative research methods.

Evgeni Varshaver holds a PhD in sociology and is an associate professor at the Higher School of Economics (Moscow, Russia) and the head of the Group for Migration and Ethnicity Research. Evgeni has conducted extensive fieldwork in Russia and internationally, and is the author of more than 40 articles on a range of topics related to migration, ethnicity, and the integration of migrants. He also teaches several courses on these topics.

Michele Vespe is a scientific officer at the European Commission, Joint Research Centre (JRC), where he coordinates a team of researchers investigating societal consequences associated with the improved availability of digital trace data, which includes research in the fields of data governance and computational social science. Previously, he was a senior scientist with NATO and a project engineer in industry. He holds a telecommunications engineering degree from the University of Florence (2003) and a PhD in signal processing from the University College London (2006). He is part of the editorial board of the journal *Data & Policy*.

Gisela Will is head of the Working Unit Migration at the Leibniz Institute for Educational Trajectories (LIfBi) in Bamberg, Germany. She is the coordinator of the project Refugees in the German Educational System (ReGES). Previously, she worked at the German National Educational Panel Study (NEPS) and was responsible for the implementation of its migration-specific aspects across various starting cohorts. Her research interests are international migration, ethnic and social educational inequality, integration of immigrants, social capital, and educational expectations and aspirations.

Acronyms/Abbreviations

A-CASI	Audio Computer-Assisted Self-Interviewing
AfD	Alternative für Deutschland (German political party)
API	Application Programming Interface
BAMF	Federal Office for Migration and Refugees (Germany)
CAPI	Computer-Assisted Personal Interviewing
CASI	Computer-Assisted Self-Interviewing
CAT (tool)	Computer-Aided Translation Tool
CCHS	Canadian Community Health Survey
CDR	Call Detail Record
CEEDD	Canadian Employer Employee Dynamics Database
CERN	European Organization for Nuclear Research
CSS	Cascading Style Sheet
CT	Census Tract
DIOC	Database on Immigrants in OECD Countries
EDS	Ethnic Diversity Survey
EU	European Union
GDPR	General Data Protection Regulation
GILT framework	Globalization, Internationalization, Localization, Translation framework
GMDAC	Global Migration Data Analysis Centre
GMP	Global Mobilities Project
GSS	General Social Survey
GT	Google Trends
HTML	HyperText Markup Language
HTTP	HyperText Transfer Protocol
IAB	Institute for Employment Research (Germany)
iCARE	Immigration Contribution Agreement Reporting Environment
ICT	Information and Communication Technology
IDP	Internally Displaced Person

IMDB	Longitudinal Immigration Database
IMISCOE	International Migration Research Network
IOM	International Organization for Migration
IRCC	Immigration, Refugees and Citizenship Canada
ISSP	International Social Survey Program
JSON	JavaScript Object Notation
KCMD	Knowledge Centre on Migration and Demography
LFS	Labour Force Survey
LisNZ	New Zealand's Longitudinal Immigration Survey
LSIA	Longitudinal Survey of Immigrants to Australia
LSIC	Longitudinal Survey of Immigrants to Canada
MPC	Migration Policy Centre
MPI	Migration Policy Institute
MT	Machine Translation
NATO	North Atlantic Treaty Organization
NGO	Nongovernmental Organization
OECD	Organization for Economic Co-operation and Development
PIAAC	Programme for the International Assessment of Adult Competencies
PISA	Programme for International Student Assessment
PRLF	Canadian Permanent Resident Landing File
RDS	Respondent-Driven Sampling
ReGES	Refugees in the German Educational System
REST	Representational State Transfer
SCIP	Socio-Cultural Integration Processes among New Immigrants in Europe
SDLE	Secure Data Linkage Environment
SHARE	Survey of Health, Aging and Retirement in Europe
SNS	Social Networking Sites
SOAP	Simple Object Access Protocol
SOEP	Socio-Economic Panel
SPARQL	SPARQL Protocol and RDF Query Language
SQL	Structured Query Language
SVI	Search Volume Index
TM	Translation Memory
TRAPD	Translation, Review, Adjudication, Pretesting, Documentation
UAV	Unmanned Aerial Vehicle
UNDESA	United Nations Department of Economic and Social Affairs
UNESCO	United Nations Educational Scientific and Cultural Organization
UNGMD	United Nations Global Migration Database
UNHCR	United Nations High Commissioner for Refugees
UNWTO	United Nations World Tourism Organization
URI	Uniform Resource Identifier

URL	Uniform Resource Locator
USA	United States of America
W3C	World Wide Web Consortium
XML	Extensible Markup Language

Chapter 1
Introduction: Using Innovative Technologies to Tackle Methodological Challenges in Migration Research

Steffen Pötzschke and Sebastian Rinken

Mobility is a defining feature of mankind, a fact w hich led some authors even to label our species *homo migrans* (Bade, 2003). In recent years, however, the surging volume of human mobility in general, and of cross-border movements in particular, has situated international migration as a major driving force of social change on a global level. The United Nations Department of Economic and Social Affairs estimates that there were 173 million international migrants in 2000 and 281 million in 2020 (UNDESA, 2020). Thus, the number of individuals who live outside their country of birth has grown by more than 60% within just two decades. These figures include international refugees and asylum seekers, whose numbers have more than doubled to about 33.8 million.[1] While (formally) voluntary migration in the pursuit of happiness, career opportunities, or improved living conditions continues to predominate, migration that is forced as a result of man-made or natural disasters has grown even faster, exposing those affected to tremendous hardships.

In this context, migration has come to occupy a prominent place in the public debates and political discourse of many countries around the globe. Human mobility, immigration, and emigration have profound social, economic, and political implications. To illustrate this complexity, we shall limit ourselves to mentioning a small selection of the most salient issues here. The aggregate economic effects of international migration are widely thought to be beneficial: for example, many states are

[1] This number includes Palestine refugees (under the mandate of the United Nations Relief and Works Agency for Palestine Refugees in the Near East) and displaced Venezuelans.

S. Pötzschke (✉)
GESIS – Leibniz Institute for the Social Sciences, Mannheim, Germany
e-mail: steffen.poetzschke@gesis.org

S. Rinken
Spanish Research Council (CSIC), Institute for Advanced Social Studies (IESA), Córdoba, Spain
e-mail: srinken@iesa.csic.es

S. Pötzschke, S. Rinken (eds.), *Migration Research in a Digitized World*, IMISCOE Research Series, https://doi.org/10.1007/978-3-031-01319-5_1

actively trying to attract (highly-) qualified migrants (Weinar & Klekowski von Koppenfels, 2020). This approach holds particularly true for countries in the Global North that use selective immigration as a strategy to counterbalance the negative labor market effects of declining birth rates. Migrants, who usually maintain continuing ties to those they left behind, also are of considerable economic importance to their countries of origin: in 2019, the combined value of remittances totaled 689 billion US-dollars, a fivefold increase since 2000 (IOM, 2019). Whether the migration of skilled workers might benefit the economic development of their countries of origin by means of a so-called "brain circulation" or, rather, constitutes a "brain drain" seems to depend on specific context factors (Atte, 2021; de Haas, 2012; Singh & Krishna, 2015). An undisputable downside, however, is that the institutional fabric of destination countries and supra-national organizations, such as the European Union, is strained by political actors who seek to exploit the increase of voluntary and involuntary migration for their own gain. Populist politicians have fanned perceptions of negative impacts on less qualified labor-market segments, conjured up cultural conflicts, and depicted redistributive policies as skewed against the interests of natives (Lucassen, 2018). Political actors and scholars aiming to preserve social inclusion have struggled to counter the false narratives that aim to stir public opinion against immigrants without, at the same time, neglecting the legitimate questions and concerns of natives.

In recent decades, these developments and a myriad of related issues have led to a considerable intensification of academic engagement with migration. This is evident not only in the surging volume of scientific projects (Isernia et al., 2018; Morales et al., 2020) and publications (Pisarevskaya et al., 2020; Pritchard et al., 2019) but also in the continuously growing number of dedicated postgraduate programs. In line with this expansion, research centers focusing specifically on migration-related issues have been established or consolidated in many countries, and the International Migration Research Network (IMISCOE) has grown from 24 founding institutes in 2004 to 61 institutional members in 2022 (IMISCOE, 2022; Levy, 2020).

Despite this increase in scientific activity, significant knowledge gaps still exist. Reliable, timely, and comparable data on the size, composition, and characteristics of immigrant and refugee populations ("stocks") and, especially, on recent migratory movements ("flows") are often lacking. Thus, key players in international migration governance have recently acknowledged the need to improve the quality and breadth of this data. Both of the United Nations' recent milestone agreements—the Global Compact for Safe, Orderly and Regular Migration and the Global Compact on Refugees—have highlighted the need to collect accurate data and commit signatory states to deepen and support corresponding efforts (UN, 2018a, objective 1, 2018b, item 3.3.). The establishment of the International Organization for Migration's (IOM) Global Migration Data Analysis Centre in 2015 (GMDAC, 2019) also bears witness to such enhanced interest in the collection and distribution of rigorous data on migrant populations and flows. The growing relevance and salience of international migration suggest that demand for such data will continue to grow in the foreseeable future, with a view to serving as input for academic research and for planning and delivering a wide range of services and policies. The ever more complex nature of international migration requires correspondingly sophisticated

data; for example, today many regions and countries are experiencing combinations of immigration, emigration, and transit mobility at the same time (Triandafyllidou, 2018).

While administrative statistics play an important role in answering many research questions, they tend to be more useful for identifying general patterns (e.g., flows and stocks of immigrants at given time points) than for revealing their underlying causes. Therefore, additional data generated by quantitative and qualitative research projects have long been indispensable for achieving a better understanding of migration. However, primary data collection is a demanding endeavor under the best of circumstances. It becomes even more challenging in a research domain in which scholars are regularly faced with manifold cultural and linguistic differences, potentially vulnerable or even traumatized target groups, and missing or incomplete sampling frames, to name just a few notorious issues. Such an abundance of difficulties and constraints implies that from a methodological viewpoint, migration research is extraordinarily complex and, by the same token, a promising breeding ground for creative and innovative solutions.

Given this situation, one might expect migration scholars to engage profusely in methodological debates. However, this is not the case. Instead, in the publications and conference presentations of migration scholars, research methods tend to play a marginal role; rarely are they the main topic.[2] In this regard, it seems indicative that—to the best of our knowledge—no international academic journal has sought explicitly and proactively to foster a debate on the methodological dimension of migration research. This assessment is substantiated by a recent bibliometric analysis of the field's development during the last decades. Pisarevskaya et al.'s (2020, Supplementary Data A) list of relevant outlets, which drew on input from senior migration scholars among other sources, did not identify a single publication channel (journal or book series) that had a methodological emphasis.[3]

Also attesting to the relative scarcity of methodological reflection among migration scholars is the fact that throughout the past decade, only a handful of edited volumes have been published in English, the lingua franca of the global scientific community, specifically on this vital dimension of migration research. None of these

[2] With a view to changing this state of affairs, the editors of this volume are among the founding members of IMISCOE's Standing Committee "Methodological Approaches and Tools in Migration Research" (Meth@Mig).

[3] Our assessment is based on the descriptions of scope provided on the websites of the journals listed in the online supplement to Pisarevskaya et al. (2020). The authors name a total of 47 journals as important outlets in the field, 44 of which continue to be published as of November 22, 2020. Only in seven cases (*Comparative Migration Studies*, *Cultural Diversity and Ethnic Minority Psychology*, *Ethnicity and Health*, *International Migration Review*, *Migration Studies*, *Movements*, and *Population Space and Place*) could we find hints suggesting that they might accept methodological contributions. Only three of those seven journals (*Comparative Migration Studies*, *International Migration Review*, and *Migration Studies*) are English-language publications with a clear international focus and interdisciplinary scope. However, even in these instances, none of their descriptions stress the importance of a thorough methodological debate in the field, present such a debate as a core concern of the journal, or specifically encourage submission of methodological contributions.

publications have focused on the new methodological options made possible by innovative information and communication technologies (ICTs).

The *Handbook of Research Methods in Migration*, edited by Carlos Vargas-Silva (2012), is perhaps the most comprehensive publication regarding the methodological dimension of migration research. It covers both qualitative and quantitative approaches and, as the "handbook" characterization suggests, touches on a broad range of topics. Its 27 chapters include introductions to different research methods, and also explore issues such as the management of large migration research projects and the translation of research findings into publications. However, most of the chapters that discuss specific research techniques focus on well-established procedures and issues, and only one (Crush et al., 2012) details the authors' experience when employing web-based resources such as personalized messages on social networking sites (SNS), in this case used for recruiting members of the African diaspora in Canada. Except for this chapter, and due perhaps to the fact that it was published a decade ago, the use of emerging technologies or data types is not a key concern of this otherwise extremely valuable collection.

The IMISCOE volume *Surveying Ethnic Minorities and Immigrant Populations: Methodological Challenges and Research Strategies* (Font & Méndez, 2013) focuses on quantitative methods, covering issues related to dedicated migrant surveys and the inclusion of migrants in general population surveys. Most contributions to this collection refer to situations in which relatively appropriate sampling frames for migrant surveys were available, or in which the migrant subsamples of general population surveys were large enough to enable their separate analysis. Complementing the *Handbook of Research Methods in Migration*, this book constitutes an important source of information for migration scholars who are interested in quantitative data collection. However, since *Surveying Ethnic Minorities and Immigrant Populations* was based on papers presented at a 2008 workshop, most of which reported on studies fielded several years earlier, it does not represent the latest state-of-the-art.

More recently, *Qualitative Research in European Migration Studies*, edited by Ricard Zapata-Barrero and Evren Yalaz (2018), was also published in IMISCOE's Research Series. As the title indicates, this book complements the volume edited by Font and Méndez in that it specifically addresses the issues raised by using qualitative research methodologies in migration studies; in this case, the territorial focus on Europe is made explicit. This publication addresses a vast range of concerns relating to the complete research process, from choosing the most appropriate methodological approach to developing policy implications. In doing so, it mainly focuses on established methods, except for one chapter that discusses research on migrants' use of ICTs, and how scholars might employ comparatively new ICT-based opportunities to collect qualitative data (Leurs & Prabhakar, 2018).

The general-population surveys discussed in Font and Méndez (2013) also attest to the broader survey research community's growing awareness of the need for generic (as opposed to dedicated, migrant-specific) data collections to address the reality of ever more diverse societies. Many survey design aspects that are of the highest importance to migration research have been discussed increasingly within

the general survey research community. These efforts have been flagged by labels such as *cross-national* or *cross-cultural survey research*. However, despite some noteworthy exceptions, debates on identical or very similar challenges have largely proceeded in parallel within the survey methods and migration research communities, respectively. Due to this disconnection, migration researchers often have failed to properly appreciate the relevance of the methodological expertise nurtured in the broader survey research community, and to seize on that know-how when preparing and conducting migration-related data collections. Relevant publications in this regard include some that discuss a broad array of topics and issues, such as *Survey Methods in Multinational, Multiregional, and Multicultural Contexts* (Harkness et al., 2010) and *Hard-to-Survey Populations* (Tourangeau et al., 2014), and others that focus specifically on linguistic and cultural issues in survey research, such as questionnaire translation and measurement equivalence in cross-cultural and cross-national studies (Behr, 2018).

Finally, a growing body of literature has highlighted the importance of communication technology for individuals and families during the planning of migratory projects, on the move, and after arrival (e.g., Akanle et al., 2021; Benítez, 2012; Borkert et al., 2018; Sanchez et al., 2018). Similarly, various authors (e.g., Martin & Singh, 2019; Rango, 2017; Sîrbu et al., 2020) have stressed that new digital data sources (often summarized under the umbrella term *big data*) can provide insights of great relevance to the study of migration. While some examples of innovative research exist in this area, such as the leveraging of mobile phone records to analyze forced migration (Salah et al., 2019), it seems fair to say that the migration research community has not generally paid much attention to the methodological opportunities and challenges arising in its field of studies due to the availability of new digital technologies and data types.

The present book aims to contribute to changing that state of affairs. It adds to the small number of extant publications on the methodological dimension of migration research by placing distinctive emphasis on the potential of technological innovations to provide new avenues for research on migrants and mobile populations. To achieve this goal, the book interconnects the expertise of both migration researchers and research methodologists. Based on the authors' hands-on involvement in empirical research, all chapters provide detailed accounts of the practical implementation of innovative research strategies. However, this publication was not conceived as a step-by-step handbook for using these strategies in future studies. Rather, our aim was to highlight various aspects and scenarios in which the research process might be enhanced by new methodological approaches and data sources. This includes the assessment of established methods, discussion of previous innovations, and identification of the challenges associated with using these new strategies and sources. The contributions to this book are diverse in their geographical scope, and they include studies based on projects realized in individual countries of the Global North and South, as well as chapters regarding data collection at a truly worldwide scale.

This edited collection is primarily aimed at migration scholars worldwide, especially, but not exclusively, those concerned with methodological issues in general and quantitative methods in particular. Yet, the topics discussed also will be of

interest to survey researchers, irrespective of their substantive focus. In addition, since methodological options are considered with regard to specific research needs, the collection will be useful for graduate courses in migration studies and related subjects, as well as graduate courses in research methodology. Finally, since the collection facilitates insights into the dynamics of using innovative methods to describe migration phenomena, it also will appeal to the migration policy community.

The nine contributions to this volume are grouped into two sections: the first is dedicated to data collections that are purposefully designed by researchers, whereas the second shifts attention to scholars' use of "found" data. The first section comprises five chapters on how new technologies might be used to address two major issues—sampling and linguistic diversity—that arise in "classical" surveys of migrant populations. The second section includes four chapters discussing a range of opportunities and challenges that arise from the use of various kinds of massive ("big") data collected, or indeed generated, for reasons alien to migration research.

In their chapter *Innovative Sample Designs for Studies of Refugees and Internally Displaced Persons*, Stephanie Eckman and Kristen Himelein discuss various strategies for sampling forced migrants in developing countries. Considering both international refugees and internally displaced persons, they focus on how to achieve probability samples of these hard-to-reach populations. Central to their exposition is the concept of *coverage error*, which they introduce to readers who might not be experts in survey research. Importantly, their chapter presents approaches that could be applied to different settings, for example, when target populations are housed in dedicated settlements (e.g., refugee camps), dispersed across urban areas, or still on the move. While building on established sampling techniques for recruiting participants of quantitative face-to-face research, the authors highlight how these techniques can be enhanced or adapted by incorporating new technologies, such as aerial images from satellites or drones, telephone trace data, and computational methods.

The chapter *Targeting on Social Networking Sites as Sampling Strategy for Online Migrant Surveys: The Challenge of Biases and Search for Possible Solutions* by Anna Rocheva, Evgeni Varshaver, and Nataliya Ivanova shifts the focus to online surveys. To address a situation in which appropriate sampling frames for migrants in their country of settlement are missing, the authors detail how advertisements on social networking sites can be used to recruit respondents. This approach enables scholars to reach hidden, and potentially geographically dispersed, populations in a timely and inexpensive manner. This chapter stands out from the still scarce literature on the implementation of this method by discussing Russian-based social media and networking sites *Vkontakte* and *Odnoklassniki* instead of focusing mainly on Facebook as the platform that monopolizes the extant literature almost completely. In addition to providing an informative description of the procedures employed in their research, Rocheva and her co-authors discuss the challenges associated with this approach and propose ways to address them.

The chapter *Web-Based Respondent-Driven Sampling in Research on Multiple Migrants: Challenges and Opportunities* by Agata Górny and Justyna Salamońska explores another innovative way of employing electronic communication and

personal networks to sample survey participants. Respondent-driven sampling has been developed specifically to recruit hard-to-reach populations. The authors discuss their experience with adapting the respondent-driven sampling (RDS) technique to the recruitment of Polish migrants who have been on the move several times (multiple-migrants), a scenario in which traditional RDS is not feasible for logistical and financial reasons. This chapter introduces the RDS method and examines the commonalities and differences of its face-to-face and web-based implementation, thus highlighting both the strengths and the weaknesses of the latter. Importantly, the authors provide first-hand experience and reflections on how the challenges they encountered when studying Polish multiple migrants might be addressed in future research.

In the chapter *Computer-Assisted Migration Research: What We Can Learn about Source Questionnaire Design and Translation from the Software Localization Field,* Dorothée Behr's point of departure is the fact that migration research is not the only kind of endeavor aimed at linguistically very diverse populations. With respect to the survey design process, she proposes, therefore, an emulation of the procedures developed by multinational corporations to adapt the user-facing part of consumer electronics software to global markets. In practice, globalization means localization—the capacity to adjust user interfaces and interactive menus to a wide range of languages, including those with different alphabets and page directionality. Behr shows how the localization industry's painstaking protocols are suitable for improving the quality of multilingual migrant surveys. Thus, she suggests that for such surveys to be successful, the following are required: very ambitious levels of logistic complexity, forward planning, multi-disciplinary co-operation, and financial resources.

The chapter *Surveying Illiterate Individuals: Are Audio Files in Computer-Assisted Self-Interviews a Useful Supportive Tool?* by Florian Heinritz, Gisela Will, and Raffaela Gentile shows how survey respondents' illiteracy in their native tongue is a relevant issue, especially in surveys that target recently arrived refugees. Traditionally, studies of such populations have relied on native speakers to conduct face-to-face interviews. However, apart from being expensive, this option may be prone to inducing response bias due to interviewer error and social desirability dynamics, which are potentially exacerbated by gender issues and ethnic cleavages. Based on their first-hand experience with a study of refugees from various Middle-Eastern and Asian countries, the authors examine the usefulness of audio recordings incorporated into digital self-interview questionnaires, an innovative tool that may enable illiterates to respond to surveys on their own.

In their chapter *Leveraging the Web for Migration Studies: Data Sources and Data Extraction*, which serves as introduction to the book's second section, Sebastian Rinken and José Luis Ortega familiarize migration researchers with web-based data sources, and encourage them to incorporate data science into their methodological repertoire. By pointing out that specific skills are required to handle the huge amounts of data available through the web, the authors seek to motivate migration scholars to widen their skill sets and cooperate actively with data scientists. First, the chapter introduces the basic concepts of *big data*, *open data*, and

linked data, and then describes the various types of data portals and repositories from which migration researchers can retrieve structured data. Finally, the authors provide a glimpse at some techniques that enable researchers to collect or access unstructured data from the web, such as the use of application programming interfaces (APIs) and web scraping.

How Canada's Data Ecosystem Offers Insights on the Options for Studying Migration in an Unprecedented Era of Information, by Howard Ramos and Michael Haan, examines how distinct data sources might contribute to migration studies. Taking the Canadian data ecosystem as an example, they begin by examining the use of censuses and national surveys, pointing out their particular strengths and weaknesses in the Canadian context and beyond. Next, they provide a thorough discussion of administrative records. While not new as such, these data sources are, to-date, highly underappreciated by migration researchers. Consequently, the authors make a compelling argument that this situation could be remedied through innovative research designs and linked data-analysis. The authors then turn their attention to new data sources, such as mobile phone records, before making a general plea for the creation of encompassing data spines, common protocols of data management, and international cooperation in this area.

The chapter *Assessing Transnational Human Mobility on a Global Scale* by Emanuel Deutschmann, Ettore Recchi, and Michele Vespe focuses on the basic fact of border-crossing that is both the foremost precondition of any more permanent settlement ("migration") and a more general trait of our contemporary world. Their contribution puts international migration into perspective by deriving a reliable estimate of global mobility from the innovative use of large, previously un-tapped datasets. To this end, they combine extant datasets on air-passenger traffic and tourism to construct a novel database of cross-country mobility, which is a complex task given that the data sources contain huge volumes of vastly different kinds of information. Thus, this chapter offers a fascinating example of how large sets of data that were not collected for research purposes can be handled and recomposed in a way that generates information on a relevant and previously under-researched topic.

Finally, in *Google Trends as a Tool for Public Opinion Research: An Illustration of the Perceived Threats of Immigration*, Reilly Lorenz, Jacob Beck, Sophie Horneber, Florian Keusch, and Christopher Antoun explore a source of big data that is available to any Internet user at just a click of the mouse. Google Trends (GT) facilitates information on the relative frequency of queries run on the world's leading search engine, scaling results from 0 to 100. Users can define the time period (starting in 2006) and geographical area (usually a country), and up to five search terms can be compared. Thus, the interface offers retrospective and/or current information regarding the relative popularity of specific search terms. The scope and timeliness of such data are evidently out of reach of any traditional research tool. However, these benefits come with some strings attached, including the unavailability of information on sociodemographic covariates. To assess the usefulness of GT for migration researchers, the authors focus on an empirical test case—the political fallout of threat perceptions associated with the so-called refugee crisis in Germany. Specifically, they examine whether the salience of various threat

perceptions, as revealed by negatively charged search terms, correlates with voting intentions for the "Alternative for Germany," a notorious anti-immigrant party.

In their conclusion, entitled *Migration Research in Times of Ubiquitous Digitization*, Sebastian Rinken and Steffen Pötzschke discuss the collection's contributions against the backdrop of a shifting data landscape in which innovative technology and "found" data are assuming increasing relevance. However, neither the book's editors nor any of the contributing authors are naïve about the manifold challenges entailed by the digital revolution. Salient issues include access to proprietary data, as well as guaranteeing the protection of personal information at all stages. That said, all the contributors are adamant that migration research, and most other fields of social research for that matter, have much to gain from the creative and responsible use of innovative data sources and data processing techniques.

References

Akanle, O., Fayehun, O., & Oyelakin, S. (2021). The information communication technology, social media, international migration and migrants' relations with Kin in Nigeria. *Journal of Asian and African Studies, 56*(6), 1212–1225. https://doi.org/10.1177/0021909620960148

Atte, F. (2021). The moral challenges of health care providers brain drain phenomenon. *Clinical Ethics, 16*(2), 67–73. https://doi.org/10.1177/1477750920946614

Bade, K. J. (2003). *Migration in European history*. Blackwell.

Behr, D. (2018). *Surveying the migrant population. Consideration of linguistic and cultural issues*. GESIS—Leibniz Institute for the Social Sciences. https://nbn-resolving.org/urn:nbn:de:0168-ssoar-58074-2

Benítez, J. L. (2012). Salvadoran transnational families. ICT and communication practices in the network society. *Journal of Ethnic and Migration Studies, 38*(9), 1439–1449. https://doi.org/10.1080/1369183X.2012.698214

Borkert, M., Fisher, K. E., & Yafi, E. (2018). The best, the worst, and the hardest to find. How people, mobiles, and social media connect migrants in(to) Europe. *Social Media + Society*. https://doi.org/10.1177/2056305118764428

Crush, J., Eberhardt, C., Caesar, M., Chikanda, A., Pendleton, W., & Hill, A. (2012). Diaspora on the web. New networks, new methodologies. In C. Vargas-Silva (Ed.), *Handbook of research methods in migration* (pp. 345–365).

de Haas, H. (2012). The migration and development pendulum. A critical view on research and policy: The migration and development pendulum. *International Migration, 50*(3), 8–25. https://doi.org/10.1111/j.1468-2435.2012.00755.x

Font, J., & Méndez, M. (2013). *Surveying ethnic minorities and immigrant populations. Methodological challenges and research strategies*. Amsterdam University Press. https://doi.org/10.1515/9789048519187

Global Migration Data Analysis Centre (GMDAC). (2019). *4th anniversary. Aims and achievements*. International Organization for Migration, Global Migration Data Analysis Centre. https://gmdac.iom.int/sites/default/files/gmdacs_4th_anniversary.pdf

Harkness, J. A., Braun, M., Edwards, B., Johnson, T. P., Lyberg, L., Mohler, P. P., Pennell, B.-E., & Smith, T. W. (2010). *Survey methods in multinational, multiregional, and multicultural contexts*. Wiley.

IMISCOE. (2022). *About IMISCOE*. https://www.imiscoe.org/about-imiscoe/mission

International Organization for Migration (IOM). (2019). *World Migration Report 2020* (World Migration Report). International Organization for Migration. https://publications.iom.int/books/world-migration-report-2020

Isernia, P., Urso, O., Gyuzalyan, H., & Wilczynska, A. (2018). *A review of empirical surveys of asylum-related migrants*. European Asylum Support Office. https://www.easo.europa.eu/sites/default/files/easo-review-surveys-1-2.pdf

Leurs, K., & Prabhakar, M. (2018). Doing digital migration studies: Methodological considerations for an emerging research focus. In R. Zapata-Barrero & E. Yalaz (Eds.), *Qualitative research in European migration studies* (pp. 247–266). Springer. https://doi.org/10.1007/978-3-319-76861-8_14

Levy, N. (2020). An IMISCOE effect? The role of a network of excellence in developing European migration research in the twenty-first century. *Comparative Migration Studies, 8*(1), 37. https://doi.org/10.1186/s40878-020-00196-z

Lucassen, L. (2018). Peeling an onion. The "refugee crisis" from a historical perspective. *Ethnic and Racial Studies, 41*(3), 383–410. https://doi.org/10.1080/01419870.2017.1355975

Martin, S. F., & Singh, L. (2019). Big data and early warning of displacement. In S. McGarath & J. E. E. Young (Eds.), *Mobilizing global knowledge. Refugee research in an age of displacement* (pp. 129–149). University of Calgary Press. http://hdl.handle.net/1880/111127

Morales, L., Saji, A., Prandner, D., Bergh, J., Bernat, A., & Méndez Lago, M. (2020). *Surveys to ethnic and migrant minorities across Europe. Identifying knowledge strengths and gaps using survey metadata* (No. 2; Ethmigsurveydata report). International Ethnic and Immigrant Minorities' Survey Data Network (Ethmigsurveydata). https://zenodo.org/record/3839677

Pisarevskaya, A., Levy, N., Scholten, P., & Jansen, J. (2020). Mapping migration studies. An empirical analysis of the coming of age of a research field. *Migration Studies, 8*(3), 455–481. https://doi.org/10.1093/migration/mnz031

Pritchard, P., Maehler, D. B., Pötzschke, S., & Ramos, H. (2019). Integrating refugee children and youth. A scoping review of English and German literature. *Journal of Refugee Studies, 32*-(Special_Issue_1), i194–i208. https://doi.org/10.1093/jrs/fez024

Rango, M. (2017). Innovative data sources. In Global Migration Group (GMG) (Ed.), *Handbook for improving the production and use of migration data for development* (pp. 21–29). Global Knowledge Partnership for Migration and Development (KNOMAD), World Bank. https://migrationnetwork.un.org/resources/handbook-improving-production-and-use-migration-data-development

Salah, A. A., Pentland, A., Lepri, B., & Letouzé, E. (2019). *Guide to mobile data analytics in refugee scenarios. The "data for refugees challenge" study*. Springer. https://doi.org/10.1007/978-3-030-12554-7

Sanchez, G., Hoxhaj, R., Nardin, S., Geddes, A., Achilli, L., & Kalantaryan, S. (2018). *A study of the communication channels used by migrants and asylum seekers in Italy, with a particular focus on online and social media*. European Commission; Migration Policy Centre. http://hdl.handle.net/1814/61086

Singh, J., & Krishna, V. V. (2015). Trends in brain drain, gain and circulation. Indian experience of knowledge workers. *Science, Technology and Society, 20*(3), 300–321. https://doi.org/10.1177/0971721815597132

Sîrbu, A., Andrienko, G., Andrienko, N., Boldrini, C., Conti, M., Giannotti, F., Guidotti, R., Bertoli, S., Kim, J., Muntean, C. I., Pappalardo, L., Passarella, A., Pedreschi, D., Pollacci, L., Pratesi, F., & Sharma, R. (2020). Human migration. The big data perspective. *International Journal of Data Science and Analytics*. https://doi.org/10.1007/s41060-020-00213-5

Tourangeau, R., Edwards, B., Johnson, T. P., Wolter, K. M., & Bates, N. (2014). *Hard-to-survey populations*. Cambridge University Press.

Triandafyllidou, A. (2018). Globalisation and migration. An introduction. In A. Triandafyllidou (Ed.), *Handbook of migration and globalisation* (pp. 1–14). Edward Elgar Publishing. https://doi.org/10.4337/9781785367519

United Nations (UN). (2018a). *Global compact for safe, orderly and regular migration. Final draft*. https://refugeesmigrants.un.org/sites/default/files/180711_final_draft_0.pdf

United Nations (UN). (2018b). *Global compact on refugees*. https://www.unhcr.org/5c658aed4
United Nations Department of Economic and Social Affairs (UNDESA). (2020). *International Migrant Stock 2020*. United Nations-Population Division. https://www.un.org/development/desa/pd/content/international-migrant-stock
Vargas-Silva, C. (2012). *Handbook of research methods in migration*. Elgar.
Weinar, A., & Klekowski von Koppenfels, A. (2020). *Highly-skilled migration. Between settlement and mobility*. Springer. https://doi.org/10.1007/978-3-030-42204-2
Zapata-Barrero, R., & Yalaz, E. (Eds.). (2018). *Qualitative research in European migration studies*. Springer. https://doi.org/10.1007/978-3-319-76861-8

Part I
Innovation in Migrant Surveys

Chapter 2
Innovative Sample Designs for Studies of Refugees and Internally Displaced Persons

Stephanie Eckman and Kristen Himelein

2.1 Introduction

The United Nations High Commissioner for Refugees (UNHCR) estimates that nearly 71 million people in the world were forcibly displaced at the end of 2018, amounting to 1 of every 108 people globally (UNHCR, 2019b). Of the more than 20 million refugees, two-thirds come from five countries—Syria, Afghanistan, South Sudan, Myanmar, and Somalia—and many of these people have been displaced for more than 5 years. Half are younger than 18, including approximately 138,600 unaccompanied minors. There are an estimated 41 million Internally Displaced Persons (IDPs), a 50% increase over the last decade. The nature of displacement has also changed: fewer people live in dedicated camps in rural areas and more in urban areas, especially in wealthier host countries (UNHCR, 2019b). This shift has led the international community to reevaluate its understanding of displacement from a humanitarian crisis to a development challenge, and donor portfolios are increasingly focused on new initiatives to aid these vulnerable groups.

With increased aid programming, however, has come the need for better quality and more rigorous data with which to design, implement, and monitor these projects. Probability samples should be the foundation of high-quality data about refugees and IDPs (or any population). Unfortunately, many or even most studies of these vulnerable populations use nonprobability methods to select cases, and some studies do not describe the selection process at all (Enticott et al., 2017; Jacobsen & Landau, 2003; Kuhnt et al., 2019). Changes in the number and situations of displaced persons necessitate a rethinking of sampling approaches. In addition, digital technology such

S. Eckman (✉)
RTI International, Washington, DC, USA
e-mail: seckman@rti.org

K. Himelein
World Bank, Washington, DC, USA

S. Pötzschke, S. Rinken (eds.), *Migration Research in a Digitized World*, IMISCOE Research Series, https://doi.org/10.1007/978-3-031-01319-5_2

as satellite photos and computer vision has led to the development of new sampling options. This chapter reviews nine probability sampling methods that have been or could be applied to refugee and IDP populations.

We focus in this chapter on samples for surveys about the living conditions of refugees and IDPs—their health, well-being, employment, etc. Such studies provide the data that guide relief and development efforts (Brown et al., 2008). We do not discuss studies which seek to count displaced persons moving into, out of, or through a given area, although such studies are important too (see Global Migration Group, n.d.; Hughes et al., 2016; Lu et al., 2016; Williams et al., 2015). The sampling methods we have chosen to highlight are those that result in probability samples (each case has a known probability of selection) and can be carried out in developing countries to study refugees and IDPs. We further focus on face-to-face surveys, although telephone and app-based surveys are also possible (Hoogeveen et al., 2019; Keusch et al., 2019).

Before describing the sampling methods, Sect. 2.2 reviews coverage error in surveys and how under- and overcoverage on a sampling frame can introduce bias into survey data. The sampling methods are then organized by the living situations of the population of interest. Section 2.3 discusses sample selection methods appropriate for populations in camps or other settlements. In the settlements, all or nearly all households or persons are eligible for the survey (absent targeting by country of origin). Because refugees and IDPs are increasingly living in urban areas rather than settlements, Sect. 2.4 considers approaches suitable for urban environments where the populations of interest live among the host population. Section 2.5 presents options for selecting migrants while they are on the move, although researchers should carefully consider the ethics of interviewing people in unstable situations. Throughout the chapter, we draw on our years of experience designing samples and conducting studies around the world with hard-to-reach and vulnerable populations (Eckman & Himelein, 2019; Himelein et al., 2014, 2017).

2.2 Coverage Error in Surveys

Several sources of error can affect surveys of refugees and IDPs, such as sampling and measurement error and bias due to nonresponse. Those designing surveys should be aware of all sources and seek to minimize them as much as possible.[1] We focus in this chapter on coverage error because it is less well known and relevant to studies of refugees and IDPs.

Coverage error refers to bias that arises in survey estimates due to under- or overcoverage on a sampling frame. Undercoverage occurs when some members of the target population cannot be selected. The cases do not appear on the frame—the list from which the sample is selected. Undercoverage happens for many reasons:

[1] For more information on error sources in surveys, see Biemer et al. (2017). For a review of errors in studies of migrants specifically, see Jacobsen and Landau (2003).

perhaps the target population is homeless, and the survey selects cases through households; perhaps the list from which dwelling are selected is outdated; or perhaps a survey excludes a province from the frame because of violence in the area.[2]

Just like nonresponse, undercoverage can lead to bias in survey estimates. Because we have no data for undercovered cases, the estimate of, for example, a mean is made only on the covered cases: $\bar{y}_c$. The amount of undercoverage bias in estimates of a mean, $bias(\bar{y}_c)$, is related to the size of the undercovered population and the difference in the mean of interest for the registered and unregistered population:

$$bias(\bar{y}_c) = (1 - CR) * [\bar{y}_c - \bar{y}_{uc}]$$

where:

CR is the proportion of the population that is covered,
$\bar{y}_c$ is the mean among the covered cases, and.
$\bar{y}_{uc}$ is the mean among the uncovered cases.

For example, a researcher might be interested in estimating the proportion of school-aged refugee children who attend school. If there is a large difference between the true rate of school enrollment in the registered and unregistered populations (that is, $\bar{y}_c - \bar{y}_{uc}$ is large), then even a small rate of undercoverage $(1 - CR)$ could lead to large bias (Lessler & Kalsbeek, 1992).

The converse of undercoverage is overcoverage—the inclusion on the frame of cases that are not members of the target population. Overcoverage is not necessarily a problem and does not always lead to bias. For example, a survey may be interested in interviewing only households with children, but the frame from which the sample is selected includes all households. In such cases, studies often screen cases to determine which are eligible. Screening involves short interviews, usually done with any available household adult, about the characteristics of those living in the household. Screening can help identify an eligible subpopulation for a study, but it does increase costs, because ineligible households must still be interviewed and screened out (Kalton, 2014; Lavallée, 2014). We discuss screening in the context of surveys of refugees and IDPs in Sect. 2.4.1.

In the following sections, we describe each sampling method and comment on its vulnerability to under- and overcoverage. Researchers should carefully consider these issues and think through whether a given sampling approach is likely to lead to coverage bias in their data.

[2]Note that undercoverage is not the same as nonresponse. Undercovered cases have no chance to participate in the survey, even if they would like to, because they are never selected. Nonrespondent cases are those that are selected and do not participate because of noncontact or refusal. We avoid the term representative because it often conflates undercoverage and nonresponse.

2.3 Living in Settlements

Many IDP and refugee populations live apart from other groups, perhaps in camps specifically built to house them. In such situations, researchers can often adapt sampling methods more commonly used for general population household surveys.

2.3.1 Registration Lists

The World Bank can often access official camp registration lists maintained by the UNHCR. For example, in the study Enquête Harmonisée Sur Les Conditions de Vie des Ménages in Chad (2018–2019), camps were selected with probability proportional to size, where the size was determined from the registration lists. Households in the selected camps were then selected directly from the registers (World Bank, forthcoming). This approach has the advantage that camp administrators are often able to assist with fieldwork by introducing the interviewers to residents.

When registers of camp residents are available to researchers, selecting samples from these lists is often a good choice. The registers often contain additional variables, such as gender, age, ethnicity, place of origin, and date of arrival, which can be used for explicit or implicit stratification. Because these variables likely correlate with the variables measured by the survey, stratification reduces the variance of sample estimates (Eckman & West, 2016).

Any persons or households that are not on the official registers will be undercovered by this approach (Lebanon Humanitarian INGO Forum, 2014). Such unregistered refugees face unique challenges earning money and accessing health care (HelpAge International and Handicap International, 2014). Because unregistered refugees have different characteristics, concerns, and outcomes than do those who are registered, a survey relying only on registration lists may produce biased estimates due to undercoverage and lead to poor policy conclusions.

Overcoverage can occur with this method if the lists include persons who are no longer in the camp. This type of overcoverage can increase data collection costs, because interviewers will spend time looking for the selected people who are not available, but it is unlikely to introduce bias. Another source of overcoverage in the registers is fraudulent registration (Lodinová, 2016). The most common issue with fraud would be multiple registrations by the same household as members try to increase their rations. Another issue is people living in the surrounding communities claiming to be IDPs to receive goods and services. When people who are not refugees or IDPs register, they may be selected and interviewed. Because they are not truly eligible, the data they give should not be part of the data set. Because those who commit fraud are likely different from those who are appropriately registered, the data collected from fraudulent registrants can introduce overcoverage bias.

Aside from coverage issues, a logistical concern with sampling from camp registration lists is that not all researchers can access the lists (Martin-Shields

et al., 2019). Although the approach of sampling from official registration lists can work well, alternatives are needed. We now turn to a discussion of these alternatives.

2.3.2 *In-Field Listing*

When registration lists are not accessible or not of high quality, in-field listing can be used. This method of frame creation is common in household surveys around the world. Clusters are first selected in one or more stages. Within the smallest clusters, field staff are sent to create a list of all households or dwellings. This process, called listing, produces a frame from which a sample can later be selected (Grosh & Muñoz, 1996; Harter et al., 2010). Although listing is commonly used for general population surveys, it can be adapted for studies of refugees and IDPs. A 2014 World Bank study in Uganda used listing to create a frame of households in refugee settlements (World Bank, 2019).

Although it may seem straightforward to create an accurate list of dwellings while walking around a selected cluster, errors of both undercoverage and overcoverage are common in listing (Eckman & Kreuter, 2013). To avoid undercoverage and other sources of bias, listing work should be done before any interviewing, preferably by different staff than those who will do the interviewing (Eckman & Koch, 2019; Manheimer & Hyman, 1949; Stoop et al., 2010). However, that approach necessitates two visits to the camps—one to do the listing and another to do the interviewing—increasing costs and introducing a delay in the data collection schedule. Another drawback to the listing approach is that it increases the amount of time that field staff spend in the camp. Listing involves walking systematically around an area, often with a tablet computer or laptop. This behavior can expose field staff to theft or even kidnapping and assault. Interviewing, because it involves travel from one randomly selected household to another, may be less dangerous (Himelein et al., 2017).

2.3.3 *Sampling from Satellite or Aerial Images*

When camp registration lists are not available or are out of date, and listing is too dangerous, alternatives are needed. Fortunately, the availability and resolution of satellite photos has increased in recent years. Many companies and governments have launched satellites to orbit the earth and take pictures of the land and ocean. The images range from high-resolution pictures, taken less frequently, to low resolution pictures, available at higher frequency. Unmanned aerial vehicles (UAVs, or drones) can also take aerial pictures, at lower cost and higher resolution. Some satellite and aerial images are available online at low or no cost, although availability and recency vary across countries. If the images are recent and high-resolution, they can be used to select a sample of dwellings in a refugee or IDP camp.

Consider the satellite images of an IDP camp in Haiti, shown in Fig. 2.1. The top image is from March 8, 2010, and the second is from April 29, 2010. Both are from Google Earth and were downloaded at no cost. The images show the development of a temporary settlement for IDPs. Between the two dates, the dwellings have been improved, perhaps because of the arrival of humanitarian aid. Let's say it is 2010 and we want to conduct a study of persons displaced by the earthquake in Haiti and living in temporary settlements such as this one. We could review satellite images to find the areas where people have resettled and then manually identify and label the dwellings and select a sample of them. We might then prepare paper or electronic copies of the satellite image with the selected dwellings marked for interviewers.

Rather than identifying dwellings manually in images, we could train an algorithm to identify them. Spectacular progress has been achieved in recent years in computer vision, the branch of computer science involved in detecting and identifying objects in images and videos. This progress is the result of three advancements: the development of advanced methods of artificial intelligence, such as deep learning and convolutional neural networks; the release of open-source machine-learning frameworks like Google's TensorFlow; and the availability of inexpensive computing resources that can run these models in hours rather than weeks. Using these tools, computers can learn to identify tents or buildings in satellite images. Wang et al. (2015) and Quinn et al. (2018) used computer vision methods to count tents in Vietnam and Turkey, with Wang et al. reporting 81% accuracy. These approaches could also be used to select a sample of tents for interviewing.

The accuracy of the computer vision detection is affected by the resolution and spectral characteristics of the images. Some dwellings are also easier to detect than others. The computer vision approach would likely work better on the dwellings in the second image in Fig. 2.1 than the first, because those in the second image are more uniform in size, shape, and color (Wang et al., 2015). With further development, these algorithms may be faster, cheaper, and more accurate than manual identification of dwellings from images.

New dwellings that have been built since the satellite photo was taken will be undercovered by both the manual and the computer vision approaches to identifying camp dwellings from imagery. If the satellite image is old or change is occurring rapidly, which is sometimes the case with refugee and IDP populations, the situation on the ground may not look like the photo at all. Unfortunately, there is no set schedule for how often satellite photos are taken and released. In times of crisis, however, satellite photos may be taken more often, to support aid efforts and reconnaissance. If updated images are not available, one solution is to use UAVs to create up-to-date photos just before they are needed (Eckman et al., 2018), although researchers should always be cautious about how UAVs would be perceived by refugees and IDPs. Another method for dealing with older images is to use a missed housing unit procedure to detect and select undercovered dwellings (Harter & English, 2018). However, these procedures are challenging to implement in the field (Eckman & O'Muircheartaigh, 2011). Overcoverage can also occur with this design if the selected cases are not dwellings or no longer exist.

Fig. 2.1 Displaced persons settlement in Haiti: March 8, 2010, and April 29, 2010. (Source: Google, Maxar Technologies)

2.4 Living in Urban Areas

Refugee and IDP populations are becoming increasingly urban. At the start of the twenty-first century, most refugees were in camp-based and rural settings. Now, more than 60% of displaced persons live in urban areas. This urbanization of refugees is particularly prevalent in middle- and high-income host countries, where nearly all refugees report living in private accommodation (UNHCR, 2019b).

Surveying these urban populations can be particularly challenging, especially if the displaced are cohabitating with established family members. However, several of the previously discussed techniques can be used or adapted to study urban refugees. For example, if high-quality registration lists such as those discussed in Sect. 2.3.1, are available, then they remain an option. However, in urban areas, it is easier to live without registering with the host country government or a nongovernmental organization. Therefore, we suspect that the share of refugees and IDPs who are missing from registration lists is larger in urban areas than in camps, and those who are registered are likely to have out-of-date information. In this section, we discuss other approaches that can capture unregistered refugees and IDPs in urban areas.

2.4.1 Household Selection with Screening

In-field listing and selection from satellite images, discussed in Sect. 2.3, are also options for surveys of refugees and IDPs living in urban areas. In-field listing was used in the Syrian Refugee and Host Community Surveys in 2015–2016 in Lebanon (Aguilera et al., 2020). Because listing methods are well-documented elsewhere (Grosh & Muñoz, 1996), we here discuss selection from satellite images in more detail. Several studies have selected buildings or dwellings from satellite and aerial images. Across sites in Senegal, South Africa, Sudan, and Zambia, Baker et al. (2019) and Lowther et al. (2009) selected buildings from images by manually marking the images. Lowther et al. found that more than 98% of the buildings they identified could be located by field workers. Dreiling et al. (2009) used a similar approach in the rural U.S.

Results from these studies point to several challenges. Undercoverage is certainly a concern, due to out-of-date images or miscoding by staff. Local staff may be better able to identify buildings from images than those who are less familiar with the area. Overcoverage was a concern in all three studies. In urban areas, many buildings contain no dwellings. Locating and inspecting these buildings for dwellings can substantially increase field costs. Dreiling et al. (2009) used trained persons with local knowledge to code buildings in images as containing dwellings or not. Most dwellings (91%) were correctly identified with about equal rates of overcoverage (structures thought to be dwellings which were not) and undercoverage (missed dwellings).

However, removing buildings which appear to contain no dwellings can also lead to undercoverage. Refugees and IDPs may be forced by circumstances to live in buildings that are not intended to be dwellings. Thus, we recommend that studies sample a portion of these buildings so that persons living there are not undercovered. Researchers could stratify buildings into those more and less likely to contain dwellings and sample at different rates from each stratum.

Another concern is that urban residential buildings may contain more than one dwelling, and it is often not possible to tell from a satellite image how many dwellings are in a building. Interviewers can be trained to select one or more dwellings from multi-unit buildings. They might use a table of random numbers[3] or (if the study uses tablet or laptops) a selection program. However, such procedures allow interviewer influence on the selection process, which can introduce bias (Eckman & Koch, 2019). In-field listing is better able to handle multi-unit buildings, because the field staff can list each dwelling separately if they are able to determine how many there are. In addition, a dwelling may be home to more than one household. Interviewers need guidance on how to handle these cases as well.

Once a sample of households is selected, the next step is to screen the selected households to determine which contain refugees or IDPs. Researchers should craft screening questions carefully. Information about how long household members have been in the country and their legal status may be considered sensitive. Respondents may be reluctant to mention household members who lack a valid residence permit and may even deliberately not report them. Any misreporting at the screener stage can lead to undercoverage bias.

The Migration and Remittances Household Surveys in Burkina Faso, Kenya, Nigeria, Senegal, South Africa, and Uganda used screening to find households where immigrants and returned migrants lived. In 2009, that survey screened approximately 56,000 households to interview about 10,000 (Plaza et al., 2011). Such large ratios between the number of households screened and the number of eligible households identified are not unusual and can increase data collection costs. To increase the efficiency of screening, researchers can stratify the sample of households into areas more and less likely to contain the population of interest. This stratification can be based on field staff observation or information from community organizations serving the population (Singh & Clark, 2013). However, some buildings or households should be selected in the strata where eligibility is expected to be low to avoid undercoverage bias.

2.4.2 Time-Location Sampling

Another technique that is available when refugees and IDPs live in host country communities is to select them through the community centers or organizations that

[3] See Figure 2a in Singh and Clark (2013) for an example.

they visit, if such organizations exist (Lee et al., 2014). This approach is known as intercept-point sampling, center sampling, or time-location sampling (Baio et al., 2011; McKenzie & Mistiaen, 2009).

This approach can result in a probability sample, if several conditions are met. First, the probabilities of selection of the organizations themselves must be known. That is, a convenience sample should not be used. Researchers should make a list of organizations where the population of interest can be found and select a sample, or select all of them, if there are not too many. They should record the probabilities of selection at this stage (for example: three organizations selected out of seven). Second, respondents should be selected from the organizations' members or visitors with known probability. Researchers should randomly select a time to visit the location and observe who is there, selecting a sample of visitors. It is important not to select only those who look approachable, to avoid biasing the sample. Third, respondents must accurately report which organizations they visit, and how often, so that weights can adjust for the higher probability of selecting those who are frequent visitors or who visit more than one organization. See Baio et al. (2011) for details on the development of weights and a discussion of an application in Italy.

Time-location sampling undercovers any members of the target population who are not members or visitors of any of the known organizations. Depending on the type of organizations, the undercovered population might include the sick and disabled, who have a harder time leaving their homes. Furthermore, any error in the reports of respondent visits to organizations will lead to inaccurate weights and thus biased estimates. We are not aware of any research into how accurately respondents can report these behaviors. One advantage to this method, however, is its low cost: in a review paper, McKenzie and Mistiaen (2009) found that surveys conducted via time-location sampling cost about half as much as surveys conducted via listing and screening.

In terms of data quality, the McKenzie and Mistiaen (2009) comparison study determined that time-location sampling overrepresented those migrants who were more connected to the community organizations and weighting only partially removed this overrepresentation. Even after weights, the survey results were biased (McKenzie & Mistiaen, 2009). However, some bias may be tolerable, given the cost savings.

2.4.3 Respondent-Driven Sampling

When a population is difficult to find in the field, but the members are well connected, respondent-driven sampling may be a good choice. This method begins with a nonprobability sample of seeds, who are members of the target population known to researchers. The seeds complete the survey and then recruit additional members to complete the survey and recruit additional members. The process continues, each seed creating a recruitment chain.

Under certain assumptions, probabilities of selection can be calculated for all cases recruited though this method (Heckathorn, 2002; Salganik & Heckathorn, 2016; Volz & Heckathorn, 2008), making it superior to snowball sampling and other methods of uncontrolled network sampling. The crucial assumptions are (1) that relationships are reciprocal (i.e., if person A recruits person B, then person B would have recruited person A, had person B been recruited first); and (2) that all respondents can accurately report how many members of the population are in their network. However, these assumptions do not always hold (Gile & Handcock, 2010). In addition, estimation works best when recruitment chains are long, but in practice, short chains are more common (UNHCR, 2019a).

Several studies have used respondent-driven sampling to study refugees (Liu et al., 2018; UNHCR, 2019a). The World Food Program used this approach to study refugees in Turkey. The results indicate the importance of the recruitment chains to the success of the project. Respondents tend to recruit others like them (from the same ethnic group, for example). To achieve a diverse sample, the seeds should represent the breadth of the population of interest (Bozdag & Twose, 2019). The book *Applying Respondent Driven Sampling to Migrant Populations: Lessons from the Field* (Tyldum & Johnston, 2014) offers practical advice to researchers who wish to implement this approach.

Undercoverage can occur with respondent-driven sampling if some members of the population are not connected to others. Unless a seed is chosen in that community, no persons can be recruited or interviewed. Thus, researchers should take care to recruit a diverse set of seeds with large networks that span different subgroups within the population of interest. For example, a study of immigrants to the United States from El Salvador, Guatemala, and Honduras should take care of recruit seeds from each country and from different demographic groups (age, gender, legal status, etc.) (Abuelafia et al., 2019).

Overcoverage can also occur with respondent-driven sampling. Because of the presence of an incentive for responding and the ability to earn a higher incentive for recruiting additional respondents, some people may claim to be population members when they are not. This type of overcoverage can introduce bias into survey data if the responses from the ineligible respondents are different from the responses of those who are truly eligible (Wright & Tsao, 1983).

2.4.4 Adaptive Sampling

This approach takes advantage of the fact that people tend to live near those who are like them. If we can identify one member of our target population, others are likely to be nearby. The adaptive cluster approach comes from wildlife surveys: where we find one zebra, we are likely to find many others (Thompson, 1990). Take the example of a survey of Syrian refugees living in a city in Turkey. It is likely that the refugees are not randomly distributed throughout the city but clustered together so that they can support each other and build a community. We first screen a random

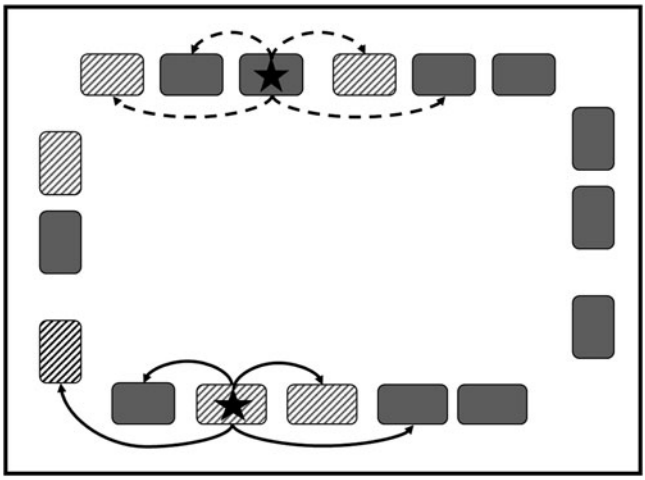


Fig. 2.2 Adaptive cluster sampling in household survey
Hatched boxes are eligible households; stars indicate those in initial sample. The lines show the additional selections that occur (solid lines) or would occur (dashed lines) when eligible households are found

sample of persons or households throughout the city, as in Sect. 2.4.1, to find a few Syrian refugees. We then put more resources into screening in the areas where we found Syrians.

The adaptive cluster approach was used by the European Union Agency for Fundamental Rights (2017) to sample Roma in the second wave of the European Union Minorities and Discrimination Survey. Figure 2.2 illustrates how the technique worked in that study. In the figure, each square represents an address. The two starred addresses are part of the initial sample and are screened. If a starred address is eligible, then the two addresses before and after it are also selected and screened. The starred address at the bottom of Fig. 2.2 is eligible, as indicated by the grey and white hatching. Because it is eligible, its four neighbours will also be screened, as shown with the solid arrows. Two of those addresses are also eligible. The starred address at the top of the figure is not eligible, so its four neighbours are not selected or screened (shown with dashed arrows). Even though two of its neighbours are eligible, they will not be screened. However, those addresses do have a probability of selection and thus are not undercovered: they have a chance to be part of the initial sample.

The European Union Minorities and Discrimination Survey found that the adaptive approach more than doubled the eligibility rate in two of the three countries where it was implemented, resulting in meaningful cost savings (European Union Agency for Fundamental Right, 2017). The efficiency gains from the approach are related to how clustered the target population is in the larger population.

Each household in Fig. 2.2 has a chance to appear in the initial sample, and many have an additional chance to be selected as the sample expands. Researchers should calculate analysis weights for each case that account for all chances that a case has to

be selected. Verma (2014) discusses weighting, and many other technical aspects of implementation of adaptive cluster sampling, in detail, including how large the neighborhood around each eligible household should be, how to introduce stratification, and when to stop expanding the sample.

We are not aware of a study using adaptive cluster sampling to study refugees or IDPs. However, the technique seems to be a good choice for populations that live in clustered urban areas. The tendency for under- and overcoverage with this approach should be the same as in any other household survey with screening (see Sect. 2.4.1).

2.5 On the Move

Although most studies of displaced populations take place once persons have settled, even temporarily, in a new location, some circumstances may require sampling mobile populations. Researchers may wish to understand the factors that compel people to continue traveling or understand how they protect themselves and maintain their livelihoods while in transit. Temporary collection areas, such as welcome centres, may offer a limited opportunity to conduct interviews or to develop a sampling frame that can be used later to recontact populations (World Bank, 2018). Below, we present two options for identifying and selecting migrants while they are on the move. However, we urge researchers interested in doing studies of migrants to carefully consider the ethics of their approaches. People who have left their homes and countries may not be in the best position to give informed consent to a survey interview.

2.5.1 Random Geographic Cluster Sampling

Random geographic cluster sampling is a method borrowed from forestry and wildlife surveys. Himelein et al. (2014) describe the method in detail and discuss an application to the Afar region of Ethiopia where livestock owners travel with their animals in search of water and food. To implement this method, we first selected random geographic points in the survey area. Geographic information system software can perform this random selection. Second, we created circles around the selected points. In the Afar implementation, the circles had radii of 0.1–5 km, with smaller radii in strata with higher likelihoods of finding people. Third, field workers travelled around inside the selected circles and interviewed all persons within them. To help guide interviewers to the selected areas and determine who was inside and outside, the points and circles were loaded onto handheld GPS devices.

This method captures persons wherever they happen to be at a given time, which makes it particularly useful for populations without permanent dwellings, such as refugees on the move, the homeless, or pastoralists. These populations are

commonly undercovered by surveys (Carr-Hill, 2013). The approach is cost-effective and can be implemented quickly, although it does require some technical sampling and mapping skills to calculate the probabilities of selection (see Himelein et al., 2014 for details). However, we are not aware of any studies using this method to sample refugees or IDPs.

All land in the target area is available for selection with this method, unless there are safety or accessibility issues. However, undercoverage of some housing units or persons can occur if interviewers do not canvass the entire circle thoroughly. As discussed with other methods, reliance on interviewers to perform selection is not ideal and opens the door to interviewer-induced undercoverage. For this reason, the circles should be kept small, 500-m radius or so. Himelein et al. (2014) had some success using Viewshed analysis to estimate how much of the selected circle interviewers observed and thus how much undercoverage there might be.

Overcoverage is also a concern; if persons travel during the field period, they can be found in more than one circle. To mitigate this issue, the field period should be short. If it is known that the population is generally traveling in a given direction (for example, from north to south), then the selected circles could be worked from south to north.

2.5.2 Mobile Phone Trace Data

Many refugees use smartphones and other mobile devices to coordinate travel and stay in touch with friends and family (Economist, 2017; Jones, 2019). Digital trace data generated by these devices can help researchers identify where the refugees are and thus where to sample for a face-to-face survey. Smartphones communicate through towers, and each tower keeps a record of the devices with which it communicates. Researchers can gain access to these call detail records, with some privacy protections applied. Most of the studies using call record data to study refugees count the stocks and flows of refugees (see, for example, Pastor-Escuredo et al., 2019). These studies often do not involve a survey—that is, they do not select and contact respondents for in-depth interviews about their living conditions, health, employment status, etc.

However, we are aware of one study that has used call detail records to design a face-to-face sample of refugees. The World Bank study of Venezuelans in Ecuador worked with a telecommunications company to obtain call detail records for each tower. They first determined which mobile phones were likely to be owned by Venezuelans: those which had made or received a call or text message from a Venezuelan number or accessed a website of interest to Venezuelans. All numbers with 30 or more such events in the prior 30 days were flagged. The towers used by those mobile phones between 8 p.m. and 6 a.m. were then identified as neighborhoods where Venezuelans lived. Neighborhoods were stratified into low, medium,

and high concentrations of Venezuelans. Within selected neighborhoods, the study used in-field listing of dwellings and screening (Muñoz et al., 2020).

With this method, the goal is not to interview the holders of the phone numbers in the call detail records—the records simply help identify where the refugees are. In fact, access to actual phone numbers is not required for this approach, which helps preserve respondent privacy. The locations of the cell towers can also be coarsened to protect privacy (Pastor-Escuredo et al., 2019).

This method seems to be a promising approach for future studies of refugees; however, it does face several challenges. Call detail records are difficult to acquire and are often maintained by several companies within a country. Researchers may need to negotiate data use agreements separately with each company; the study of Venezuelans was unable to obtain records from the largest telecommunications company in Ecuador (Muñoz et al., 2020). Any delay in receiving and processing the data may mean that the persons of interest have moved on. However, the method described above, which used the records only to identify neighborhoods where refugees are, is likely somewhat less vulnerable to changes over time—neighborhoods change less quickly than individuals' residences.

Undercoverage should be about the same as it is in any other survey using in-field listing and screening (see Sect. 2.4.1). Although there may be refugees who do not have mobile phones, or whose mobile phones are not flagged, those individuals are not undercovered, because the call data are used only for stratification, not selection. IDPs, however, would likely be harder to identify with this method, because their call, text, and browsing habits may be similar to other country residents who are not displaced.

2.6 Discussion

We have discussed nine sample selection methods for studies that wish to conduct face-to-face surveys of refugees and IDPs. For each approach, we have reviewed previous studies and summarized the advantages and disadvantages. We have also discussed the patterns of undercoverage and overcoverage that may result from each method.

The best approach for a given study will depend on the situation of the population of interest, the data collection budget, and the tolerance for under- and overcoverage bias. The highest quality surveys will use a high coverage method such as in-field listing with screening but they will also be the most expensive. Respondent-driven and time-location sampling can be logistically and statistically challenging but are generally faster and less expensive. Adaptive sampling, random geographic cluster sampling, and sampling from images using computer vision may be best for studies with technical staff.

We have not discussed surveys of host communities in this chapter, mostly due to space constraints. However, some of the studies we have cited also involved surveys with members of the host communities, to understand how they were affected by the

presence of refugees and IDPs. Often innovative designs are not needed for studies of the host community because they are well captured by census data and are the majority of the residents in their neighborhoods.

We foresee many more studies of refugees and IDPs in the coming years, as their numbers unfortunately continue to grow. We hope that the discussion in this chapter helps researchers in designing future studies. We also encourage continued methodological development to improve sample selection and survey data quality as new data sources and data collection methods become available. Survey researchers can help to improve the conditions of refugees and IDPs by collecting high-quality data about their living situations, which can support policy and aid responses.

References

Abuelafia, E., Del Carmen, G., & Ruiz-Arranz, M. (2019). *Tras los pasos del migrante: Perspectivas y experiencas de la migracion de El Salvador, Guatemala y Honduras en Estados Unidos*. International Development Bank. Available at https://publications.iadb.org/publications/spanish/document/Tras_los_pasos_del_migrante_Perspectivas_y_experiencias_de_la_migraci%C3%B3n_de_El_Salvador_Guatemala_y_Honduras_en_Estados_Unidos.pdf

Aguilera, A., Krishnan, N., Muñoz, J., Russo Riva, F., Sharma, D., & Vishwanath, T. (2020). Sampling for representative surveys of displaced persons. In J. Hoogeveen & U. Pape (Eds.), *Data collection in fragile states: Innovations from Africa and beyond* (pp. 129–151). Palgrave Macmillan.

Baio, G., Blangiardo, G. C., & Blangiardo, M. (2011). Centre sampling technique in foreign migration surveys: A methodological note. *Journal of Official Statistics, 27*(3), 451–465.

Baker, S., Ali, M., Deerin, J. F., Eltayeb, M. A., Espinoza, L. M. C., Gasmelseed, N., Im, J., Panzner, U., Kalckreuth, V. V., Keddy, K. H., Pak, G. D., Park, J. K., Park, S. E., Sooka, A., Sow, A. G., Tall, A., Luby, S., Meyer, C. G., & Marks, F. (2019). The Typhoid Fever Surveillance in Africa Program: Geospatial sampling frames for household-based studies: Lessons learned from a multicountry surveillance network in Senegal, South Africa, and Sudan. *Clinical Infectious Diseases, 69*(Suppl._6), S474–S482. https://doi.org/10.1093/cid/ciz755

Biemer, P., de Leeue, E., Eckman, S., Edwards, B., Kreuter, F., Lyberg, L., Tucker, N. C., & West, B. (2017). *Total survey error in practice*. Wiley.

Bozdag, I., & Twose, A. (2019). *Reaching hidden populations with an innovative two-stage sampling method: A case study from the refugee population in Turkey*. World Food Programme. Available at https://docs.wfp.org/api/documents/WFP-0000104292/download/

Brown, V., Guerin, P. J., Legros, D., Paquet, C., Pécoul, B., & Moren, A. (2008). Research in complex humanitarian emergencies: The Médecins Sans Frontières/Epicentre experience. *PLoS Medicine, 5*(4), e89. https://doi.org/10.1371/journal.pmed.0050089

Carr-Hill, R. (2013). Missing millions and measuring development progress. *World Development, 46*, 30–44. https://doi.org/10.1016/j.worlddev.2012.12.017

Dreiling, K., Trushenski, S., Kayongo-Male, D., & Specker, B. (2009). Comparing household listing techniques in a rural midwestern Vanguard Center of the National Children's Study. *Public Health Nursing, 26*(2), 192–201. https://doi.org/10.1111/j.1525-1446.2009.00770.x

Eckman, S., & Himelein, K. (2019). Methods of geo-spatial sampling. In J. Hoogeveen & U. Pape (Eds.), *Data collection in fragile states: Innovations from Africa and beyond* (pp. 103–128). Palgrave Macmillan.

Eckman, S., & Koch, A. (2019). Interviewer involvement in sample selection shapes the relationship between response rates and data quality. *Public Opinion Quarterly, 83*(2), 313–337. https://doi.org/10.1093/poq/nfz012

Eckman, S., & Kreuter, F. (2013). Undercoverage rates and undercoverage bias in traditional housing unit listing. *Sociological Methods & Research, 42*(3), 264–293. https://doi.org/10.1177/0049124113500477

Eckman, S., & O'Muircheartaigh, C. (2011). Performance of the half-open interval missed housing unit procedure. *Survey Research Methods, 5*(3), 125–131.

Eckman, S., & West, B. (2016). Analysis of data from stratified and clustered surveys. In C. Wolf, D. Joye, T. Smith, & Y. Fu (Eds.), *The SAGE handbook of survey methodology* (pp. 477–487). Sage.

Eckman, S., Eyerman, J., & Temple, D. (2018). *Unmanned aircraft systems can improve survey data collection*. RTI Press. https://doi.org/10.3768/rtipress.2018.rb.0018.1806

Economist. (2017, February 11). Migrants with mobiles: Phones are now indispensable for refugees. *The Economist*. Available at https://www.economist.com/international/2017/02/11/phones-are-now-indispensable-for-refugees

Enticott, J. C., Shawyer, F., Vasi, S., Buck, K., Cheng, I.-H., Russell, G., Kakuma, R., Minas, H., & Meadows, G. (2017). A systematic review of studies with a representative sample of refugees and asylum seekers living in the community for participation in mental health research. *BMC Medical Research Methodology, 17*(1), 37. https://doi.org/10.1186/s12874-017-0312-x

European Union Agency for Fundamental Rights. (2017). *Second European Union minorities and discrimination survey: Technical report*. Luxembourg. Available at https://fra.europa.eu/sites/default/files/fra_uploads/fra-2017-eu-midis-ii-main-results_en.pdf

Gile, K. J., & Handcock, M. S. (2010). Respondent-driven sampling: An assessment of current methodology. *Sociological Methodology, 40*(1), 285–327. https://doi.org/10.1111/j.1467-9531.2010.01223.x

Global Migration Group. (n.d.). *Handbook for improving the production and use of migration data for development*. Available at https://www.un.org/en/development/desa/population/migration/events/coordination/15/documents/Final%20Handbook%2030.06.16_AS4.pdf

Grosh, M. E., & Muñoz, J. (1996). *A manual for planning and implementing the living standards measurement study survey*. World Bank. Available at http://documents.worldbank.org/curated/en/363321467990016291/pdf/multi-page.pdf

Harter, R., & English, N. (2018). Overview of three field methods for improving coverage of address-based samples for in-person interviews. *Journal of Survey Statistics and Methodology, 6*(3), 360–375. https://doi.org/10.1093/jssam/smx037

Harter, R., Eckman, S., English, N., & O'Muircheartaigh, C. (2010). Applied sampling for large-scale multi-stage area probability designs. In *Handbook of survey research* (2nd ed., pp. 169–198). Emerald Group Publishing.

Heckathorn, D. D. (2002). Respondent-driven sampling II: Deriving valid population estimates from chain-referral samples of hidden populations. *Social Problems, 49*(1), 11–34. https://doi.org/10.1525/sp.2002.49.1.11

HelpAge International, & Handicap International. (2014). *Hidden victims of the Syrian crisis: Disabled, injured and older refugees*. HelpAge International and Handicap International.

Himelein, K., Eckman, S., & Murray, S. (2014). Sampling nomads: A new technique for remote, hard-to-reach, and mobile populations. *Journal of Official Statistics, 30*(2). https://doi.org/10.2478/jos-2014-0013

Himelein, K., Eckman, S., Murray, S., & Bauer, J. (2017). Alternatives to full listing for second stage sampling: Methods and implications. *Statistical Journal of the IAOS, 33*(3), 701–718. https://doi.org/10.3233/sji-160341

Hoogeveen, J. G., Rossi, M., & Sansone, D. (2019). Leaving, staying or coming back? Migration decisions during the northern Mali conflict. Journal of Development Studies, 55(10), 2089–2105. 10.13140/RG.2.2.22454.50248.

Hughes, C., Zagheni, E., Abel, G. J., Wisniowski, A., Sorichetta, A., Weber, I., & Tatem, A. J. (2016). *Inferring migrations: Traditional methods and new approaches based on mobile phone, social media, and other big data*. European Commission. Available at https://ingmarweber.de/

wp-content/uploads/2016/08/Inferring-Migrations-Traditional-Methods-and-New-Approaches-based-on-Mobile-Phone-Social-Media-and-other-Big-Data.pdf

Jacobsen, K., & Landau, L. B. (2003). The dual imperative in refugee research: Some methodological and ethical considerations in social science research on forced migration. *Disasters, 27*, 185–206. https://doi.org/10.1111/1467-7717.00228

Jones, C. (2019). 'The cellphone does everything': Smartphones, internet access are key tools of 21st century migration. *Cronkite News*, 2019.

Kalton, G. (2014). Probability sampling methods for hard-to-sample populations. In R. Tourangeau, B. Edwards, T. Johnson, K. Wolter, & N. Bates (Eds.), *Hard-to-survey populations* (pp. 401–423). Cambridge University Press.

Keusch, F., Leonard, M. M., Sajons, C., & Steiner, S. (2019). Using smartphone technology for research on refugees: Evidence from Germany. *Sociological Methods & Research, 004912411985237*. https://doi.org/10.1177/0049124119852377

Kuhnt, J., Martin-Shields, C., & Wedel, R. (2019). Challenges and possible solutions to conducting quantitative surveys with displaced populations. *Briefs on Methodological, Ethical and Epistemological Issues, 13*. Available at https://www.die-gdi.de/en/others-publications/article/challenges-and-possible-solutions-to-conducting-quantitative-surveys-with-displaced-populations/

Lavallée, P. (2014). Indirect sampling for hard-to-reach populations. In R. Tourangeau, B. Edwards, T. Johnson, K. Wolter, & N. Bates (Eds.), *Hard-to-survey populations* (pp. 445–467). Cambridge University Press.

Lebanon Humanitarian INGO Forum. (2014). *Background paper on unregistered Syrian refugees in Lebanon*. Available at http://lhif.org/uploaded/News/d92fe3a1b1dd46f2a281254fa551bd09LHIF%20Background%20Paper%20on%20Unregistered%20Syrian%20Refugees%20(FINAL).pdf

Lee, S., Wagner, J., Valliant, R., & Heeringa, S. (2014). Recent developments of sampling hard-to-survey populations: An assessment. In R. Tourangeau, B. Edwards, T. Johnson, K. Wolter, & N. Bates (Eds.), *Hard-to-survey populations* (pp. 424–444). Cambridge University Press.

Lessler, J. T., & Kalsbeek, W. D. (1992). *Nonsampling error in surveys* (1st ed.). Wiley-Interscience.

Liu, M., McCann, M., Lewis-Michl, E., & Hwang, S. A. (2018). Respondent driven sampling in a biomonitoring study of refugees from Burma in Buffalo, New York who eat Great Lakes fish. *International Journal of Hygiene and Environmental Health, 221*(5), 792–799. https://doi.org/10.1016/j.ijheh.2018.04.014

Lodinová, A. (2016). Application of biometrics as a means of refugee registration: Focusing on UNHCR's strategy. *Development, Environment and Foresight, 2*(2), 91–100.

Lowther, S. A., Curriero, F. C., Shields, T., Ahmed, S., Monze, M., & Moss, W. J. (2009). Feasibility of satellite image-based sampling for a health survey among urban townships of Lusaka, Zambia. *Tropical Medicine and International Health, 14*(1), 70–78. https://doi.org/10.1111/j.1365-3156.2008.02185.x

Lu, X., Wrathall, D. J., Sundsøy, P. R., Nadiruzzaman, M., Wetter, E., Iqbal, A., et al. (2016). Unveiling hidden migration and mobility patterns in climate stressed regions: A longitudinal study of six million anonymous mobile phone users in Bangladesh. *Global Environmental Change, 38*, 1–7. https://doi.org/10.1016/j.gloenvcha.2016.02.002

Manheimer, D., & Hyman, H. (1949). Interviewer performance in area sampling. *Public Opinion Quarterly, 13*(1), 83–92. https://doi.org/10.1086/266043

Martin-Shields, C. P., Camacho, S., Taborda, R., & Ruhe, C. (2019). *Digitalisation in the lives of urban migrants: Evidence from Bogota*. Deutsches Institut für Entwicklungspolitik.

McKenzie, D. J., & Mistiaen, J. (2009). Surveying migrant households: A comparison of census-based, snowball and intercept point surveys. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 172*(2), 339–360. https://doi.org/10.1111/j.1467-985X.2009.00584.x

Muñoz, J., Muñoz, J., & Olivieri, S. (2020). *Big data for sampling design: The Venezuelan migration crisis in Ecuador* (Policy Research Working Paper; No. 9329). World Bank. © World Bank. https://openknowledge.worldbank.org/handle/10986/34175 License: CC BY 3.0 IGO

Pastor-Escuredo, D., Imai, A., Luengo-Oroz, M., & Macguire, D. (2019). Call detail records to obtain estimates of forcibly displaced populations. In A. Salah, A. Pentland, B. Lepri, & E. Letouzé (Eds.), *Guide to mobile data analytics in refugee scenarios*. Springer.

Plaza, S., Navarrete, M., & Ratha, D. (2011). *Migration and remittances household surveys in SubSaharan Africa: Methodological aspects and main findings*. World Bank.

Quinn, J. A., Nyhan, M. M., Navarro, C., Coluccia, D., Bromley, L., & Luengo-Oroz, M. (2018). Humanitarian applications of machine learning with remote-sensing data: Review and case study in refugee settlement mapping. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Science, 376*(2128). https://doi.org/10.1098/rsta.2017.0363

Salganik, M. J., & Heckathorn, D. D. (2016). Sampling and estimation in hidden populations using respondent-driven sampling. *Sociological Methodology, 34*(1), 193–240. https://doi.org/10.1111/j.0081-1750.2004.00152.x

Singh, G., & Clark, B. D. (2013). Creating a frame: A spatial approach to random sampling of immigrant households in inner city Johannesburg. *Journal of Refugee Studies, 26*(1), 126–144. https://doi.org/10.1093/jrs/fes031

Stoop, I., Billiet, J., Koch, A., & Fitzgerald, R. (2010). *Improving survey response: Lessons learned from the European social survey*. Wiley.

Thompson, S. K. (1990). Adaptive cluster sampling. *Journal of the American Statistical Association, 85*(412), 1050–1059. https://doi.org/10.1080/01621459.1990.10474975

Tyldum, G., & Johnston, L. G. (2014). *Applying respondent driven sampling to migrant populations*. Palgrave Macmillan.

United Nations High Commissioner for Refugees (UNHCR). (2019a). *Crossing paths – A respondent driven sampling survey of migrants and refugees in Nouadhibou, Mauritania*. Available at https://data2.unhcr.org/en/documents/details/71198

United Nations High Commissioner for Refugees (UNHCR). (2019b). *Global trends: Forced displacement in 2018*. United Nations High Commissioner for Refugees: The UN Refugee Agency.

Verma, V. (2014). *Sampling elusive populations: Applications to studies of child labour*. International Labour Office.

Volz, E., & Heckathorn, D. D. (2008). Probability based estimation theory for respondent driven sampling. *Journal of Official Statistics, 24*, 79–97.

Wang, S., So, E., & Smith, P. (2015). Detecting tents to estimate the displaced populations for post-disaster relief using high resolution satellite imagery. *International Journal of Applied Earth Observation and Geoinformation, 36*, 87–93. https://doi.org/10.1016/j.jag.2014.11.013

Williams, N. E., Thomas, T. A., Dunbar, M., Eagle, N., & Dobra, A. (2015). Measures of human mobility using mobile phone records enhanced with GIS data. *PLoS One, 10*(7), e0133630. https://doi.org/10.1371/journal.pone.0133630

World Bank. (2018). *Asylum seekers in the European Union: Building evidence to inform policy making*. World Bank.

World Bank. (2019). *Informing the refugee policy response in Uganda: Results from the Uganda refugee and host communities 2018 household survey*. Available at http://documents.worldbank.org/curated/en/571081569598919068/pdf/Informing-the-Refugee-Policy-Response-in-Uganda-Results-from-the-Uganda-Refugee-and-Host-Communities-2018-Household-Survey.pdf

World Bank. (Forthcoming). *Building the evidence base on forced displacement in Chad*.

Wright, T., & Tsao, H. J. (1983). A frame on frames: An annotated bibliography. In T. Wright (Ed.), *Statistical methods and the improvement of data quality* (pp. 25–72). Academic. https://doi.org/10.1016/C2013-0-11729-1

Chapter 3
Targeting on Social Networking Sites as Sampling Strategy for Online Migrant Surveys: The Challenge of Biases and Search for Possible Solutions

Anna Rocheva, Evgeni Varshaver, and Nataliya Ivanova

3.1 Introduction

Migrant surveys present a number of methodological challenges, including a possible lack of sampling frames, migrants' status as a hard-to-reach and mobile population, and, in most cases, multilingualism. Depending on the national context, the extent of these difficulties may vary. As far as sampling frames are concerned, few countries maintain centralized population registers that provide accurate information on migrants. The situation in Denmark, Sweden, Italy, and Spain is better than in most countries, since their registers provide data on inhabitants' characteristics, including various categories of non-citizens. Nevertheless, using these registers as sampling frames demands some caution (Careja & Bevelander, 2018; UNECE, 2019), since they tend, on the one hand, to underrepresent short-term migrants and migrants whose legal status is unclear, meaning they are not readily "observable" by the State. On the other hand, these registers tend to overrepresent foreign-born persons whose emigration from a receiving country is not always easy to track. Population registers in many other countries are even more problematic with respect to migrant sampling. They may, for example, provide a limited number of categories that can be used to identify migrants, or they enforce rules that do not enable or

We thank our colleagues Ilya Schurov (Associate Professor at the Department of Higher Mathematics at the Higher School of Economics) and Irina Maslyakova (Senior Lecturer at the Department of Higher Mathematics at the Russian University of Economics) for their consultation. We also thank the Mail.ru Group for its cooperation.

A. Rocheva (✉) · N. Ivanova
Group for Migration and Ethnicity Research and RANEPA, Moscow, Russia
e-mail: anna.rocheva@gmail.com

E. Varshaver
HSE and Group for Migration and Ethnicity Research, Moscow, Russia

S. Pötzschke, S. Rinken (eds.), *Migration Research in a Digitized World*, IMISCOE Research Series, https://doi.org/10.1007/978-3-031-01319-5_3

motivate migrants to register. In some cases, there is no national register but several registers specific to particular communities (Salentin & Schmeets, 2017; Sanguilinda et al., 2017), and in the absence of a population register, scholars can use censuses and other statistical data as sampling frames (Kühne & Kroh, 2017; Reichel & Morales, 2017). Even migration-focused statistics do not, in most cases, represent the entirety of the migrant population of a receiving country accurately, especially in cases where a considerable share of migrants are undocumented, as in the USA (Hoefer et al., 2012).

In situations in which population registers or other data sources do not contain categories that enable an explicit differentiation between migrants and non-migrants, onomastic sampling can be used. While this method is effective in some contexts (Prandner & Weichbold, 2019; Salentin, 2014), it does not work in others, for example, in countries with a large non-migrant multiethnic population. Russia is one such case, since it has large numbers of people without a migrant background who have names and surnames that nonetheless closely resemble those of migrants from Central Asia and the Caucasus, due to the shared nomenclature traditions of Islamic culture.

Other methods such as random route walking (Reichel & Morales, 2017), time-location sampling (Agadjanian & Zotova, 2012), the area cluster approach (Vigneswaran, 2009), and respondent-driven sampling (Zotova et al., 2016) have limited efficiency when migrants come from different countries and are spatially dispersed in the receiving state. Moreover, the conditions of work and accommodation for low-qualified migrants in certain contexts can imply limited access for researchers. Such contexts may include "closed" workplaces, e.g., construction sites, large bazaars, and the "back-offices" of the catering industry. Long working hours, workplace residency, or employer-organized dormitories may constitute further barriers, as may low levels of trust and a reluctance to open apartment doors to unfamiliar persons. Random digit dialing can be too laborious if the share of migrants in the population is low, necessitating a high number of calls and screening to recruit a sufficient number of survey participants.

With respect to these circumstances, a new source of hope for migration scholars has been the development of digital information and communication technologies (ICT) and the rapid increase of the Internet penetration rate (World Bank, 2019). The active utilization of ICT both in the "mainstream" survey industry (Toepoel, 2015) and among migrants (Bucholtz, 2018; Dekker et al., 2015) has paved the way for the use of ICT in migration studies. As of today, this is done in a variety of ways. Researchers conduct ethnographic studies of online migrant communities (Mateos & Durand, 2012) and use big data on social networking sites (SNS) and other online services to estimate the number of migrants from specific countries of origin in destination countries (Spyratos et al., 2018; Zagheni et al., 2017), or even to assess the degree of their integration (Dubois et al., 2018; Herdağdelen et al., 2016). Significantly, researchers also conduct surveys online. In this latter case, SNS assume a greater importance (Hu & Wang, 2015; Wei & Gao, 2017) as a venue for recruiting respondents through posts in specific interest groups (Moreh, 2019),

snowball sampling (Herz, 2015) or using special targeting instruments provided by SNS for advertising purposes (du Plooy et al., 2018; Pötzschke & Braun, 2017).

The advantages of sampling migrants using SNS and surveying them online are clear: researchers can contact a population that is geographically dispersed within a short timeframe (McGhee et al., 2017; Sue & Ritter, 2012). Another advantage is that online, researchers can interact with a population that may otherwise be difficult to reach for various reasons. As mentioned previously, these difficulties may include the "closed" character of migrants' workplaces and accommodation or inadequate documentation (c.f. research on Internet surveys of illicit drug users [Miller & Sønderlund, 2010; Temple & Brown, 2011]). Importantly, web-based surveys are better equipped to address sensitive questions than are other survey modes (Milton et al., 2017), which can be important for migration studies (as, for example, concerning questions about migrants' documentation or lack thereof). Moreover, online surveys enable an easier coordination and organization of surveys across multiple countries simultaneously. Using the multilingual interfaces of survey software, such surveys do not require the deployment of interviewers who speak all of the migrants' native languages. Indeed, interviewers are not required at all. Researchers conducting such surveys would most likely require translators during the development stage of a questionnaire, the testing of its online implementation, and when cleaning and coding the data (unless, of course, the researcher is proficient in the languages spoken by the target population).

The advantages of web-based surveys in combination with SNS-based recruitment make this method promising for a significant number of research contexts, including Russia, since it does not have a population register, its census data on migrants tend to be inaccurate and incomplete (Mkrtchyan, 2011), and its statistics on foreign citizens are too limited to serve as a sampling frame, not to mention the even more inadequate data on persons with a migrant background (those who are naturalized or who are migrants' descendants). Migrants in Russia are dispersed across the country, regions, and cities.[1] At the same time, the Internet and SNS penetration rate among migrants is quite high. According to our bilingual face-to-face survey of Kyrgyz migrants in Moscow in 2014 (Varshaver et al., 2014), 63% of respondents used SNS. Today, several years later, we can expect this figure to be higher for three reasons: the penetration rate of the Internet has continued to rise (World Bank, 2019), there are more SNS today, which were not included in our questionnaire in 2014 (e.g., Instagram), and migrants to Russia are mostly young[2] and thus generally more internet "savvy." In total, from 2016 onwards, we

[1] Since current migration to Russia is primarily economically driven, more migrants are living in cities than in rural areas, and, more broadly, a higher concentration of migrants are living in the more economically developed regions. Nevertheless, migrants reside in all regions of Russia.

[2] According to data from the Federal State Statistics Service (2017–2018) and from the former Federal Migration Service (2016). See https://showdata.gks.ru/report/278008/ (accessed: 17.04.2020).

conducted five web-based surveys that employed SNS-based targeting of first- and second-generation migrants from Central Asia and the South Caucasus.[3]

While research has indicated the advantages of sampling migrants using targeting on SNS (du Plooy et al., 2018; Pötzschke & Braun, 2017), the drawbacks of this method—and, even more so, the proposal of potential solutions—constitute an almost unexplored field. Among the most significant of these drawbacks is the possibility that the method may result in biased samples. We have succeeded in finding only one paper that explicitly addresses the problem of biases: in their study of Polish migrants in the UK, McGhee et al. (2017) gathered data online via Polish-language Facebook groups and Polish online media. Before presenting their substantial results, the authors compared the composition of their sample with the Annual Population Survey ("the largest ongoing household survey in the UK, based on interviews with the members of randomly selected households" [Nomis. Official labour market statistics, 2020]). They documented several discrepancies in the socio-demographic characteristics of the two samples in terms of gender, age, and education, but concluded that the offline Annual Population Survey also did not provide a representative picture of migrants. Significantly, they did not propose any further steps to remedy the problem.

In this chapter, we step away from a focus on the "bright side" of the method and instead explore the biases it may present and propose solutions that may help to countervail them. We base our analysis on material from the five web-based migrant surveys we conducted using targeting on SNS. First, we describe the procedure for surveying migrants by targeting them on SNS, then we provide an outline of the major challenges we identified, and lastly, we delineate possible solutions, which we illustrate with material from one of the surveys. We conclude that, at present, the range of biases remains more considerable than the opportunities to adjust for them. Thus, it could be time to concede to this difficulty and instead direct our efforts to exploring other approaches to data analysis and presentation that are more suitable for the contexts of uncertainty.

[3] The chapter is based on the following five surveys, which recruited migrant participants using SNS advertising between 2016 and 2018: (1) a trilingual survey of Uzbek and Tajik migrants in Russia within a study of migration legislation (2016, N = 2412), (2) a bilingual survey of Uzbek migrants in Russia as part of a project on migrants' labor market participation (2017, N = 375), (3) a survey of Armenian and Azeri migrant children aged 18–35 y.o. (2017, N = 302), (4) a survey of migrant children aged 18–35 y.o. from Armenia, Azerbaijan, Uzbekistan, Tajikistan, Kyrgyzstan, and Ukraine, in Russia, together with their non-migrant peers (2018, N = 12,524), and (5) a trilingual survey of female migrants from Kyrgyzstan and Tajikistan in Russia aimed at studying their reproductive behavior (2018, N = 1000). Projects 1 and 3–5 were conducted with the financial support of the Government of the Russian Federation; project 2 was conducted with the financial support of the World Bank. The opinions, findings, conclusions, and recommendations expressed in this publication are those of the authors and do not necessarily reflect the view of the World Bank or the Government of the Russian Federation. Hereafter, we consider *first-generation migrants* to be those who were born outside of Russia and *second-generation migrants* to be those who have at least one migrant parent and who themselves were either born in Russia or moved at an early age, so that they graduated from a school in Russia.

3.2 Online Survey with Targeting on SNS: Description of the Procedure

In this section, we provide a general outline of the procedure for conducting surveys using targeting methods on SNS. This includes creating a questionnaire and an advertisement, as well as defining targeting criteria.

A starting point is uploading a questionnaire to a special online service for conducting surveys (e.g., SurveyMonkey) and creating an advertisement that includes a link to the survey on an SNS. An advertisement consists of an image, explanatory text, and a motivational button (Fig. 3.1). In our surveys, the pictures may contain national symbols (such as flags) or photographs of people and landscapes (Fig. 3.2). The explanatory text is intended to appeal to a target audience and thus to serve as an additional sorting and attracting mechanism alongside the specified targeting variables. If the rules of the SNS allow for it, migrants' native languages can be used in the advertisement.[4] After clicking on the advertisement, a user is directed to the survey's landing page, which includes a language selection menu (if necessary) and a welcome text providing information about the survey. On the last page of the questionnaire, researchers may express their gratitude to the survey participants and share links to their website and SNS pages for the project and/or research team.

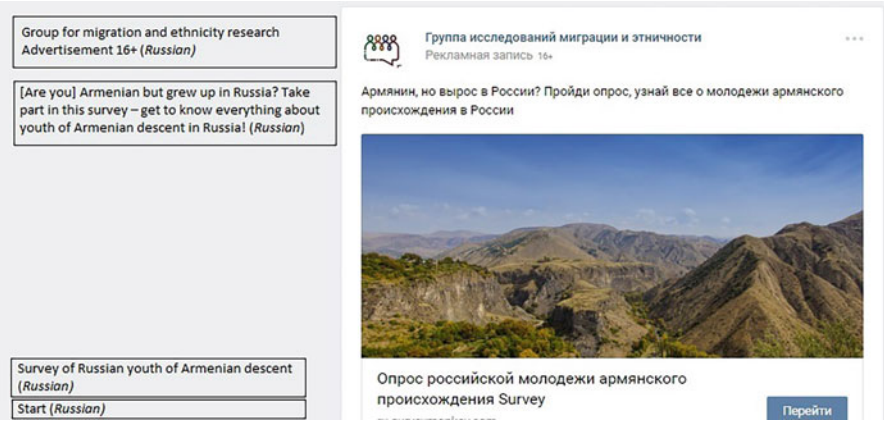


Fig. 3.1 Example of a banner used for male respondents of Armenian descent to advertise a survey of second-generation migrants on SNS (2017)

[4] The most popular SNS in Russia—Vkontakte and Odnoklassniki—adhere to Russian legislation for advertising and thus limit the usage of languages other than Russian. By contrast, Facebook and Instagram do not have such limitations.

Fig. 3.2 Example of a banner used to advertise a survey of Tajik female first-generation migrants on SNS (2018)

The next step involves choosing targeting criteria that effectively define the audiences on a given SNS that the researchers seek to reach via advertisements. SNS differ in the targeting options they support. The most popular SNS worldwide, Facebook, shares an advertising platform with Instagram. This platform offers several ready-made targeting criteria that can be appropriate for migration studies, such as "expats from country *X* in country *Y,*" but they do not cover all nationalities comprehensively. For example, criteria for migrants from Central Asian countries in Russia are not available. In such cases, researchers can use other features of targeting, for example, the "interests" category. When conducting a Facebook survey of second-generation migrants from Central Asia, South Caucasus, and Ukraine in Russia, we chose an intersection of two features: location (Russia) and interests (interests that can be described using keywords related to the country or culture of their parents). As an example, our list of interests for targeting second-generation migrants from Kyrgyzstan included *Kyrgyzstan, Kyrgyz language*, and three major cities—*Bishkek, Osh*, and *Talas*.

However, in some countries Facebook may be significantly less popular as compared with other social media or altogether unavailable. Such cases include China and Iran (where Facebook is banned) and the post-Soviet countries where Facebook does operate, but on a smaller scale than other mostly local social media. In Russia, as of 2018, the most popular SNS are the following (in descending order): Vkontakte (36 million users who posted at least once during the month preceding the study), Instagram (24 million), Odnoklassniki (16 million), Facebook (2 million), Twitter (0.8 million), Moi Mir (0.099 million) (Brand Analytics, 2018). The prominence of Odnoklassniki, Vkontakte, and Instagram remains true for the three Central Asian states sending migrants to Russia—Kyrgyzstan, Tajikistan, and Uzbekistan (The Open Asia, 2020). In the South Caucasus, the situation is different: in Armenia, Vkontakte is considered the most popular SNS, but Facebook outstrips Odnoklassniki (Sputnik Armenia, 2018), whereas in Azerbaijan the largest share of the social media market is held by Facebook and Instagram (Midia.Az, 2018). Data on SNS usage among migrants is unavailable, but we can hypothesize that their

situation would broadly reflect the SNS ratings in Russia and their country of origin. Thus, depending on the focus of the study, we used a combination of Odnoklassniki, Vkontakte, Instagram, and Facebook for our various surveys of first- and second-generation migrants.

The two main Russian SNS, Vkontakte and Odnoklassniki, are owned by the Mail.ru Group and offer quite similar targeting options (both differ from Facebook). Mail.ru Group has its own advertising platform, MyTarget, which until 2019 was the only means of advertising on Odnoklassniki (since then Odnoklassniki has developed its own advertising provisions). The users of MyTarget can access all of the advertising options provided by Odnoklassniki but only some of the options provided by Vkontakte. The latter, however, also offers separate advertising options. Neither Vkontakte nor MyTarget support targeting criteria such as "expats from *X* in *Y*." Moreover, their "interests" categories are in a fixed format (menu selection) and do not contain any country/culture-related options. However, from our fieldwork, we know that at least some people with a migrant background in Russia (both first- and second-generation) participate in social media groups with ethnic connotations, such as "Uzbeks in Moscow" or a group of Tajik humorous anecdotes. Therefore, we compiled lists of such groups on Vkontakte and Odnoklassniki and used them as a basis for targeting along with other requirements relevant for each situation (e.g., age, gender, location, or place of residence). In terms of geographical criteria, both Vkontakte and MyTarget differ from Facebook. When setting up an advertisement campaign on Facebook, advertisers (or researchers) select locations and define whether targeted users reside there, have recently been there, or currently are travelling there. Vkontakte and MyTarget offer two options with respect to geography. The first involves setting up an ad campaign that will be disseminated across specific countries, regions, or cities, but without an option to specify users' relations with these locations. In choosing users according to this criterion, Vkontakte analyzes the information that users provide on their profiles, whereas MyTarget analyzes IP addresses, i.e., users' current locations. The second option involves selecting one or several dots on a map with a radius of up to 10 km (MyTarget) and up to 40 km (Vkontakte) and defining users' relations with these locations. When launching a national campaign on Vkontakte or MyTarget, it is not feasible to select the specific relations users have with locations, and so the possible choices are restricted to either a user's designated place of residence (Vkontakte) or current location (MyTarget). In certain contexts, current location may serve as a useful parameter, whereas in others, it may be too crude, or even misleading. An example of the latter situation would be research contexts involving significant tourist flows between countries, alongside migrant flows. However, this is not the case for migration flows between Central Asia or the South Caucasus and Russia where economic migrants constitute the vast majority of flows.

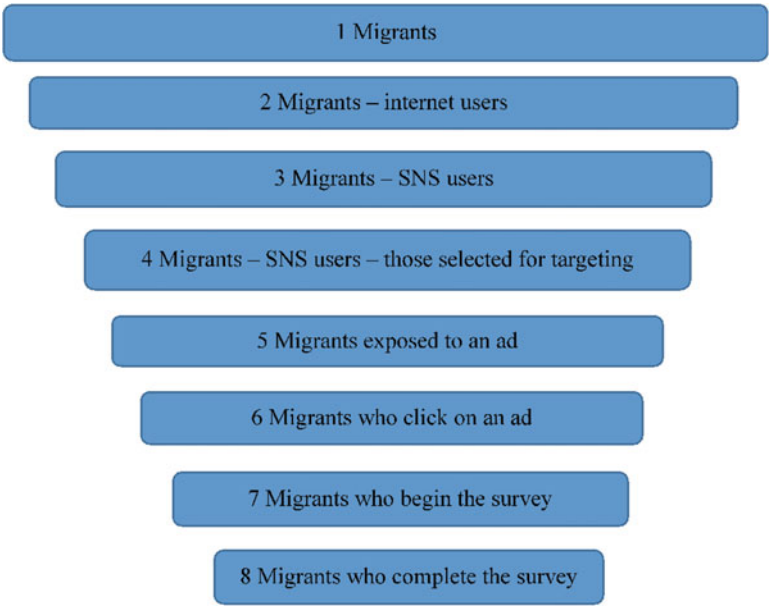


Fig. 3.3 Biases of a web-based migrant survey using targeting on SNS

3.3 Biases and the Search for Solutions

One of the most serious challenges involved in targeting respondents with a migrant background on SNS for web-based surveys concerns biases. In this section, we discuss these biases and propose possible solutions for mitigating them.

The biases manifest in the sampling of migrants on SNS can be shown as a complicated structure with several layers (Fig. 3.3). To begin, not all migrants use the Internet (1–2). Among those who do, not all are registered on SNS, and some of those who are registered do not use their accounts (2–3). Those who have accounts on SNS may follow different patterns of online behavior, and therefore, may be classified in different ways by the SNS. Such differences may structure their potential to be selected as part of the target audience for a given advertisement (3–4). Moreover, migrants may have different habits in their SNS usage (3–5), for example, an individual may have accounts on several SNS, whereas another may have only one SNS account; and another may be highly active on SNS, participating in multiple groups, whereas another may log in only to exchange messages. Due to these variations, individuals' chances of seeing an ad are not equal. In addition, some users who qualify to be targeted and could be exposed to an ad are not selected by the advertisement algorithms of the SNS (4–5). Users also may either click on an ad, skip it, or miss it altogether while inattentively scrolling their news feed (5–6). Finally, not all of the users who click on a given ad necessarily proceed to the survey questionnaire (6–7) and, among those who do, not all will go on to complete it (7–8).

Table 3.1 Results of logistic regression analysis (dependent variable: "usage of SNS," 0 no, 1 yes)

	B	Exp(B)
Gender (0-female, 1-male)	−0.434	0.648
Age	−0.133***	0.876
Income	0.000	1.000
Number of trips to Russia	−0.040	0.961
Region of birth in Kyrgyzstan (0-south, 1-north)	0.033	1.034
Place of birth (0-rural, 1-urban)	−0.092	0.912
Russian language level (ref = "speaks fluent Russian")		
Does not speak any Russian	−1.230*	0.292
Speaks some Russian	−0.781*	0.458
Education (ref = "secondary education")		
Vocational education	0.052	1.053
Higher education	0.655	1.925
Constant	4.957	142.223
Pseudo R2	0.350	
N	314	

***$p < 0.001$, **$0.001 \leq p < 0.01$, *$0.01 \leq p < 0.05$

Based on these observations, we can hypothesize that the respondents who eventually complete a survey could differ considerably from our hypothetical sampling frame.

Addressing this complicated set of biases demands different approaches. To begin, we need a better understanding of migrants' Internet and SNS usage and engagement, which would help with estimating the probability of the participation of various categories of migrants in surveys. This probability is determined by various factors: one migrant does not use the Internet or social media at all, one indicates that she/he comes from a specific country of origin, one hides this fact, another spends a lot of time online, one pays scant attention to advertisements when scrolling their feed, and so on. These factors correspond to layers 1–6 in Fig. 3.3. To the best of our knowledge, no comprehensive study on this topic has been conducted with respect to the Russian context, and internationally the number of such studies remains low (Madianou & Miller, 2013; Law & Chu, 2008). However, we can use the small amount of data that we do have to hypothesize which characteristics differentiate those migrants who use SNS from those who do not (layers 1–3 in Fig. 3.3). Thus, we performed a regression analysis on data from a face-to-face survey of Kyrgyz migrants in Moscow in 2014[5] (Varshaver et al., 2014) where we asked whether our respondents used Vkontakte, Odnoklassniki, Facebook, or Moi Mir (the most popular SNS at the time) (Table 3.1). As aforementioned, 63% of these respondents used SNS. In our regression model, the dependent variable was having an account on at least one of the four specified SNS. The independent variables included age,

[5]This was a bilingual survey conducted within a 500 meters radius of 50 randomly selected metro stations in Moscow (N = 350).

Russian language proficiency, education level, gender, urban/rural place of birth, region of birth in Kyrgyzstan (south or north), income, and number of trips to Russia.[6]

The two statistically significant factors are age and Russian language proficiency: SNS are more actively used by those who speak fluent Russian and those who are younger. While it is not surprising that age is a significant factor in these calculations, we can only hypothesize as to why Russian language proficiency matters, especially given that education does not. One plausible explanation is that even though currently, all the popular SNS offer an option to set up an interface in almost any language of the former Soviet Republics, when the SNS were first introduced in the post-Soviet space, they were initially only supported in Russian and, therefore, were most accessible for those with Russian proficiency. If this is true, the significance of the Russian language may now be less high than it was in 2014. However, if other explanations are plausible, we can expect the significance of the Russian language to remain high. When assessing these findings, it is important to bear in mind the limitations of the 2014 survey. First, the data are by now somewhat outdated, since the SNS landscape is very dynamic. Second, since the survey was limited to Kyrgyz migrants and only those residing in Moscow, we can hypothesize that other migrant groups in different locations could entail quite different outcomes. Regarding other ethnic groups, e.g., Tajik and Uzbek, gender also may factor in SNS usage due to the different constructions of gender relations among the Kyrgyz, Tajik, and Uzbek migrants (Rocheva & Varshaver, 2017). Third, since the survey was conducted in the vicinity of randomly selected metro stations, it may have omitted those migrants who rarely use the metro, e.g., drivers (including taxi drivers), janitors who use bikes, housewives, and so forth. Nevertheless, the results of this survey indicate that online surveys using SNS-based sampling can yield findings that are biased towards those who are younger and speak better Russian.

Another approach to mitigating a selection bias stemming from SNS targeting methods would be to conduct two surveys with the same questionnaire and target population: one would be a face-to-face or telephone survey structured as closely as possible to be random, and the other would be an online survey using SNS-based sampling. A comparison of the results would enable the calculation of propensity weights for subsequent use (c.f. Lee, 2006; Terhanian & Bremer, 2012). We can hypothesize two designs for a propensity score adjustment that appear to be feasible, although resource-intensive, for the Russian context. The first would involve conducting a random face-to-face survey of foreign students—since statistics exist that reveal the distribution of foreign students across Russian universities—in parallel with an online survey of foreign students on several SNS. The second

[6]Migration from Central Asia, including Kyrgyzstan, to Russia is transnational and involves a variety of migration regimes, which include seasonal migration, spending a year in Russia with summer vacations at home, permanent residence in Russia, as well as more complicated patterns. The aims of migration vary as well. However, by "number of trips," we mean all trips to Russia, irrespective of their goals and length.

would be to study several locations with a high concentration of migrants, using both an online survey with targeting on SNS and a face-to-face random survey.

Moreover, it would be useful to study how SNS construct their different target variables (e.g., how Facebook defines who is an "expat from Poland in the UK" or who has interests related to "Armenia, Yerevan, and the Armenian language") and how SNS select which users are exposed to an ad among all those to whom the specified target variables apply. This study would correspond with layers 3–5 in Fig. 3.3. To date, SNS have refrained from disclosing this information, since they consider it to be commercially sensitive and have not shown much interest in cooperating with researchers regarding these matters.

Finally, yet importantly, we need to further explore the interaction between a respondent and a questionnaire, including its welcome text, so to understand better who is more likely to leave a survey page before starting the questionnaire, or to drop out of a survey without completing it (layers 6–8 in Fig. 3.3). The matter of suitable questionnaire length is one example of conventional wisdom on these matters, but there may be further issues of specific relevance for respondents with a migrant background.

These directions for further research would appear to be long-term goals. Are there any "tactical" steps apart from them that we can take to enhance the results from research using SNS targeting methods? We think there are.

First, we can assess whether any dropout bias exists by comparing "completers" with those who drop out. Second, we can do an external validation that compares our results with data from a different source, such as available statistics. Where these statistics are lacking, such validation can take less conventional forms. For example, when we carried out a survey of second-generation migrants and their local peers in Russia using SNS-based sampling, and we lacked statistics on second-generation migrants, we compared the distribution of our respondents across various Russian regions, and the distribution of their ethnicities, with the corresponding characteristics of the Russian population provided in the Russian census. Drawing on this comparison, we checked our results for any significant discrepancies and identified only those we had expected and were able to account for.

A third possible tactical intervention involves weighting the data according to the results of a dropout analysis or external validation. A set of methods are available for reducing bias in online surveys, which are based on comparing the survey data with other data considered representative of the target population and adjusting it accordingly. Adjustment experiments reveal that a basic procedure of weighting obtains almost the same results as other more complicated procedures. Moreover, it has been found that the complexity of statistical procedures is less important to an effective bias reduction than is the choice of variables for the adjustment (Mercer et al., 2018). With respect to the choice of variables, optimal results are achieved when researchers take into account not only demographic variables, but also behavioral and attitudinal ones (e.g., political views, health, internet usage, etc.) (Taylor, 2000; Mercer et al., 2018). However, one problem is that in migration studies, the range of external representative data is very limited, and so it is difficult to devise an

extensive set of behavioral and attitudinal variables for weighting. Nonetheless, weighting can be based on limited statistics and results from a dropout analysis.

These tactical actions can be undertaken in different combinations, depending on the data. In the next section, we further illustrate their applicability by commenting on a survey of migrants from Uzbekistan in Russia.

3.4 Illustrative Example: Survey of Migrants from Uzbekistan in Russia

In March–June 2017, we conducted research into the labour market participation of migrants in Russia. This included a survey of Uzbek migrants (May 2017) who constitute the largest group of labour migrants in the country. We implemented the online questionnaire in SurveyMonkey and disseminated advertisements targeting participants on two SNS—Odnoklassniki and Vkontakte. We targeted SNS users who were at least 18 years old and who participated in or liked groups/pages with ethnic connotations (e.g., "Uzbeks in Moscow"). The advertisements were shown to those users who resided or were located at the time in Russia. We were less concerned with the risk of getting too many people who were just visiting Russia as guests or tourists, since the majority of the flows from Uzbekistan to Russia, at the time, was, and still is, comprised of people travelling for economic reasons.

The choice of SNS is an important step. Ideally, we would have known the shares of Uzbek migrants using different SNS and set up our campaigns to reach the necessary number of respondents on each SNS. However, so far, we only know the number of users of different SNS in Uzbekistan. Over two million Uzbek users of Odnoklassniki visit this SNS monthly (77% are male) (Odnoklassniki, 2019), whereas Vkontakte has one million monthly Uzbek users (Infocom, 2018). Instagram and Facebook have 0.89 and 0.72 million Uzbek users each, respectively (Infocom, 2018). This ranking was also supported in our interviews for previous research projects. Therefore, for this survey, we chose the two most popular SNS, Odnoklassniki and Vkontakte, and decided to allow for a "natural flow" by assuming that readiness to respond would be equivalent on the two SNS and that each would therefore contribute a number of respondents proportional to the popularity of these sites among these migrants.

Language is another issue. Odnoklassniki and Vkontakte limit the usage of languages other than Russian. This limitation was problematic, since we wanted to stress the Uzbek character of our advertisement. To circumvent these restrictions, we chose to use the colors of the Uzbek flag in the advertisement and included our question, within the image, in Uzbek: "Do you work in Russia?" («Россияда ишлаяпсизми?»).[7] However, the main text accompanying the ad was in Russian:

[7] Although the official alphabet in Uzbekistan changed from being Cyrillic-based to Latin-based in 1993, we chose to use the Cyrillic version, since it still has widespread use in the country.

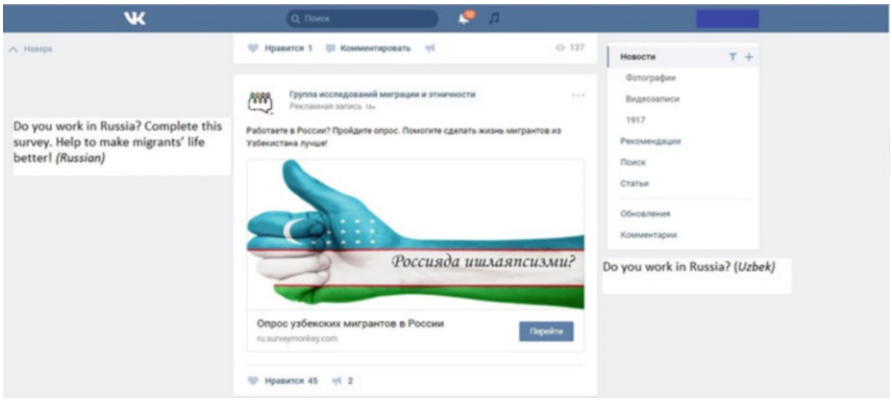


Fig. 3.4 Screenshot of an ad for our survey of Uzbek migrants in Russia

"Do you work in Russia? Complete this survey. Help to make migrants' life better!" (Fig. 3.4). At the beginning of the survey, respondents could choose whether they wanted the questionnaire to be displayed in Uzbek or Russian.

In total, 1099 individuals chose a language on the survey landing page (51% selected Uzbek and 49% Russian), and 865 responded to the first question. A total of 388 went on to complete the survey. With respect to the 1099 individuals, 66% (729 individuals) were recruited through Odnoklassniki and 34% (370 individuals) through Vkontakte.

The accuracy of our targeting can be measured in four dimensions. First, of the 803 individuals who responded to a question regarding their current location, only 5 (0.6%) said they were not presently in Russia. Second, out of the 865 respondents who provided their year of birth, 17 (2%) were underage (less than 18 years old). The third dimension relates to migrant background and the Uzbek/Uzbekistan connection, which can be measured in two ways: place of birth and citizenship. Out of the 865 respondents who replied to the question asking for their place of birth, 77% were born in Uzbekistan and another 16% in neighboring countries in Central Asia. Out of the 865 respondents who provided an answer to the question about citizenship, 655 (76%) had Uzbek citizenship. Last, we targeted individuals who were not just visiting Russia. Out of the 701 respondents who answered the question as to what they were doing in Russia, 661 people (79%) said they were working, 10% working and studying, 6% studying, 4% were occupied with housework; 1% chose the option "other," mentioning that they simply lived in Russia, and only 1 person defined himself as a guest. Thus, our cleaned dataset included respondents who were at least 18 years of age, who were located or resided in Russia, who were not just visiting, and who had Uzbek citizenship and/or were born in Uzbekistan. Regarding the entire dataset, 540 respondents fit this description, but not all of them completed the survey. A total of 303 respondents fit this description and completed the survey, so they are included in our final database. Overall, regardless of their characteristics,

388 respondents completed the survey. The ratio between 303 and 388 (78%) can be understood as accuracy rate. Who are the 85 respondents who completed the survey, but did not fit the selection criteria? The majority (67 out of 85) consisted of individuals from two countries that neighbor Uzbekistan—Tajikistan and Kyrgyzstan—of whom 26 self-identified as Uzbeks. So, if we take into account respondents who self-identified as Uzbeks yet come from other Central Asian countries, the accuracy rate rises to 84% (326 respondents out of 388). Nonetheless, in the subsection that follows, we use our stricter criteria for inclusion.

3.4.1 Dropout Rate and Language Bias

The completion rate for the respondents who fit our selection criteria was 56% (303 responded to the last question, out of the 540 individuals who responded to the minimal list of questions necessary to identify them as a fit). Although a conventional standard does not exist for the completion/dropout rate for web-based surveys, and even less so for migrants, some researchers have indicated 60% as an acceptable completion rate for web-based non-probability panels for the general population (Liu & Wronski, 2018). As an example, a web-based survey of Polish migrants in other European countries had a completion rate of 72% (Pötzschke & Braun, 2017), but web-based surveys of "hidden" populations, such as drug users, have had completion rates as low as 38.3% (Temple & Brown, 2011).

As we hypothesized previously, the completion rate could become a source of bias if respondents with different characteristics drop out, to different degrees. To test this hypothesis, we performed a logistic regression analysis (Table 3.2). The dependent variable was whether or not participants responded to the final question, whereas the independent variables included gender, age, questionnaire language, education level, and whether a respondent came to the survey via MyTarget/

Table 3.2 Results of a regression analysis (dependent variable "presence of response to the final question," 0 no, 1 yes)

	B	Exp(B)
Gender	0.289	1.336
Age	−0.001	0.999
Language (0 Russian, 1 Uzbek)	−0.636**	0.530
Education (ref = "secondary")		
Vocational	0.377	1.458
Higher	0.049	1.050
SNS (0 MyTarget/Odnoklassniki, 1 Vkontakte)	0.281	1.325
Constant	0.017	1.017
N	540	
Pseudo R2	0.053	

***$p < 0.001$, **$0.001 \leq p < 0.01$, *$0.01 \leq p < 0.05$

Odnoklassniki or Vkontakte. The only significant variable was language: if a respondent filled out a questionnaire in Russian, they were twice as likely to complete it.

3.4.2 External Validation

The two sets of data to which we compared our sample were the limited migration statistics and exposure data provided by the SNS. The official statistics provided by the former Federal Migration Service (FMS) of the Russian Federation until 2016[8] indicate the number of male and female foreign citizens of different age groups who were in Russian territory on a specific date, but do not disaggregate them by goals of entry. These statistics also do not include those who obtained Russian citizenship. Nevertheless, we can use the data for Uzbek citizens who were in Russian territory in 2016. The exposure data reveal how many users with different demographic characteristics were exposed to a given advertisement.

Our data (N = 303) included 72% male and 28% female respondents, but a comparison with the migration statistics found that women were overrepresented in our survey (Table 3.3). Migration from Uzbekistan and other Central Asian countries to Russia has a mostly male character (Rocheva & Varshaver, 2017). As regards age groups, our dataset included fewer respondents aged 40 years and older than were provided by migration statistics reports for this specific group, and more respondents aged 30–39 years. At the same time, quite surprisingly, the share of the youngest group aged 18–29 was almost the same in the statistics and our dataset.

The overrepresentation of women in our dataset can be explained by several factors, including women's more active usage of SNS or SNS groups with ethnic connotations, or their higher inclination to participate in (and complete) online surveys, as is suggested in previous research (Smith, 2008). In our dropout/

Table 3.3 Comparison of our dataset with migration statistics according to gender and age groups

	Survey			Migration statistics		
Age	Male	Female	Total (N)	Male	Female	Total (N)
18–29	42%	10%	52%	44%	6%	51%
30–39	18%	13%	31%	19%	5%	25%
40–49	11%	4%	15%	14%	4%	17%
50+	1%	2%	2%	5%	3%	7%
Subtotal	72%	28%	100% (303)	82%	18%	100% (1,646,098)
Total (N)	100% (303)			100% (1,646,098)		

[8]The Federal Migration Service was dissolved in 2016, and a new body, the Main Directorate for Migration Affairs, was established. Since then, available migration statistics have changed. Now, the data are disaggregated according to goal of entry and country of origin, but do not indicate gender or age.

Table 3.4 Conversion rates for Odnoklassniki and Vkontakte according to gender

		Male	Female	Total
Odnoklassniki	(A) SNS users exposed to an advertisement	96,285 (72%)	36,825 (28%)	133,110 (100%)
	(B) Respondents who completed questionnaires	123 (64%)	70 (36%)	193 (100%)
	Conversion rate (B/A)	0.13%	0.19%	0.15%
Vkontakte	(A) SNS users exposed to an advertisement	92,338 (70%)	40,274 (30%)	132,613 (100%)
	(B) Respondents who completed questionnaires	95 (86%)	15 (14%)	110 (100%)
	Conversion rate (B/A)	0.10%	0.04%	0.08%

completion analysis (Table 3.2), we found that gender was not a significant factor. We can assess whether women are more likely than men to participate in a survey if exposed to an advertisement by calculating a conversion rate, i.e., the ratio of respondents who completed a questionnaire to those who were exposed to an advertisement on SNS (Table 3.4). Whereas the Odnoklassniki conversion rate was higher for females than for males, the exact opposite was true in Vkontakte. Thus, we cannot conclude that females are more likely to participate in a survey after being exposed to an advertisement.

Interestingly, the share of women, among the users who were exposed to an advertisement, was similar (28–30%) on both SNS, and higher than the share of women among migrants in the available statistics. This finding leads us to suggest that women may be more active users of SNS or groups with ethnic connotations on SNS, which, in turn, may contribute to an explanation for the overrepresentation of women in our dataset as compared with the migration statistics.

3.4.3 Weighting

We have been able to identify several biases in our sample. First, respondents who selected the Uzbek language questionnaire were less likely to complete the survey. Second, women were more prone to participate in the survey than men on Odnoklassniki, whereas the opposite was true for Vkontakte. Third, in comparison with migration statistics, there were more females and people aged 30–39 and fewer people aged 40 and older in our dataset. Ideally, we would adjust our dataset according to these identified biases— sequentially, one by one—as if these were "layers" we wanted to restore. However, the procedure of weighting allows for only one step, not several. Moreover, we could not accommodate gender differences according to both the migration statistics and conversion rates within this step. Therefore, we opted for a combination of language dropout data and gender and age proportions from the migration statistics. Before we describe the weighting procedures, we need to make an assumption. Since migration statistics are based

on current nationality, and since our database contained citizens of both Uzbekistan and Russia, we had to assume that the proportions of men and women of different ages, among those who retained Uzbek citizenship and those who were naturalized, were equivalent.

The three variables we used for weighting were age, gender, and language. Since our sample was not that large, we used three age groups instead of four: 18–29 years old, 30–39 years old, and 40 years and older. We used the following formula to calculate the weighting coefficients (w):

$$w = \frac{k * m}{n}$$

where k is the share of the women/men of a specific age group in the total number of Uzbek citizens according to statistics; m is the share of female/male respondents of a specific age group who selected the Uzbek/Russian language, even though they might not have answered the rest of the questions; n is the share of female/male respondents of a specific age group in our final sample who selected the Uzbek/Russian language.

After weighting, the dataset included a higher proportion of those who selected the Uzbek language, fewer women, and more respondents of an older age (Table 3.5).

It may have been productive to compare the characteristics of the weighted database with some other migration statistics (regional distributions, occupations, etc.), but we used all the available statistics variables (gender and age) for weighting. However, we were able to check the changes of the variables in the dataset. We found that the weighting changed the distributions of the variables that were connected with a general orientation towards Russia or country of origin. First, the amount of remittances increased from 15,301 rubles before the weighting to 16,586 rubles after the weighting (approximately $212 and $229 correspondingly). Second, after the weighting, there was a larger share of those willing to return to the country of origin and a smaller share of those willing to stay in Russia or to live transnationally in two countries (Table 3.6).

Table 3.5 Results of weighting

	Before weighting		After weighting	
	Russian language	Uzbek language	Russian language	Uzbek language
Male 18–29	24%	18%	20%	24%
Male 30–39	8%	10%	8%	11%
Male 40+	6%	6%	8%	11%
Female 18–29	8%	2%	5%	2%
Female 30–39	7%	6%	3%	3%
Female 40+	3%	3%	3%	3%
Subtotal	55%	45%	47%	53%
Total (N)	100% (303)		100% (303)	

Table 3.6 Comparison of plans for the future, before and after the weighting

	Before weighting	After weighting
Live in Russia	16%	12%
Live in the country of origin	55%	61%
Live in Russia and the country of origin	25%	23%
Leave for another country (neither Russia nor country of origin)	4%	4%
Total	100% (303)	100% (303)

To summarize, targeting the participants of the Uzbek-connotated groups and pages on the two SNS popular in Russia and Uzbekistan proved to be an efficient method of sampling. We were able to get responses from our target group: those who were born in Uzbekistan and/or had Uzbek citizenship, who were at least 18 years of age, who were currently located or resided in Russia, and who were neither guests nor tourists. However, this method is associated with some biases, which we were able to partially compensate for using weighting. The weighting of the dataset according to gender, age, and survey language altered the distributions of some variables. Among these, some were associated with orientations towards Russia or Uzbekistan, namely, remittance behavior and migration intentions.

3.5 Discussion and Conclusion

This chapter contributes to the growing body of literature demonstrating the effectiveness of targeting on SNS as a sampling strategy for online migrant surveys in various contexts (Pötzschke & Braun, 2017). Its main goal, however, was to foster a discussion of the serious challenges of biased samples associated with this method, and to propose possible approaches to address this problem. The range of methods developed by scholars to adjust for biases in non-random surveys of other target populations (Baker et al., 2013) have limited applicability in the field of migrant studies due to a lack of sampling frames and, more generally, to limited knowledge about the characteristics of this target population. Methods such as propensity score adjustment are not easily applicable in our field, at least thus far. We have demonstrated how weighting based on dropout analysis and external validation can be used as hands-on solutions. Still, we need to note that weighting can in some cases exacerbate biases, if the underlying assumptions are incorrect. For example, our calculations of the weighting coefficients showed that we needed to increase the share of older respondents in accordance with migration statistics. This increase was based on the assumption that the older respondents in our sample did not differ significantly from older migrants who did not take part in the survey. However, if this assumption were to prove incorrect—for example, if significant differences existed between the older migrants who used SNS and those who did not (and

thus did not have an opportunity to participate)—our weighting would not have been an effective method of adjustment.

Transparency in the assumptions made and, more broadly, in the descriptions of the design and data analysis, is deemed an essential element for non-random survey designs. Thus, it is necessary for the assessment of a given study and its results, as well as for the advancement of the method (Baker et al., 2013). In the case of a study using targeting on SNS, scholars alone cannot provide fully transparent descriptions of their sampling strategy, since there remains an important piece of the puzzle which is lacking with respect to the operations of the SNS. For now, SNS do not disclose how they construct target variables or how they select users who are exposed to an advertisement among all those who fit the targeting criteria. Besides improving transparency, this information would help researchers to estimate the probability of different users' participation in a survey.

A realistic assessment of the present situation leads us to concede that the number of biases is considerable, whereas our opportunities to adjust for them are rather modest. If we resign ourselves to this fact, we might instead direct our efforts to the exploration of other approaches to analyze and present the data that may be a better fit for contexts involving high levels of uncertainty. At least two potential sources of inspiration stand out for this strand of exploration: the fuzzy set theory and Bayesian statistics.

Fuzzy set theory was introduced in the 1960s (Zadeh 1965), and since then, it has attracted the attention of different fields, including the social sciences (Ragin & Pennings, 2005; Smithson & Verkuilen, 2006). As a soft computing method, fuzzy set theory was developed to work with linguistic categories that often have blurred boundaries, as well as to deal with imprecise and incomplete data. Unlike classical (crisp) sets—in which an element belongs (1) or does not belong (0) to a set—a fuzzy set's belonging can vary on an interval from 0 to 1 (with 0.5 being the point of least certainty), which is described with a membership function. In addition to being applicable to categories such as "poor" (Lemmi & Betti, 2006) or "migrant," which do not always allow for strict definitions (using crisp logic, spending one more day in another country could change your status), fuzzy logic can help scholars to depart from the conventional manner of working with data. Such conventions imply the provision of exact figures for "well integrated" migrants or "average remittances." Using fuzzy logic can enable researchers to aim for a formulation of tendencies and approximate assessments. Thus, instead of providing a "precise" figure for the average income of a migrant, a scholar can define limits or provide an approximate figure, perhaps indicating a "credible" interval, so to emphasize the character of the data and, more broadly, the context of uncertainty.

In recent decades, Bayesian statistics has advanced due to improvements in computing technologies and algorithms, yet its usage in the social sciences remains, to date, very modest (Lynch & Bartlett, 2019; Western, 1999). Unlike more conventional frequentist statistics, analysis using Bayesian logic implies not only working with a specific dataset, but also taking into account "priors," which may include results from previous studies, as well as experts' or scholars' assessments of probability. Thus, Bayesian statistics are used to "blend" several data sources, for

example, SNS data and migration statistics to estimate the number of migrants in different states in the USA (Alexander et al., 2020), or various opinion polls regardless of their representativeness (Roshwalb et al., 2012). Regarding an online migrant survey that uses targeting on SNS, priors may include the results of a study of migrants' SNS usage practices or exposure data provided by the SNS. Bayesian statistics can also be useful when a survey is implemented that targets individuals on several SNS, where respondents using various SNS differ considerably in their characteristics.

To conclude, we have demonstrated a complicated set of biases that scholars face when conducting an online migrant survey based on SNS targeting, as well as extant possibilities for adjusting for some of these biases. However, these remedies seem insufficient to redress the potentially strong and rather unpredictable distortions caused by those biases. Therefore, we surmise that it might be time to explore other avenues of working with such data in contexts of uncertainty, for example, fuzzy set theory and Bayesian statistics.

References

Agadjanian, V., & Zotova, N. (2012). Sampling and surveying hard-to-reach populations for demographic research: A study of female labor migrants in Moscow, Russia. *Demographic Research, 26*, 131–150.

Alexander, M., Polimis, K., & Zagheni, E. (2020). Combining social media and survey data to nowcast migrant stocks in the United States. *arXiv preprint arXiv*:2003.02895.

Baker, R., Brick, J. M., Bates, N. A., Battaglia, M., Couper, M. P., Dever, J. A., Gile, K. J., & Tourangeau, R. (2013). Summary report of the AAPOR task force on non-probability sampling. *Journal of Survey Statistics and Methodology, 1*, 90–143.

Brand Analytics. (2018). *Social networking sites in Russia* (autumn 2018). https://br-analytics.ru/blog/wp-content/uploads/2018/12/Sotsseti-Rossiya-osen-2018.pdf. Accessed 10 Sept 2019.

Bucholtz, I. (2018). Bridging bonds: Latvian migrants' interpersonal ties on social networking sites. *Media, Culture & Society*, 016344371876457.

Careja, R., & Bevelander, P. (2018). Using population registers for migration and integration research: Examples from Denmark and Sweden. *Comparative Migration Studies, 6*(1), 19.

Dekker, R., Engbersen, G., & Faber, M. (2015). The use of online media in migration networks. *Population, Space and Place, 22*(6), 539–551.

du Plooy, D. R., Lyons, A., & Kashima, E. S. (2018). The effect of social support on psychological flourishing and distress among migrants in Australia. *Journal of Immigrant and Minority Health*, 1–12.

Dubois, A., Zagheni, E., Garimella, K., & Weber, I. (2018). Studying migrant assimilation through Facebook interests. In *International conference on social informatics* (pp. 51–60). Springer.

Herdağdelen, A., Adamic, L., & Mason, W. (2016, May). The social ties of immigrant communities in the United States. In *Proceedings of the 8th ACM conference on web science* (pp. 78–84). ACM.

Herz, A. (2015). Relational constitution of social support in migrants' transnational personal communities. *Social Networks, 40*, 64–74.

Hoefer, M., Rytina, N. F., & Baker, B. (2012). *Estimates of the unauthorized immigrant population residing in the United States: January 2011* (pp. 1–7). Department of Homeland Security, Office of Immigration Statistics.

Hu, J., & Wang, Z. (2015). Exploring the associated factors of elevated psychological distress in a community residing sample of Australian Chinese migrants. *Australian Journal of Psychology, 68*(2), 116–122.

Infocom. (2018). *Publication of the most popular services, application and phones of Internet users in Uzbekistan*. http://infocom.uz/2018/01/22/opublikovany-samye-populyarnye-servisy-prilozheniya-i-telefony-internet-polzovatelej-uzbekistana/#gallery-3. Accessed 10 Sept 2019.

Kühne, S., & Kroh, M. (2017). *The 2015 IAB-SOEP migration study M2: Sampling design, nonresponse, and weighting adjustment* (No. 473). SOEP Survey Papers.

Law, P. L., & Chu, W. C. R. (2008). ICTs and migrant workers in contemporary China. *Knowledge, Technology & Policy, 21*(2), 43–45.

Lee, S. (2006). Propensity score adjustment as a weighting scheme for volunteer panel web surveys. *Journal of Official Statistics, 22*(2), 329.

Lemmi, A. A., & Betti, G. (Eds.). (2006). *Fuzzy set approach to multidimensional poverty measurement* (Vol. 3). Springer.

Liu, M., & Wronski, L. (2018). Examining completion rates in web surveys via over 25,000 real-world surveys. *Social Science Computer Review, 36*(1), 116–124.

Lynch, S. M., & Bartlett, B. (2019). Bayesian statistics in sociology: Past, present, and future. *Annual Review of Sociology, 45*(1), 47–68. https://doi.org/10.1146/annurev-soc-073018-022457

Madianou, M., & Miller, D. (2013). *Migration and new media: Transnational families and polymedia*. Routledge.

Mateos, P., & Durand, J. (2012). Residence vs. ancestry in acquisition of Spanish citizenship: A netnography approach. *Migraciones Internacionales, 6*(4), 9–46.

McGhee, D., Moreh, C., & Vlachantoni, A. (2017). An "undeliberate determinacy"? The changing migration strategies of Polish migrants in the UK in times of Brexit. *Journal of Ethnic and Migration Studies, 43*(13), 2109–2130.

Mercer, A., Lau, A., & Kennedy, C. (2018). *For weighting online opt-in samples, what matters most*. Pew Research Center. https://www.pewresearch.org/methods/2018/01/26/for-weighting-online-opt-in-samples-what-matters-most. Accessed 10 Sept 2019.

Midia.Az. (2018). *Какие соцсети наиболее популярны в Азербайджане? Данные и рейтинг* [What social networks are most popular in Azerbaijan? Data and rating]. https://media.az/society/1067721256/kakie-socseti-naibolee-populyarny-v-azerbaydzhane/. Accessed 06 June 2020.

Miller, P. G., & Sønderlund, A. L. (2010). Using the internet to research hidden populations of illicit drug users: A review. *Addiction, 105*(9), 1557–1567.

Milton, A. C., Ellis, L. A., Davenport, T. A., Burns, J. M., & Hickie, I. B. (2017). Comparison of self-reported telephone interviewing and web-based survey responses: Findings from the second Australian young and well national survey. *JMIR Mental Health, 4*(3), e37.

Mkrtchyan, N. V. (2011). Population dynamics of Russia's regions and the role of migration: Critical assessment based on the 2002 and 2010 censuses. *Regional Research of Russia, 1*(3), 228.

Moreh, C. (2019). *Online survey design and implementation: Targeted data collection on social media platforms*. SAGE.

Nomis. Official labour market statistics. (2020). *Annual population survey*. https://www.nomisweb.co.uk/home/release_group.asp?g=16. Accessed 06 June 2020.

Odnoklassniki. (2019). *The results of the year*. https://insideok.ru/blog/itogi-goda-odnoklassnikov-v-belarusi-uzbekistane-i-kazahstane. Accessed 10 Sept 2019.

Pötzschke, S., & Braun, M. (2017). Migrant sampling using Facebook advertisements: A case study of Polish migrants in four European countries. *Social Science Computer Review, 35*(5), 633–653.

Prandner, D., & Weichbold, M. (2019). Building a sampling frame for migrant populations via an onomastic approach–lesson learned from the Austrian immigrant survey 2016. *Survey Methods: Insights from the Field* (SMIF).

Ragin, C. C., & Pennings, P. (2005). Fuzzy sets and social research. Special Issue. *Sociological Methods & Research, 33*(4), 423–573.
Reichel, D., & Morales, L. (2017). Surveying immigrants without sampling frames–evaluating the success of alternative field methods. *Comparative Migration Studies, 5*(1), 1.
Rocheva, A., & Varshaver, E. (2017). Gender dimension of migration from Central Asia to the Russian Federation. *Asia-Pacific Population Journal, 32*(2), 87–136.
Roshwalb, A., El-Dash, N., & Young, C.A. (2012). *Towards the use of Bayesian credibility intervals in online survey result.* Ipsos White Paper.
Salentin, K. (2014). Sampling the ethnic minority population in Germany. The background to "migration background". *Methods, data, analyses, 8*(1), 28.
Salentin, K., & Schmeets, H. (2017). Sampling immigrants in the Netherlands and Germany. *Comparative Migration Studies, 5*(1), 21.
Sanguilinda, I. S., di Belgiojoso, E. B., Ferrer, A. G., Rimoldi, S. M. L., & Blangiardo, G. C. (2017). Surveying immigrants in Southern Europe: Spanish and Italian strategies in comparative perspective. *Comparative Migration Studies, 5*(1), 17.
Smith, G. (2008). *Does gender influence online survey participation? A record-linkage analysis of university faculty online survey response behavior* (ERIC Document Reproduction Service No. ED 501717).
Smithson, M., & Verkuilen, J. (2006). *Fuzzy set theory: Applications in the social sciences* (No. 147). Sage.
Sputnik Armenia. (2018). *Один в пролете, другой – в шоколаде: какие соцсети предпочитают армяне* [One in the span, the other in chocolate: which social networks do the Armenians prefer]. https://ru.armeniasputnik.am/society/20180202/10384192/odin-v-prolete-drugoj-v-shokolade-kakie-socseti-predpochitayut-armyane.html. Accessed 06 June 2020.
Spyratos, S., Vespe, M., Natale, F., Weber, I., Zagheni, E., & Rango, M. (2018). *Migration data using social media*. https://ingmarweber.de/wp-content/uploads/2018/06/Migration-Data-using-Social-Media-a-European-Perspective.pdf. Accessed 10 Sept 2019.
Sue, V. M., & Ritter, L. A. (2012). *Conducting online surveys*. Sage.
Taylor, H. (2000). Does internet research work? *International Journal of Market Research, 42*(1), 1–11.
Temple, E. C., & Brown, R. F. (2011). A comparison of internet-based participant recruitment methods: Engaging the hidden population of cannabis users in research. *Journal of Research Practice, 7*(2), D2–D2.
Terhanian, G., & Bremer, J. (2012). A smarter way to select respondents for surveys? *International Journal of Market Research, 54*(6), 751–780. https://doi.org/10.2501/IJMR-54-6-751-780
The Open Asia. (2020). *Какие социальные сети популярны в центральной Азии* [What social networks are popular in central Asia]? https://theopenasia.net/ru/post/kakie-sotsseti-populyarny-v-tsentralnoy-azii. Accessed 06 June 2020.
The World Bank. (2019). *Individuals using the Internet (% of population)*. https://data.worldbank.org/indicator/IT.NET.USER.ZS. Accessed 10 Sept 2019.
Toepoel, V. (2015). *Doing surveys online*. Sage.
UNECE. (2019). *Guidance on data integration for measuring migration*.
Varshaver, E., Rocheva, A., Kochkin, E., & Kuldina, E. (2014). *Kyrgyz migrants in Moscow: Results of a quantitative research on integration tracks*. Russian Presidential Academy of National Economy and Public Administration. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2425312. Accessed 10 Sept 2019
Vigneswaran, D. (2009). Residential sampling and Johannesburg's forced migrants. *Journal of Refugee Studies, 22*(4), 439–459.
Wei, L., & Gao, F. (2017). Social media, social integration and subjective well-being among new urban migrants in China. *Telematics and Informatics, 34*(3), 786–796.

Western, B. (1999). Bayesian analysis for sociologists: An introduction. *Sociological Methods & Research, 28*(1), 7–34. https://doi.org/10.1177/0049124199028001002

Zadeh, L. A. (1965). Fuzzy sets. *Information and Control, 8*(3), 338–353.

Zagheni, E., Weber, I., & Gummadi, K. (2017). Leveraging Facebook's advertising platform to monitor stocks of migrants. *Population and Development Review, 43*(4), 721–734.

Zotova, N., Agadjanian, V., Isaeva, J., & Kalandarov, T. (2016). *Implementation of respondent driven sampling for hard-to-reach populations: A survey of female migrants in Nizhniy Novgorod, Russia*. http://www.academia.edu/download/44507367/PAA_2016_paper.pdf. Accessed 10 Sept 2019.

Chapter 4
Web-Based Respondent-Driven Sampling in Research on Multiple Migrants: Challenges and Opportunities

Agata Górny and Justyna Salamońska

4.1 Introduction

Migrants are an example of what survey researchers call a "hard-to-survey population" (Tourangeau et al., 2014). Challenges in surveying migrants can occur at various stages of the survey process. In many countries, sample frames for surveying migrants are not available, which makes this group *hard-to-sample*. If migrant populations are hidden, subject to social stigma or even persecution (as may be the case with undocumented migrants), they also remain *hard-to-identify*. Migrants also may be *hard-to-reach,* since they may be highly mobile people or international elites living in gated communities or they may be less privileged migrants living in more vulnerable areas difficult to access by researchers. Additionally, issues such as language barriers may make migrants *hard-to-persuade* to take part in research, and, even once they agree to take participate, *hard-to-interview*.

These challenges, however, make up for the one of the most fascinating features of migration studies: as researchers we constantly learn about and test-and-try to address these issues. This chapter testifies to this learning process by focusing on a particularly challenging sub-group of migrants—the multiple migrants who have migrated more than once and to more than one destination country. We know relatively little about multiple migrants, but we can assume that they are a rather small population and flexible with respect to changing their destinations (Jancewicz & Salamońska, 2020). Consequently, researching these migrants involves a wide array of challenges related to hard-to-survey populations, which demands an innovative methodological approach beyond the traditional survey methods.

To study multiple migrants from Poland (within the MULTIMIG project, which we describe in more detail in one of the following sections on survey design and

A. Górny (✉) · J. Salamońska
University of Warsaw, Warsaw, Poland
e-mail: a.gorny@uw.edu.pl; jj.salamonska@uw.edu.pl

S. Pötzschke, S. Rinken (eds.), *Migration Research in a Digitized World*, IMISCOE Research Series, https://doi.org/10.1007/978-3-031-01319-5_4

implementation), we employed a Web-based survey to deal with their spatial dispersion and high mobility. Due to the lack of a sampling frame, we decided to use the respondent-driven sampling (RDS) method (originally introduced by Douglas Heckathorn, see Heckathorn, 1997) with the aim to improve the representativeness of the sample. Such an approach, referred to in the literature as Web-based RDS (Bauermeister et al., 2012), seemed to address most of the challenges involved in studying Polish multiple migrants. Owing to implementation of the Web-based RDS survey, we learned about the characteristics of the mobility of Polish multiple migrants, but above all, we learned about the extent to which Web-based RDS can actually work in the field in general, and in the case of this group in particular.

Thus, the goal of this chapter is to address the challenges and opportunities of using the Web-based RDS method for migration studies and beyond. Although in our particular case, we used the method to analyze international mobility, we believe that it can be of value more generally in a variety of thematic contexts.

The chapter proceeds as follows: first, it introduces the RDS survey method by pointing to the differences between traditional and Web-based approaches and by outlining the strengths of the Web-based RDS approach for the migration research. Next, we describe the research design of the MULTIMIG project and provide an overview of the survey sample we obtained. Then, we turn to our assessment of the effectiveness of the Web-based RDS method in our study of multiple migrants. Finally, we outline the challenges to overcome and the outlook for future uses of Web-based RDS in studies on migration and beyond.

4.2 Respondent Driven Sampling in Migration Studies: Face-to-Face and Web-Based Approaches

4.2.1 RDS Assumptions

The RDS method is a modified version of chain-referral sampling with a double incentive system. Respondents are remunerated for taking part in the research (primary incentive) and for recruiting a peer (secondary incentive). This system facilitates the recruitment process and reduces some of the biases that have been found to occur with "traditional" snowball sampling (for an elaboration, see Heckathorn, 1997). RDS starts with a selection of a small number of individuals (so-called *seeds*) who initiate the recruitment chains. These initial seeds should be diversified with regard to the characteristics that influence how social ties are formed (including, age, social standing, and geographical location). They also should possess wide networks and be highly motivated to take part in the research (Wejnert & Heckathorn, 2008). The seeds and subsequent recruiters are allowed to recruit only a limited number of persons each (usually two to three) by physically passing them coupons with a unique code (Tyldum & Johnston, 2014: 13). Therefore, the recruitment within chains proceeds without the intervention of researchers, and is

based on the social networks of respondents. Consequently, face-to-face RDS surveys have limited geographical coverage. Respondents, while physically passing coupons to persons they invite to the study, have a tendency to approach the people living in their neighborhood (McCreesh et al., 2011). At the same time, RDS is not only a recruitment method, but also a broader analytical approach based on a number of rigorous assumptions that allow for the procurement of unbiased estimators for the studied populations (Heckathorn, 1997, 2007; Volz & Heckathorn, 2008).

The RDS conditions and assumptions relate to the network structure of the studied population (linked also to the definition of the studied group) and the sampling procedure (Gille et al., 2015). First, the population has to be densely networked, and the relations between its members should be reciprocal. The definition of the *studied group* should be formulated with respect to two interlinked conditions—the members of the studied group should be capable of identifying other members of this group, and they should share some feeling of belonging to this group. Both assumptions are important prerequisites for effective RDS recruitment. Moreover, respondents must be able to report the number of peers in their network who belong to the target population. Researchers use reported information to assess the bias of the obtained RDS data. In the studied group, barriers for contacts and bottlenecks in the network may exist, but the network has to constitute a single component.[1] *Homophily*—understood as a tendency of an individual to hold ties with others who are like them—within the network should not be too strong, so to avoid the tendency that individuals will only recruit people who are similar to them. Another precondition relates to the sampling procedure that individuals must randomly recruit peers from their networks. These two conditions ensure that a fundamental precondition of RDS is met: that the characteristics of the final sample are independent of the seeds' selection. However, this precondition can be guaranteed only if the chains in the sample are long enough (for a more elaborate description of the RDS assumptions, see the introduction in Tyldum and Johnston (2014), Heckathorn (1997) and Volz and Heckathorn (2008)).

The RDS method was designed to study populations whose members usually do not want to openly admit their affiliation to the studied group, and thus remain "invisible" to researchers (Heckathorn, 1997). For most such groups, respective registries are incomplete or inexistent, for example, injection drug users, commercial sex workers, and men having sex with men (Malekinejad et al., 2008; Montealegre et al., 2013). The RDS method also has been increasingly used to study migrant populations (Tyldum & Johnston, 2014; Schenker et al., 2014) and has proved to be effective in many contexts, including research on Polish migrants (Tyldum & Johnston, 2014). A comparison of two surveys based on quota sampling and RDS carried out among Ukrainian migrants in the larger Warsaw area demonstrated that the RDS survey provided a more diversified sample with a higher representation of temporary migrants and a broader geographical coverage (Górny & Napierała,

[1] This condition enables the passing of the coupon (via connections in the network) from one person to any other randomly chosen person in the network, independently of the seeds' selection.

2016). However, it also has been found that the RDS method may lead to an underrepresentation of highly skilled, well-off migrants and those with limited contacts with other migrants (e.g., spouses of natives) (Górny, 2017).

4.2.2 Web-Based RDS vs Face-to-Face RDS

In many respects, the Web-based RDS procedure is similar to the face-to-face RDS version. However, the recruitment process is based on the virtual ties of respondents, which only partly intersect with their personal, physical social networks. On the one hand, the Web-based approach may involve more dense networks of weaker, virtual ties, and on the other, it may limit the possibilities of recruiting those potential participants who are less embedded in the digitalized world. To begin a Web-based RDS study, researchers identify a limited number of seeds in the target population, to whom they send an individualized link to the online questionnaire. While face-to-face RDS makes use of paper and pencil or computer assisted personal interviews, the Web-based RDS interviewees fill in the questionnaires online. On completion of the questionnaire, the respondents receive e-coupons, which they pass on electronically to other persons in the network. Importantly, the Web-based RDS does not have an interviewer who explains the logics of further recruitment to the study. The Web-based RDS uses a dual incentive system, but, unlike the face-to-face RDS with a "cash in hand" transfer, Web-based RDS respondents can receive remuneration only if they provide their personal details, which may prove to be an issue for some respondents. This problem can be reduced when the reward for participating in the study is transferred as a donation for a charity organization. However, the money donation solution might be unattractive for some respondents. When selecting incentives, researchers need to make sure that the form of the reward is suitable for an online transfer (for examples of such rewards, see Bauermeister et al., 2012; Bengtsson et al., 2012; Lachowsky et al., 2016; Strömdahl et al., 2015).

A Web-based version of the RDS builds on the strengths of online surveys and the RDS method. Potentially the Web-based RDS enables the reaching of large populations in a relatively short time, since the recruitment does not require a physical passing of the coupon (instead, it is passed via the Internet) and interview completion occurs without the need to physically go to a specific location (Tyldum & Johnston, 2014; Wejnert & Heckathorn, 2008). Thus, the Web-based RDS does not pose geographical limits on a study, unlike traditional RDS studies (Bengtsson et al., 2012). Perhaps even more importantly with respect to migration research, the selection of research sites is made by the respondents themselves who choose further study participants from their networks, participants who reside in various places (Salamońska & Czeranowska, 2018). Therefore, the Web-based RDS reduces the financial resources required to carry out a study in terms of interview venue rent, interviewers' remuneration, and questionnaire printing costs (all relevant for the face-to-face version of the RDS). The online interview mode also may be natural to

migrants who navigate the online world in their daily lives to stay in touch with their family and friends who live in other countries.

Among the weaknesses of the Web-based RDS, as opposed to the face-to-face RDS, the former assumes a certain level of digital competences among the target population and Internet access, which may introduce a possible underrepresentation bias (Wejnert & Heckathorn, 2008) with respect to groups such as poor or older migrants. Another weak point of the Web-based RDS is generally understood as the limited control researchers have over the research process. Since the Web-based RDS recruitment process progresses quickly online, they may not be able to react in a timely manner if oversampling of specific sub-groups occurs. As in face-to-face RDS surveys, such potential biases are difficult to identify if they do not take an extreme form (e.g., only men being recruited to a sample). This situation may be explained by the fact that the RDS method usually is employed when official statistics and adequate sample frames are unavailable. Importantly, Web-based RDS researchers have virtually no control over who participates in an interview (Wejnert & Heckathorn, 2008) or over the quality of respondent answers. Moreover, an interviewer is not available to explain the logic of further recruitments to the study or the transfer of e-coupons. At the same time, the risks of duplicated responses and of persons from outside the target population filling in the questionnaires (overusing the survey to reap the prizes) is higher with the Web-based RDS than with other online surveys, since respondents can earn more money by filling out more Web-based RDS questionnaires. Also, in the case of targeting a worldwide population, a research team may find it challenging to design an incentive system that could operate equally efficiently in various countries, a situation in which the same reward will have a different purchasing power, depending on the country (Salamońska & Czeranowska, 2018).

Nevertheless, researchers can attempt to track these misuses of Web-based RDS. For example, researchers can restrict participants to only one questionnaire from any one IP, although this strategy would not block people from repeatedly participating in questionnaires by using various electronic devices. Adding internal checks in a questionnaire, with an aim to detect inconsistencies, would be an option for identifying respondents who do not belong to the studied group.

The Web-based RDS was developed and tested across various studies on "hidden populations," but, to our knowledge to date, Web-based RDS studies on migrants have not been done. Observations regarding the efficacy of sampling with the Web-based RDS method are mixed and depend on the character of the target population. It can be argued that in the case of groups that have a comparatively strong affiliation to a given community, the Web-based RDS method has been relatively effective, for example, in studies of men who have sex with men living in Vietnam (Bengtsson et al., 2012) or marijuana users in Oregon (Crawford, 2014), and American youth (Bauermeister et al., 2012). However, in other Web-based RDS surveys, researchers have frequently struggled with a low propensity of respondents to recruit their peers and with the problem of short referral chains (e.g., Lachowsky et al., 2016; Strömdahl et al., 2015; Truong et al., 2013). It is clear that methodological research is still needed regarding this domain.

4.3 Web-Based RDS Survey of Polish Multiple Migrants: Research Design and Overview of the Sample

4.3.1 Survey Design and Implementation

A Web-based RDS survey on Polish multiple migrants was carried out in 2018 by the Centre of Migration Research, University of Warsaw in a project entitled 'In search of a theory of multiple migration. A quantitative and qualitative study of Polish migrants after 1989' (MULTIMIG).[2] The project was designed to study the migration trajectories of Polish multiple migrants via a mixed methods design, including a Web-based RDS survey and a qualitative panel study.

The Web-based RDS *target population* was defined as adults who were born in Poland and residing abroad at the time of the survey and who had lived for at least 3 months in each of at least two countries other than Poland. No limits were imposed on the country of residence, i.e., the research covered Polish multiple migrants worldwide. The research team also decided to make this definition inclusive of Poles who migrated at different stages of their life courses (also including those who migrated as children) and in different historical periods. Return migrants were not part of the target population.

The data on multiple migrants are scarce, but, for example, according to results of a survey of Polish migrants by the National Bank of Poland in 2016, around 11% of all Polish migrants in four European countries were multiple migrants (with percentages varying by country; see Jancewicz & Salamońska, 2020). Thus, multiple migrants constitute a small fraction of the overall population of Polish emigrants. Multiple migrants also are a hidden and presumably highly geographically dispersed population, which poses additional challenges. They may be quite mobile and yet maintain connections to Poland and other international context(s) in which they have lived. Needless to say, the project design lacked a sampling frame for this group. In other words, for some countries, we had some sample frames that we could use for Polish migrants, but none were available for Polish multiple migrants specifically.

Designing the survey as a Web-based RDS seemed to respond well to the aforementioned challenges related to the sampling of this particular target population. However, this assessment would hold only if the two RDS assumptions were met with respect to the existence of the specific virtual network of Polish multiple migrants and the sense of belonging to the group (as required by RDS methodology). Overall, the implementation of the Web-based RDS was supposed to be its methodological test in this research field.

An external market research company specializing in online surveys implemented the Web-based RDS. Our research team had access to a platform that enabled the tracking of the recruitment chains in real time so as to control the recruitment process. We designed the questionnaire in Polish, so knowledge of the Polish

[2]This project was funded by the National Science Centre, Poland, under a Sonata Bis Grant, 2016–2021 (ID: 2015/18/E/HS4/00497).

language was an additional target group screening criterion, although not explicitly. The questionnaire consisted of the following sections: migrant trajectories, professional trajectories, life course events, social relations, and identity. It also featured RDS-related questions that aimed to assess the size of respondents' networks of other multiple migrants. Overall, the research topic—tracing the migratory trajectories of multiple migrants and their correlates—required detailed and diversified information from respondents. We designed the questionnaire to take up to 30 min to complete, based on a successful example of an online survey of Latvian emigrants (see Mierina, 2019). We assumed that this length of questionnaire would not increase the number of survey dropouts because of the engaging topic. Additionally, we relied on the expertise of the market research company implementing the survey, which considered the length of our questionnaire as acceptable for an online survey.

We aimed to reach 500 respondents with our survey. At the beginning of the study, our research team identified a small number of potential seeds (four recruited in July 2018) based on the personal networks of researchers and social and professional networking sites. We sent a personalized questionnaire link to these initial seeds. One of the last questionnaire screens included a request for help to recruit three additional respondents to take part in the study. Once the respondents agreed to recruit others (by marking the adequate answer in the questionnaire), we provided them with specific links to the questionnaires and instructed them that they could send these links to up to three other adult Polish multiple migrants living outside of Poland. The provided information included explanations that links were personalized so that they could be used only once and by one person each. The screen with invitation links included information about the reward for recruiting further respondents to the study. It also reminded the participants of the definition of the *target population*. In hindsight, these explanations did not adequately stress the outstanding importance of the referral system to the study's success.

For completing an interview, each respondent received PLN40 (equaling to around EUR10). With respect to each successful recruitment of another respondent (a completed interview), a respondent would earn an additional PLN20. So each respondent could earn a maximum of PLN100 (around EUR25). Respondents also could choose PayPal transfer as an incentive form (although this required them to submit personal data when they completed an interview), a charity donation (a reward transfer to one of three charities working with persons with disabilities, animals or older persons), or no prize at all.

We started the fieldwork on July 15th 2018, and it proceeded until December 13th 2018, a period of 5 months in total, which was much longer than expected. A pilot survey that preceded the fieldwork was carried out between May 29th 2018 and June 14th 2018. It involved recruitment of two seeds who could recruit as many respondents as possible for the study. This pilot study was set up to test the questionnaire and also the Web-based RDS recruitment component. Only one of the pilot seeds recruited one respondent for the study, which resulted in three interviews collected at the pilot stage. As a consequence of this small pilot study, we amended the questionnaire, and carefully checked the procedure of passing links for technical problems.

4.3.2 Sample Overview

We closed the online survey after 515 respondents had replied. During the data cleaning process, we discovered that 35 migrants—although having declared at least two migration experiences in two or more countries at the screening stage—pointed to repeated migration experiences in one country only in the migration trajectories section of the questionnaire. These migrants were excluded from the overview of the sample presented below (i.e., the final sample of Polish multiple migrants was 480 respondents).

The Polish multiple migrants who participated in the Web-based RDS survey mostly had experiences of living in two different countries outside of Poland (55%), whereas a further 28% had lived in three countries and 10% had lived in four countries. Only 6% of our respondents' trajectories involved living in five or more countries. The countries in which the multiple migrants lived at the time of the interview were predominantly European Union member states. The UK, Germany, Czechia, Spain, Australia, Denmark, Norway, Switzerland, the Netherlands, and the US were the top 10 countries in which the multiple migrants resided (in descending order). The top countries, the UK and Germany, hosted about 9% of the sample each. In the case of Switzerland, the Netherlands, and the US, the sample percentage was around 3% each. As regards the motivations for leaving Poland, the respondents indicated work (40%), family reasons (35%), and simply a wish to live in a new country (35%). Also, some declared mixed motivations for moving.

The vast majority of multiple migrants who participated in the survey were women (75%). The respondents also were relatively young, with a mean age of 34, and the majority was between 25 and 34 years old. The respondents also were highly educated, with over 75% completing a third level education (obtained in Poland or abroad). Most (80%) were in a relationship (either registered or not), and about half had children. About 70% worked in the destination country (this proportion was even higher for men), 15% were caretakers for home and family (these percentages were higher for women), and about 7% were students. Overall, the study enabled reaching quite a heterogeneous sample, with the majority holding high levels of human capital.

4.4 Effectiveness of Recruitment in the Web-Based RDS

4.4.1 Overview of the Recruitment Process

We assessed the effectiveness of the recruitment for this Web-based RDS survey on Polish multiple migrants on the basis of the whole obtained sample, i.e., 515 questionnaires,[3] which enabled the tracking of the dynamics of the data collection. This

[3] Since for 13 respondents the recruitment tracking was missing due to technical issues, we excluded these cases from our analysis of the recruitment process.

number refers to the completed interviews only. We did not include incomplete interviews in the database because they, by and large, involved situations in which the respondents did not pass a screener—they did not satisfy the definition of the target group. Overall, while the planned number of questionnaires (500) was achieved, the number of respondents who qualified as seeds (regarding the RDS methodology) amounted to 395 persons, i.e., the vast majority (79%) of the sample. One third of all respondents refused outright to recruit anybody, either for a lack of adequate persons in their network or due to other reasons.[4]

Only 3% of the sample was recruited during the fourth or later wave (at least four persons in the chain recruited somebody). All these respondents belonged to one "exceptional" recruitment chain encompassing 22 persons. This chain started from a young man (25 years old) with secondary education residing in Germany who had migration experiences also beyond Europe. Interestingly, this chain consisted of relatively diverse individuals in terms of, for example, their countries of residence (10 different countries, almost all European) and age (persons born between 1957 and 1995). These migrants often undertook their first migration for study (40.9% vs. 33.8% in the total sample). However, at the time of the survey, they were performing jobs that covered almost the whole occupational ladder in the destination countries. What they did have in common was that, except for one person, they selected a PayPal transfer as the remuneration for an interview. Nevertheless, it would be difficult to identify one clear distinctive feature of the persons who belonged to this single "super-chain," and the group also was too small for carrying out a meaningful statistical analysis. However, it is worth noting that these 22 interviews were collected over 1 month (from November 5th until December 8th), so at a relatively slow pace.

Overall, one sixth of the obtained overall sample was recruited within the first wave (i.e., by seeds invited to participate in the survey by researchers), 3% in the second wave, and only 1% in the third wave (see Table 4.1). Consequently, for the majority of seeds invited to the study by researchers (82% or 325 seeds),[5] no recruitment chains were formed, and the chains we did obtain were very short. Thus, this recruitment was not satisfactory at all, given the precondition of the RDS method that to obtain a frequency equilibrium of a sample, recruitment chains should be "long enough" (Heckathorn, 1997). So, even if we treat this precondition in a flexible manner, the recruitment process still did not meet the precondition of the RDS method, since the majority of the sample was recruited by researchers.

It became clear during the fieldwork that the recruitment process based on the RDS method was not effective enough to reach a sample of an adequate size. During the first 3 months of fieldwork, the number of collected questionnaires barely

[4] Before receiving the invitation links, respondents were asked if they would agree to recruit somebody to the study.

[5] This number might be lower due to problems with the transfers of links in some invitations (up to 50 cases), so the recruitment process might not have been registered appropriately. We discuss this problem in a later part of this section.

Table 4.1 Recruitment of respondents by wave

Wave	Number of respondents	Share in the total
Initial seeds	395	79%
1st wave	70	14%
2nd wave	14	3%
3rd wave	6	1%
4th or later wave	17	3%
Total	502	100%

Notes. For 13 cases, information on the number of the wave was missing
Source: MULTIMIG Web-based RDS survey 2018

exceeded 10 per month. This was a period when only a limited number of seeds (10 seeds) was recruited, as demanded by the RDS methodology, which were expected to induce long recruitment chains. When at the end of the third month (October 13th) only 32 questionnaires had been collected, the research team changed the strategy by introducing a more intensive recruitment of respondents-seeds. At this stage, the research team still assumed that inducing a more intensive recruitment of respondents was possible and that some of those numerous respondents-seeds would initiate recruitment chains. However, we need to stress that the decision about increasing the recruitment of respondents was an initial step towards relaxing the rigid methodological assumptions of the RDS method. We placed the invitations to participate in the study on more than 120 Facebook groups. Persons interested in participating had to contact a researcher to acquire a personalized link to the questionnaire. This move visibly enhanced the collection of interviews. During the following month of fieldwork, we collected almost 90 questionnaires, i.e., 2.8 questionnaires per day. However, the recruitment driven by the survey participants still was not satisfactory, since the vast majority of the seeds did not recruit anyone.

Therefore, we introduced another modification to the fieldwork—we substantially shortened the questionnaire. The research team saw this change as a methodological test that took into consideration the recommendation of the survey-methods literature to use shorter questionnaires in a web-mode survey (e.g., Hoerger, 2013). We reduced the number of questions from 139 to 91. These cuts included the deletion of questions about exact dates (month and year) of migration moves. According to comments made by respondents, these questions were posing particular difficulties, since they demanded a detailed recollection of mobility trajectories.

We introduced the shorter version of the questionnaire on November 22nd and boosted the collection of data. Until the end of the fieldwork on December 13th, the average speed of data collection was 14.8 questionnaires per day. Among the 311 questionnaires collected during the last 3 weeks of fieldwork, 53 were the longer versions of the questionnaire; they were still circulating on the Internet because the research team decided not to interrupt the potential recruitment process. Overall, the total number of questionnaires in the longer and shorter versions were 257 and

Table 4.2 Methodological phases of the fieldwork

Phase description	Time	Number of completed questionnaires	Mean daily number of completed questionnaires
Limited recruitment of seeds, long questionnaire	15.07–12.10.2018	32	0.4
Intensive recruitment of seeds, long questionnaire	13.10–13.12.2018	225	3.7
Intensive recruitment of seeds, short questionnaire	22.11–13.12.2018	258	12.3
Fieldwork—total	**15.07–13.12.2018**	**515**	**3.4**

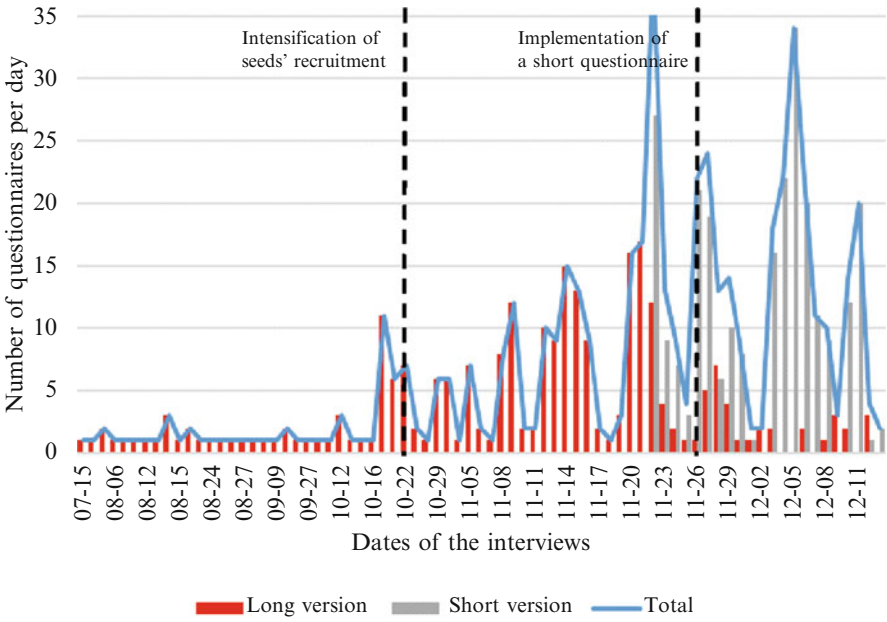
Source: MULTIMIG Web-based RDS survey 2018

258, respectively. Table 4.2 provides a more detailed description of the three methodological phases of the project.[6]

It is challenging to assess the influence of the shortening of the questionnaire on the speed of data collection. First, respondents could fill in the questionnaire in stages and return to it as many times as they wanted. Consequently, the measurement of the time devoted to the completion of the questionnaire was problematic (the total time between the start and end of filling in the questionnaire). However, the differences between the median duration of time to answer the two versions of the questionnaires (long and short) were visible: 37 min for the longer version and 23 min for the shorter one (the latter was closer to the recommended length in the literature for an online questionnaire). Second, it is difficult to disentangle the effect of the intensity of the recruitment on Facebook groups and other channels carried out by researchers and the external market research company. As portrayed on Fig. 4.1, this active recruitment was crucial in shaping the pace of data collection. Peaks in the numbers of filled-in questionnaires correlated first of all with the timing of the invitation placements on the Internet (Czeranowska, 2019). Over 60% of respondents declared that they found information about the study on the Internet (55% referred directly to Facebook). The maximum daily number of completed long questionnaires was 17, whereas for short questionnaires, it was 34, which suggests that the shortening of the questionnaires had an impact on the readiness of target group members to participate in the study.

Reducing the length of the questionnaire did not impact the effectiveness of the follow-up recruitment by respondents (the key precondition of the RDS method). The overwhelming majority of the respondents who filled in the shorter version of the questionnaire were seeds (84%), and none were recruited during the fourth or later waves, whereas those recruited within the second or third waves constituted only 2% of the analyzed subsample. Moreover, among the one third of respondents who refused to recruit anybody (did not mark an adequate answer in the

[6]This third methodological phase of the project was accompanied by an additional mailing to over 300 Polish organisations worldwide. However, this mailing was not very effective due to a number of inactive addresses and the small response rate from these organisations (IQS, 2018).



Source: MULTIMIG Web-based RDS survey 2018

Fig. 4.1 Numbers of filled-in questionnaires by the date of an interview and the version of the questionnaire (long or short)

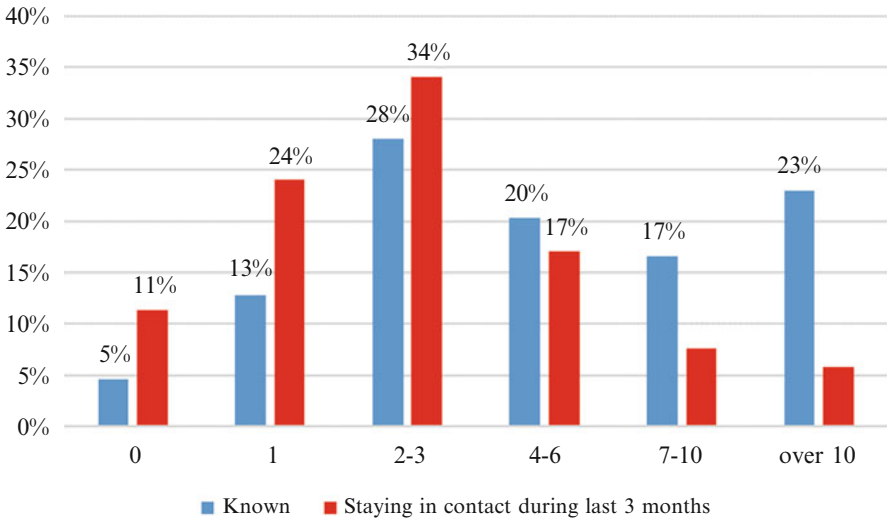
questionnaire; see Sect. 3.1), the short and long versions of the questionnaires were evenly distributed (50:50). Thus, the factors that restrained the recruitment by respondents did not relate strongly to the length of the questionnaire. Overall, in our study of multiple migrants, the intensification of the seeds' recruitment and the shortening of the questionnaire produced an increase in data collection speed, although these changes did not stimulate the recruitment by respondents.

4.4.2 Potential Barriers for the Effective Recruitment

The effectiveness of follow-up recruitment is conditioned by two main factors—the capability to recruit new respondents and the motivation to do so. The former is directly linked to the character of the definition of the *target group*, which should be clear to the respondents when recruiting others. Furthermore, the criteria should be clear enough to make it simple for them to identify, in their social networks, persons who can be invited to participate in the study. It can be argued that the definition of the *multiple migrant* employed in the study is rather straightforward: Poles who spent at least 3 months in at least two destination countries. However, determining whether respondents were always aware of all their friends' migrations and

migration durations is less obvious. Possibly, respondents had to check with their acquaintances about the details of their mobility before inviting them to the study. Additionally, instead of checking these details with their friends, respondents may have chosen not to invite anybody, which could have created a barrier to the speed and effectiveness of the recruitment.

Consequently, respondents could have either under- or over-estimated the number of their friends who could be invited to participate in the study. At the same time, a sufficiently high number of other members of the target group within the social networks of respondents was an essential precondition for RDS recruitment to be effective (Gille et al., 2015). This requirement was not fully met in the discussed study. Although only 5% of studied migrants did not know any multiple migrants, 11% had not been in contact (as declared by the respondents in the questionnaire) with such persons during 3 months preceding the survey (Fig. 4.2). This fraction of the sample had almost no capability to recruit new respondents. Over half of the sample was in touch with up to three multiple migrants, but for one fourth, it was only one multiple migrant. The respondents who knew more than 10 multiple migrants constituted another fourth of the sample, but only 8% had been in touch with that many members of the target population during the last 3 months preceding the survey. Therefore, most respondents did not have large Polish multiple migrant networks. Moreover, their ties within these networks were rather loose and weak,



Notes. [1]Defined as persons who spent at least three months in at least two countries outside Poland.
Source: MULTIMIG Web-based RDS survey 2018

Fig. 4.2 Number of Polish multiple migrants in respondents' networks according to two indicators: persons known by respondents and persons with whom the respondents were in touch within 3 months prior to the study (%)

since they involved only occasional contacts, which may have been a crucial barrier for recruitment using the Web-based RDS survey approach.

The recruitment capabilities of respondents also relate to the means for distributing invitations, which explains the limited territorial coverage of face-to-face RDS surveys (McCreesh et al., 2011). Although territorial boundaries do not exist on the Internet, Web-based recruitment requires efficient technical solutions. For example, at one point in our survey, the transfer of links (from an inviting person to an invitee) was found to be problematic. Apparently, the personalized links to the questionnaires were too long, which sometimes resulted in copy-and-paste mistakes in the invitation emails to respondents' friends (Czeranowska, 2019). Moreover, the correspondence of the research team with invited seeds indicated that it was not always clear to seed-respondents that they should copy the links from their questionnaires. Some of the seeds thought that they needed to provide their friends' personal data in the questionnaires, which made them uncomfortable. This anxiety may have caused them to refuse to recruit anybody. Others, especially those filling in questionnaires on mobile phones, lost their links after switching to their email screen. In case of problems with managing links, respondents were helped by the researchers who sent another link. However, this solution was only possible if they asked for help in the first place. It is difficult to assess how many did not ask for help (ibid.).

The eagerness and engagement of respondents to recruit their peers for a study is, besides their capability to do so, an important precondition for successful recruitment in RDS studies. Thus, it is crucial to create a positive atmosphere around a study (Górny & Napierała, 2016) whether it be carried out in a face-to-face or online setting. Creating such an atmosphere is more difficult for researchers doing Internet surveys, since they do not have any, or at most, have very limited personal contacts with their respondents. Such limited contact usually occurs only with those respondents who get in touch due to problems during the completion of their questionnaire.

Consequently, the research topic and the quality of research tools are the main drivers of the study's image in an online setting. Our study on multiple migrants seemed attractive to our respondents, although the fact-enumeration style of the questionnaire may have been fatiguing for some of them. In particular, recalling all their migration events was burdensome for those with particularly rich migratory histories (Czeranowska, 2019). In addition, some respondents reported as problematic the lack of the possibility to return to some earlier questions (due to technical issues) to correct answers (e.g., about migrations) (ibid.). These findings suggest that, when attempting to reconstruct migration trajectories by survey questionnaires, an option to enable respondents to correct their earlier answers is a necessity.

The afore-mentioned issues indicate that the questionnaire design can negatively influence the recruitment rate. This finding supports the general rule that Internet surveys need to be simple and engaging, which is even more important with respect to Web-based RDS studies. In this setting, overly demanding questionnaires might not only be a factor that causes them to drop out at some point, but also poses a crucial barrier to respondent recruitment. To put it more clearly, even the respondents who complete a survey might be reluctant to invite their peers to participate if they experienced the survey process as burdensome. Nevertheless, we want to stress

that shortening the questionnaire and removing potentially problematic questions in the third phase of the field period did not have a positive effect on the recruitment process by respondents. Only the speed of the seeds' recruitment increased.

Finally, the incentives used to increase the readiness of respondents to recruit their peers constitute a core element of the method, but their management is always a challenge in RDS studies, particularly as carried out with Web-based surveys. Thus, while money constitutes the most neutral type of incentive, its transfer usually violates the anonymity of respondents in some way. Most research institutes and companies are not allowed to pay money without a signed receipt. With respect to the Internet, the possibility of securing the anonymity of respondents when remunerating them is even more limited, which was the case in this study. Less than half of the respondents (49%) chose to receive a PayPal transfers (which required them to share their personal data with the research company), and 44% decided to donate their renumeration to a charity organization. In fact, respondents valued having the latter option, which is an important methodological observation from the study. At the same time, some respondents expressed their anxiety, in messages to the researchers, about sharing personal data with the research company. It is difficult to evaluate how much this anxiety discouraged some of them from inviting additional participants. Furthermore, regarding the PayPal transfers, some respondents declared (in messages sent to the researchers) that they would not invite their friends to the study until they received a PayPal transfer for the questionnaire they filled-in themselves. Such attitudes may have impeded (time lost waiting for payment) or even restrained (after some time passed, respondents may have forgotten or lost interest in the recruitment) the recruitment of new respondents.

4.5 Discussion: Challenges to Overcome and Outlook for the Future Use of the Web-Based RDS in Migration Studies and Beyond

The methodological considerations stemming from our study on Polish multiple migrants worldwide using a Web-based RDS strategy relate to the usefulness of online studies in general and the RDS method in particular. On the one hand, the study found that devoting enough time and effort to advertising a survey on the Internet can lead to the collection of a satisfactory number of questionnaires with multiple migrants. On the other hand, inducing the further recruitment of participants by respondents is a challenge when doing Web-based RDS surveys in general (e.g., Lachowsky et al., 2016; Strömdahl et al., 2015; Truong et al., 2013), and, according to our methodological observations, with multiple migrants in particular.

We found that the identified barriers to chain-referral recruitment are of two main types: those linked to RDS assumptions and those related to fieldwork design. The first category refers to the relatively small density of migrants' networks with other Polish multiple migrants. It also relates to the fact that multiple migrants apparently

may not consider themselves a distinct group, i.e., they do not identify themselves with other multiple migrants in terms of a salient social identity. However, the latter is only our supposition because we did not examine group identity in our Web-based RDS survey. This issue could be researched more in-depth in the ongoing MULTIMIG project (in which Web-based RDS was only one of the components), i.e., in a qualitative panel on multiple migrants. Nevertheless, respondents were not necessarily able to identify Poles who had experiences with multiple migration to other destination countries. It seems that Polish multiple migrants do not form a particular social network, but rather seem to be dispersed across various migrant networks (i.e., they are not necessarily directly linked with each other). In other words, although the features of the Web-based RDS method appear to address some challenges related to the sampling of multiple migrants very well, the definition of this migrant group does not fully fit the assumptions of the RDS method in that the respondents were not able to easily identify other members of the target group. An obstacle of a similar kind also has been reported in a study of the early integration patterns of recent migrants (including Poles) in several countries (Platt et al., 2015) in which the RDS recruitment was ineffective due to the limited (direct) connectedness between recent migrants and the fact that the seeds did not refer further respondents—recent migrants—from within their networks.

Barriers to RDS recruitment that stem from the design of the implemented online survey of Polish multiple migrants include a not very effective scheme of distribution of personalized questionnaire links for invitees, the format of the questionnaire (length and the fact-enumeration questions), and last but not least, a reduction of anonymity with respect to PayPal transfers. Moreover, our results suggest that the value of the implemented incentives was of little importance to the recruitment dynamics. The purchasing power of the incentive reward was different depending on in which country the respondents were based. In addition, almost half of the respondents chose the charity donation or no remuneration at all for participating in the survey, which suggests that they had non-financial motivations. Consequently, the dual incentive system, which is integral to recruitment with the RDS method, may have been inefficient in the case of our survey respondents. This tentative finding is in line with the observation that highly-skilled migrants (most of the multiple migrants in our sample) are less likely to participate in RDS surveys because the incentives are relatively unattractive to them, and thus they are not interested in participating and recruiting new participants (Górny, 2017).

Our study of multiple migrants is a useful contribution to ongoing methodological debates in migration studies and beyond. A prerequisite of the RDS method is the recruitment of new participants by respondents themselves, and this is also the main challenge we identified with respect to Web-based RDS surveys. In this regard, our most appealing and straightforward conclusions relate to the limitations of the research design, which can be reduced by intensive testing of the research tool (thus obtaining a relatively short and user-friendly questionnaire), and even more importantly, ensuring smooth operation of the e-coupons transfer. However, we would argue that a more attractive questionnaire format is not always the solution for convincing people to participate in a study, and more importantly, to invite new

persons to participate in it. Simply put, not all topics are appropriate for the Web-based RDS approach. The more engaging and salient the topic of the questionnaire is for respondents, the higher the chances of a successful implementation of a Web-based RDS survey, i.e., fast and efficient data collection resulting in long referral chains. While this suggestion is not new, with respect to Web-based RDS surveys it is crucial in order to obtain a sample that would go beyond a mere convenience sample. This, in turn, would not allow for the computation of RDS weights that would enable the obtaining of unbiased estimators for a studied population.

With regard to another element of the research design—the e-coupons form and the procedure of their transfer by respondents—further substantive testing is needed. Prospects for a satisfactory implementation of this procedure should be a pivotal criterion in the choice of the implementation mode of the survey and, if necessary, in the selection of a research company capable of conducting a Web-based RDS survey. Such an approach requires securing substantive resources for the pilot stage in the budget of the Web-based RDS survey because this exploratory stage requires the testing of different recruitment scenarios (to procure a variety of potential seeds) and of various technical solutions. Also during the pilot stage, it would be advisable to do cognitive interviews with respondents, which should focus not only on the comprehension and wording of questions, but also on how an e-coupon transfer occurs in the context of the respondents' social networks. These interviews should address the recruitment challenges the respondents point to, the solutions to overcome them, the attractiveness of incentives, and the motivations for recruitment. Overall, the e-coupons transfer procedure should be simple and self-governed. We also strongly recommend that researchers provide instructions on how to pass on the coupons, and information on the importance of the referral process for the success of the study. An innovative measure to consider might be the usage of short videos to explain to respondents how to pass on the coupons and to stress the importance of their role as persons inviting new respondents. In face-to-face RDS, this is the interviewers' role, so, in this regard, the face-to-face RDS version is less demanding.

Another Web-based RDS component that requires attention is the form of incentives used. Both their value (not too high and not too low) and form (easily transferable online) can contribute to the success of the study. On the basis of our review of earlier studies (Bengstsson et al., 2012; Truong et al., 2013), it seems that introducing a lottery element in the incentives scheme will have a positive impact on the eagerness of respondents to participate in a study.

Regarding the RDS assumption-related challenges, the *target group* should be defined in the simplest way possible, share some common affiliation, and have dense reciprocal ties that bind its members. We would claim that these requirements are even more important in the Web-based RDS, since researchers have less influence on the research process than in face-to-face RDS surveys. This claim directly implies that the formative stage, in preparation for an actual study, should not be neglected in the online versions of RDS surveys. If possible, a pre-study should involve mixed methods and mixed mode approaches, i.e., face-to-face interviews, phone interviews, and online surveying that address the topics important to RDS survey success, such as group affiliation, social networks, readiness to pass coupons, attractiveness of incentives, etc.

To conclude, our methodological test of the Web-based RDS online survey on the population of multiple migrants resulted in several important recommendations regarding such studies' research design, in particular, with respect to the organization of the e-coupon transfer and the incentives form. It also paid attention to the fact that the Web-based RDS survey, although saving some money on fieldwork (e.g., the costs of organizing the research site, salaries of interviewers), requires substantial funding for an exploratory and testing phase. More studies of this kind are needed to cross-check and validate these recommendations in practice. Finally, the experience we gained during the study led to two important general observations. First, the implementation of established field research methods in a virtual environment provides for new opportunities, but also comes with added challenges. Second, it always is necessary to consider carefully whether the chosen study design procures a sufficiently good fit between its inherent methodological requirements and the particular characteristics of the envisioned target group.

References

Bauermeister, J. A., Zimmerman, M. A., Johns, M. M., Glowacki, P., Stoddard, S., & Volz, E. (2012). Innovative recruitment using online networks: Lessons learned from an online study of alcohol and other drug use utilizing a web-based, respondent-driven sampling (webRDS) strategy. *Journal of Studies on Alcohol and Drugs, 73*(5), 834–838.

Bengtsson, L., Lu, X., Nguyen, Q. C., Camitz, M., Le Hoang, N., Nguyen, T. A., Liljeros, F., & Thorson, A. (2012). Implementation of web-based respondent-driven sampling among men who have sex with men in Vietnam. *PLoS One, 7*(11). https://doi.org/10.1371/journal.pone.0049417

Crawford, S. S. (2014). Revisiting the outsiders: Innovative recruitment of a marijuana user network via web-based respondent driven sampling. *Social Networking, 3*, 19–31.

Czeranowska, O. (2019). *Raport z rekrutacji do badania ilościowego w ramach projektu 'W poszukiwaniu teorii migracji wielokrotnych. Ilościowe i jakościowe badanie polskich migrantów po 1989 roku'*. Centre of Migration Research, University of Warsaw, Warsaw (unpublished report).

Gille, K. J., Johnston, L., & Salganik, M. J. (2015). Diagnostics for respondent-driven sampling. *Journal of the Royal Statistical Society, 178*(1), 241–269.

Górny, A. (2017). All circular but different: Variation in patterns of Ukraine-to-Poland migration. *Population, Space and Place, 23*(8), 1–10.

Górny, A., & Napierała, J. (2016). Comparing the effectiveness of respondent-driven sampling and quota sampling in migration research. *International Journal of Social Research Methodology, 19*(6), 645–661.

Heckathorn, D. D. (1997). Respondent-driven sampling: A new approach to the study of hidden populations. *Social Problems, 44*(2), 174–199.

Heckathorn, D. D. (2007). Extensions of respondent-driven sampling: Analyzing continuous variables and controlling for differential recruitment. *Sociological Methodology, 37*(1), 151–207.

Hoerger, M. (2013). Participant dropout as a function of survey length in internet-mediated university studies: Implications for study design and voluntary participation in psychological research. *Cyberpsychology, Behaviour, and Social Networks, 13*(6). https://doi.org/10.1089/cyber.2009.0445

Jancewicz, B., & Salamońska, J. (2020). Migracje wielokrotne w Europie: polscy migranci w Wielkiej Brytanii, Holandii, Irlandii i Niemczech. *Studia Migracyjne—Przegląd Polonijny, 2*(176), 7–28. https://doi.org/10.4467/25444972SMPP.20.009.12325

Lachowsky, N. J., Lal, A., Forrest, J. I., Card, K. G., Cui, Z., Sereda, P., Rich, A., Raymond, H. F., Roth, E. A., Moore, D. M., & Hogg, R. S. (2016). Including online-recruited seeds: A respondent-driven sample of men who have sex with men. *Journal of Medical Internet Research, 18*(3). https://doi.org/10.2196/jmir.5258

Malekinejad, M., Johnston, L. G., Kendall, C., Kerr, L. R. F. S., Rifkin, M. R., & Rutherford, G. W. (2008). Using respondent-driven sampling methodology for HIV biological and behavioral surveillance in international settings: A systematic review. *AIDS and Behavior, 12*(Suppl), 105–130.

McCreesh, N., Johnston, L. G., Copas, A., Sonnenberg, P., Seeley, J., Hayes, R. J., Frost, S. D. W., & White, R. G. (2011). Evaluation of the role of location and distance in recruitment in respondent-driven sampling. *International Journal of Health Geographics, 10*, 10–56.

Mierina, I. (2019). An integrated approach to surveying emigrants worldwide. In K. R. Mierina (Ed.), *The emigrant communities of Latvia. National identity, transnational belonging, and diaspora politics* (pp. 13–34). Springer.

Montealegre, J. R., Johnston, L. G., Murrill, C., & Monterroso, E. (2013). Respondent driven sampling for HIV biological and behavioral surveillance in Latin America and the Caribbean. *AIDS and Behaviour, 17*, 2313–2340.

Platt, L., Luthra, R., & Frere-Smith, T. (2015). Adapting chain referral methods to sample new migrants: Possibilities and limitations. *Demographic Research, 33*(24), 665–700.

Salamońska, J., & Czeranowska, O. (2018). *How to research multiple migrants? Introducing web-based respondent-driven sampling survey* (CMR Working Papers 110/168). Warsaw University, Warsaw.

Schenker, M. B., Castañeda, X., & Rodriguez-Lainz, A. (Eds.). (2014). *Migration and health. A research methods handbook*. University of California Press.

Strömdahl, S., Lu, X., Bengtsson, L., Liljeros, F., & Thorson, A. (2015). Implementation of web-based respondent-driven sampling among men who have sex with men in Sweden. *PLoS One, 10*(10). https://doi.org/10.1371/journal.pone.0138599

Tourangeau, R., Edwards, B., & Johnson, T. P. (Eds.). (2014). *Hard-to-survey populations*. Cambridge University Press.

Truong, H. H. M., Grasso, M., Chen, Y. H., Kellogg, T. A., Robertson, T., Curotto, A., Steward, W. T., & McFarland, W. (2013). Balancing theory and practice in respondent-driven sampling: A case study of innovations developed to overcome recruitment challenges. *PLoS One, 8*(8). https://doi.org/10.1371/journal.pone.0070344

Tyldum, G., & Johnston, L. (Eds.). (2014). *Applying respondent driven sampling to migrant populations. Lessons from the field*. Palgrave Macmillan.

Volz, E., & Heckathorn, D. D. (2008). Probability based estimation theory for respondent driven sampling. *Journal of Official Statistics, 24*(1), 79–97.

Wejnert, C., & Heckathorn, D. D. (2008). Web-based network sampling: Efficiency and efficacy of respondent-driven sampling for online research. *Sociological Methods & Research, 37*(1), 105–134.

Chapter 5
Computer-Assisted Migration Research: What We Can Learn About Source Questionnaire Design and Translation from the Software Localization Field

Dorothée Behr

5.1 Introduction

In the digital age, surveys are often conducted computer-assisted, which can either be administered by an interviewer or self-administered in the presence of an interviewer, or online without any interviewer present. In the following, I will use the term "computerized surveys" for all these scenarios. Among the great advantages of computerized surveys is the fact that the collected data is immediately digitized and does not require further processing. On the design side, the computerization comes with an increased interplay and interconnectivity between questionnaire and survey software (e.g., through the use of fills or placeholders, complex routing, automatic pop-up error messages, etc.), which has to be carefully considered when implementing computerized surveys.

The trend towards computerization also applies to migration research. While computerized surveys facilitate certain processes, such as reaching the target population (Pötzschke & Braun, 2017), the typically multilingual and multicultural character of migration surveys adds a layer of complexity to computerized implementation. This chapter aims to tackle the topic of questionnaire translation in migration research mainly from this technical perspective of multilingual survey implementation. General translation approaches and requirements can be found in various publications on questionnaire translation (e.g., Behr, 2018a; Behr & Shishido, 2016; Harkness et al., 2010b; Mohler et al., 2016). However, technical aspects related to translation often remain in the background, even though they can become quite crucial for data quality (Wang et al., 2017). It is only recently that challenges resulting from the interplay between translation and survey software have been brought to the fore (Pan et al., 2020; Wang et al., 2017). Although highly

D. Behr (✉)
GESIS – Leibniz Institute for the Social Sciences, Mannheim, Germany
e-mail: dorothee.behr@gesis.org

S. Pötzschke, S. Rinken (eds.), *Migration Research in a Digitized World*, IMISCOE Research Series, https://doi.org/10.1007/978-3-031-01319-5_5

informative, these recent works do not attempt to build on and connect to related research in other disciplines, such as in translation studies. A notable exception is Upsing et al. (2011), who describe processes for large-scale competency assessment with reference to standards in software translation and adaptation – typically called "software localization" – using PIAAC as an example, the OECD Program for the International Assessment of Adult Competencies. Localization hereby refers to the processes of "adapting [digital] content linguistically, culturally, and technically" (Gambier, 2016, p. 891). It is the explicit aim of this chapter to draw on frameworks from software localization to foster knowledge transfer. I should add that the content of this chapter is applicable to multilingual computerized survey research in general, whether it is migration, cross-cultural or cross-national research.

In the following, I will provide a brief overview of good practices in questionnaire translation in general (Sect. 5.2). I will then introduce frameworks from software localization and transfer these to the survey research field (Sect. 5.3). With the help of these frameworks, I will outline major decisions in multilingual survey design and challenges that can arise when translating questionnaires for computerized surveys (Sect. 5.4, 5.6 and 5.7). Major players in multilingual research (e.g., Upsing et al. 2011; Dept et al., 2017; the consortia implementing the OECD studies PISA[1] and PIAAC) already operate according to the principles referred to in this chapter. Implementing computerized studies with a high level of comparability across diverse countries and cultures would otherwise not be possible. The author is not cognizant of computerized migration research that systematically addresses technical challenges right from the start. This can also be due, however, to a general scarcity of documentation on translation procedures and challenges in migration research.

5.2 Questionnaire Translation: What Aspects Lead to High Quality?

Migration research frequently requires the translation of a source questionnaire into various target languages in order to reach the intended respondent population. The term translation, as understood in this chapter, shall also include different levels of cultural adaptation (e.g., substantial modifications to items) to make an item work better in a new context (Behr & Shishido, 2016; van de Vijver & de Leung, 2011). Unfortunately, questionnaire translation is still too often seen as a step outside of the scientific process (Smith, 2004). Moreover, it is often regarded as a mere "word-by-word substitution, a problem of dictionaries" by those not familiar with translation (Gambier, 2016, p. 887). However, good translation significantly differs from such a view. It draws on interacting competencies, including linguistic and textual competencies in both the source and target language, as well as on cultural, substantive, information acquisition, tools, and general translation know-how (Behr, 2018a).

[1] PISA: Program for International Student Assessment.

When seen from this angle, it becomes clear that a simplistic view of translation leads to little consideration of translation needs, to small budgets for translation teams, and/or to unfeasible timelines. A reduced quality of the translation and ultimately of the research output is likely to result.

Much has been written in the cross-cultural survey methodology literature on good practice translation procedures. Seminal work is included in the edited volumes by Harkness et al., (2003), Harkness et al. (2010a), Johnson et al. (2018) as well as in the *Cross-Cultural Survey Guidelines* (Survey Research Center, 2016). The main guidance in cross-cultural psychology is provided by the *ITC Guidelines for Translating and Adapting Tests* (International Test Commission, 2017). The health research field is quite divers in the guidance it offers. Aquadro et al. (2008), Wild et al. (2005), Wild et al. (2009), and Eremenco et al. (2018) represent useful literature to get started in this field. Practices and challenges, specifically for migration research, are summarized in an edited volume by Behr (2018b).

The different disciplines coincide in emphasizing that a multi-step process is needed in order to ensure comparability and quality of the newly produced instrument. Using the example of the TRAPD model (Harkness, 2003), good practice calls for double translation of the questionnaire by two independently working translators (T), team reviews to arrive at a final version (R = Review and A = adjudication), pre-testing (P) among the target population, and a thorough documentation of all these steps for both internal and external quality control (D).[2] The various steps should include input from experts from different fields since a combination of expertise (in translation, questionnaire/survey design, and the substantive topic) is deemed crucial for producing a high-quality translation that fulfills both the needs of a good translation and those of a properly functioning measurement instrument (see Behr & Shishido, 2018a, b; Harkness, 2003).

Furthermore, translation teams need to be briefed on the task at hand. That is, they need to be given concrete information on the study and the translation goal (e.g., are cultural adaptations allowed or should the translation adhere to the source text?) so that they can make appropriate decisions in line with the overall objective of the study. Behr et al. (2018) also speak of "input documentation" in this context, as opposed to "output documentation." The latter includes the translated questionnaires and comments on these (e.g., in case of difficult decisions or needed cultural adaptations) and a description of the overall process implemented. At a minimum, the briefing – or "input documentation" – should include information on the study, the translation goal, the target group, the survey mode, the employed translation and assessment processes, and expectations linked to each of these steps. It can – and even should – be expanded by information on key terms, on the questionnaire structure in case of complex instruments, reference materials, etc. Information can be conveyed in written form and additionally through (web) trainings (Behr, 2018a; Behr et al., 2018; Dept et al., 2017). Input documentation intended for translators and

[2] Modifications of this double translation and team model are illustrated, for instance, in Dept et al. (2017), Martinez et al. (2006), and Goerman et al. (2018).

output documentation intended for research teams, alongside with open communication channels for all types of queries and issues, is particularly important in situations where research teams responsible for a study do not themselves speak the languages of a study and thus need to rely on external translators.

Another crucial cornerstone when it comes to ensuring translation quality and comparability is the source questionnaire itself. It is already during the development phase of a source questionnaire that the way is paved for comparability. Cross-national or cross-cultural research collaboration at the development stage of a questionnaire is vital to ensure that different cultural and linguistic realities are considered and sufficiently taken on board. The wording of source questions should be kept as simple as possible and allow a "relatively" easy transfer from one language to the other. Questions can also be annotated for translation or specifically earmarked for adaptation (Behr & Scholz, 2011). Furthermore, pretesting and translatability assessments – or alternatively advance translations – help to assess the questionnaire's suitability for a multilingual and multicultural study before the source questionnaire is finalized (Acquadro et al., 2018; Dept et al., 2017; Dorer, 2020; Smith, 2004). The translatability criteria summarized in Aquadro et al. (2018) or the advance translation scheme by Dorer (2011; later updated in Dorer, 2020) highlight what can be considered when reviewing a source questionnaire for translatability (e.g., issues pertaining to culture, language or item construction).

Challenges that need to be considered when implementing a computerized survey in more than one language and/or cultural group can partly be deduced from these schemes but they are not explicitly mentioned. Since the technical set-up of a computerized survey, in particular the way how it is programmed, will impact on translation and may lead to problems with the translation later on, this topic shall receive heightened attention in this chapter. Readers should bear in mind, however, that these more technical aspects always need to be considered alongside cultural, linguistic, and design issues that can impact translation and comparability.

Against this backdrop, I now want to introduce key frameworks and approaches from software localization. The ultimate goal is to transfer these to the survey research field. The software localization field has summarized the technical challenges with multilingual software that they have encountered over decades into best practice frameworks. Transferring this knowledge across disciplines – and adapting it where needed – avoids making the same mistakes again.

5.3 Software Localization: Frameworks and Approaches

With the advent of the personal computer in the 1980s, the localization industry began to develop, tasked with "adapting [digital] content linguistically, culturally, and technically" for new regional markets (Gambier, 2016, p. 891). The industry started with the localization of software and websites, and it has now moved on to also include the localization of mobile phones and video games (Gambier, 2016).

The activities related to providing such products for new linguistic and cultural markets largely exceeded the requirements linked to translation as exercised and known before. After all, extensive project management, software or graphics engineering, content management systems, etc. are all needed in this endeavor. Hence, a new term was coined: localization. Simultaneously to these developments, computer-aided translation (CAT) tools rose to prominence, which help, for instance, in the separation of programming code on the one hand and content for translation on the other hand, and in consistent use of reoccurring text elements. CAT tools are nowadays an integral part of the translation environment of professional translators (Sin-wai, 2016).

In a more fine-grained and process-driven view, the localization industry operates according to the GILT framework, which subsumes the four interdependent activities globalization, internationalization, localization, and translation. *Globalization* stands for all activities related to marketing a product in various regional markets. *Internationalization* stands for preparing a product at the technical level for localization. As such, it is the "process of generalizing a product so that it can handle multiple languages and cultural conventions without the need for redesign" (LISA cited by Esselink, 2000, p. 2). *Localization* stands for the process of modifying a product for a specific market. *Translation* is in fact already part of localization, because localization includes both adaptation and translation. The better a product is internationalized, the more cost- and time-effective localization can be carried out (Gambier, 2016; Sandrini, 2008; Valli, 2019). Overall, the GILT framework highlights that technology and its requirements are one decisive pillar besides language and culture.

The internationalization and localization processes can further be broken down into five core elements (Schäler, 2010): (1) *Analysis* refers to a set of key questions that need to be asked prior to localization, for instance, whether it makes sense to localize the content at all, whether all the text that needs to be translated is accessible for translators, or whether it is hidden in program code that cannot be modified? (2) *Preparation* refers to preparing a so-called localization kit for everyone working on the project, including source materials, reference materials, guidelines, milestones, etc. (3) *Translation* takes place in a highly computerized environment, which is nowadays standard for many translators. Translators work with CAT tools that include translation memories[3] (TMs), terminology databases, machine translation (MT) functionalities, automatic checking routines, etc. Sometimes, preview functions allow viewing the translated text in the actual software environment, which is even more important when text strings have to be translated out of context. (4) *Engineering and testing* involves assessing the content in terms of linguistic correctness, interface layout, and functionality. (5) A *review* closes the localization project. Lessons learned are collected for future projects.

[3] Translation memories display previously translated (similar) segments and thus both speed up the translation process and ensure consistency.

Table 5.1 Merging frameworks and approaches across survey methodology and localization

Survey methodology	GILT framework	Core elements of a localization process	Covered in this chapter
Design and implementation of the source questionnaire	Internalization	Analysis	Sect. 5.4
		Preparation	Sect. 5.5
Translation, incl. Adaptation	Localization (translation)	Translation	Sect. 5.6
Technical pretesting		Engineering and testing	Sect. 5.7
Substantive pre-testing. (Please note: For reasons of completeness, substantive pretesting that examines wording, content, and comparability (e.g., through cognitive interviewing and/or pilot testing) is listed here. It should complement the process of questionnaire translation since it evaluates to what extent the questions indeed measure what they are supposed to measure. However, due to the technical focus of this chapter, substantive pretesting is not further covered. More information on different forms of pretesting can be found in: Goerman and Caspar (2010), Pan et al. (2010) or Willis (2005, 2015, 2016))	n/a	n/a	n/a

In the following, these frameworks and approaches from the localization field are merged to outline steps that should be taken when preparing for and implementing translations in multilingual and multicultural computerized migration surveys (see Table 5.1).[4]

The chapter will focus on technical issues that can affect or interact with translation. For more comprehensive guidelines on technical questionnaire design, Hansen et al. (2016) should be consulted. In the subsequent descriptions, we will assume that survey programming is implemented centrally and that a source questionnaire, as programmed, can serve as a template for the intended target languages. I will not go into details of programming but rather point to general issues to consider.

[4]Please note: For reasons of completeness, substantive pretesting that examines wording, content, and comparability (e.g., through cognitive interviewing and/or pilot testing) is listed here. It should complement the process of questionnaire translation since it evaluates to what extent the questions indeed measure what they are supposed to measure. However, due to the technical focus of this chapter, substantive pretesting is not further covered. More information on different forms of pretesting can be found in: Goerman and Caspar (2010), Pan et al. (2010) or Willis (2005, 2015, 2016).

5.4 Design and Implementation of the Source Questionnaire

This section is dedicated to the design phase of a survey questionnaire. It highlights 'internationalization' decisions that need to be made prior to and during technical questionnaire design. In line with best practice in multilingual and multicultural research, namely that the design process is decisive for translation quality later on (Behr & Zabal, 2019), this preparatory process should receive special attention in a survey.

5.4.1 Software Fit

Migration research is oftentimes multilingual. The languages chosen depend on the target population. The German IAB-BAMF-SOEP Survey of Refugees, for instance, is implemented in German and (alongside it in a bilingual fashion) in English, Arabic, Farsi, Pashto, Urdu, and Kurmanji (Jacobsen, 2018). In such multilingual surveys, the first necessary technical clarification refers to the survey software. It needs to support the required scripts and character sets, language directions (i.e., left-to-right, right-to left, vertical, bi-directional, see Hansen et al., 2016), fonts, etc. that are needed for a specific survey. This means that the decision about the required target language(s) should precede any decisions about survey software, or at least it should go hand in hand with software decisions. The right-to-left implementation needed for Arabic[5] is particularly important nowadays given that major refugee studies around the world field their surveys in Arabic, amongst other languages (e.g., Jacobsen, 2018; AIFS, 2018). Sometimes, audio-assisted self-interviewing (ACASI) is meant to compensate for the lack of multilingual interviewers – this should then also be supported by the software (see the above-mentioned studies for examples).

In the following, we will look into more specific aspects that need to be considered when designing – and translating – multilingual and multicultural computerized surveys.

[5] For ease of communication, we simply refer to "Arabic" in these lines. Readers should be aware, however, that there are many different ways to translate a questionnaire into Arabic, taking into account different regional dialects and the difference between written and colloquial Arabic. The target group should be the decisive factor when determining which type(s) of Arabic should be used for a given survey project.

5.4.2 Culture-Driven Response Formats with Technical Implications

Some survey features are culture-dependent. Hence, the survey should be designed in such a way that it allows for cultural adjustments. These adjustments could affect the following (Hansen et al., 2016; Valli, 2019, Maroto & De Bortoli, 2001; Pym, 2011):

- *Date formats:* e.g., different positions regarding day, month, and year such as mm/dd/yyyy in the US and dd/mm/yyyy in many European countries;
- *Time formats:* e.g., 12-hour vs. 24-hour clock;
- *Name formats:* e.g., two surnames in Spanish-speaking countries;
- *Address formats:* e.g., different sequence of information or type of information required (state, province, etc.);
- *Telephone number formats*: e.g., including or excluding local prefixes;
- *Number formats*: e.g., different decimal, thousand, etc. separators, such as 20,5 in German (DE) vs. 20.5 in English (US);
- *Currency formats:* e.g., currency symbol after or before the relevant currency entry;
- *Measurement units*: e.g., metric vs. imperial units for distances, Celsius vs. Fahrenheit, different clothing units, etc.

These formats are not only relevant when it comes to programming individual questions but also when information from these questions is automatically inserted in follow-up questions in a survey (*fills*). The way how fills are programmed needs to ensure cultural particularities such as addressing a person with the appropriate order of names (Wang et al., 2017) or "piping in" the date in the culturally appropriate way (see Sect. 5.4.4 for more information on fills.)

These or similar formats also play a role when defining out-of-scope answers that automatically trigger error messages popping up on the screen. For instance, the allowable ranges for feet or meter, when it comes to size, will be different across cultural groups who use metric vs. imperial units (e.g. km vs mile); or the need for inclusion or exclusion of commas or full stops will vary depending on language. Wang et al. (2017) share their experiences from the Chinese 2016 Census Test Internet Instrument: Respondents had to set up security questions. A valid answer had to contain at least three characters. In Chinese, the names of people, locations or items often consist of only two characters. Hence, respondents entering those names were confused with automatic validation checks, which were appropriate for the English source language but not for Chinese.

Moreover, if pre-coded response lists are provided, these will need to be adapted to the respective needs. For instance, pre-coded response lists of time will look different in different languages and cultures.

Valli (2019), speaking for software localization, argues that during internationalization cultural assumptions should be removed from software design. In particular, software should not include *hard-coded* culture-specific formats (e.g., date

formats) that cannot be changed. Hard-coded stands for text that is directly part of a source code and typically not accessible for translators. Transferred to the survey world, we can say that anything that could require a cultural accommodation should be soft-coded and/or be made editable in some other way.

5.4.3 Non-linguistic Adaptations with Technical Implications

Graphics, icons or images can be embedded in a survey for different reasons. They may serve a measurement purpose when they are an integral part of questions. Additionally, they could serve to represent the survey sponsor, survey agency or the study itself, for instance in the header of an online survey. For measurement-related graphics, icons or images, their cultural suitability for the target population should be assessed to ensure that cultural norms are not transgressed. For instance, Hansen et al. (2016) show, using the 2007 International Social Survey Program (ISSP), that body shapes can be presented with figures wearing only boxer shorts or bikinis in Austria whereas in the Philippines they wear clothes covering larger parts of their body. In technical terms, such images should be soft-coded so that they can be replaced if needed.

Similarly for icons representing a sponsor, agency or the study itself: For instance, when conducting multilingual web surveys within the contexts of cross-cultural web probing studies (Behr et al., 2019), the survey icon representing our institute, *GESIS – Leibniz Institute for the Social Sciences,* was uploaded and integrated into the survey with the German institute name for the study conducted in Germany and with the English institute name for all other surveys.[6] At a minimum, graphics, icons or images should contain editable text in case translation teams need to implement a change (Valli, 2019).

Also, links to external websites (e.g., on the survey introduction page linking to further information) may need to be replaced so that they directly link to a website in the respective target language. Similarly, Sha et al. (2018) describe how entry pages (websites) to a multilingual survey should best be designed and also adapted in order to ensure participation across multilingual groups in a society. In their case study, the authors were interested in limited English speakers' entry to U.S. Federal Government internet surveys.

In technical terms, all of the information referred to in this section should be accessible for translation teams so that the content can be adapted, if needed.

I should add here that also colors (e.g., background colors of a survey or of a logo) should be thoroughly checked in terms of cultural meaning and associations (Hansen et al., 2016).

[6] We started from one common source questionnaire that needed to be translated and adapted; the survey software itself (EFS) catered for multilingual implementation in the sense that once a source instrument was programmed it could easily be replicated in different languages.

5.4.4 Linguistic Differences with Technical Implications

Computerized surveys have certain features that make them unique compared to paper-and-pencil surveys. One of these features is the possibility to use responses given earlier in the survey to adapt survey text in later questions (*fills*). For instance, questions can be tailored to refer to a previously mentioned male or female partner; or questions can be asked in present or past tense depending on whether a situation currently applies or whether it applied in the past. With survey software taking on such adjustments, interviewers in interviewer-administered surveys can focus on the interviewing task itself and do not have to adapt text to a given respondent (Latour et al., 2013). In self-administered surveys, the respondents can focus on relevant text for their situation without being distracted by irrelevant information.

The multi-country *Survey of Health, Aging, and Retirement in Europe* (SHARE) (Das et al., 2005) used fills, such as the automatic insertion of 'he' or 'she' depending on the gender of a partner as indicated earlier. The multilingual implementation of the English source version proved challenging, however:

> At first sight this seemed to be straightforward, but because of country specific [sic!] grammar and syntax it became complicated. In later versions of the CAPI instrument generic fill texts used in multiple question texts were no longer used. Instead, each question had its own fills, using question-specific fill names. (p. 17)

Also in the Programme for the *International Assessment of Adult Competencies* (PIAAC), fills – or dynamic text, as they called it – was used to accommodate different respondent situations.[7] In this study, too, the researchers experienced challenges across languages. For instance, for the source question "In your ^JobLastjob, how often ^DoDid you usually . . . read directions or instructions?" it is sufficient to insert the words "current job" and "do" for the text indicated through ^ and the result is a perfect sentence in present tense in the English language. When "last job" and "did" are inserted, the resulting sentence successfully captures past respondent activities. However, in many languages, a close translation of the question, including a literal translation of fills, did not work, because past and present tense are not formed in the same manner as in the English language. Oftentimes, other solutions, including translating the entire question (or larger parts thereof) for all respondent conditions, had to be resorted to (Latour et al., 2013).

Fills are not only difficult for these linguistic-technical reasons, however. It can also be difficult to convey to translators unfamiliar with questionnaires how they are supposed to understand and translate these fills. On the other hand, if questions are completely written out for different respondent conditions, special attention needs to be directed to the briefing of translators so that they understand the difference between similarly worded sentences, their respective role in the survey, and consistency needs. For instance, in a GESIS study with tuberculosis patients from Somalia

[7] The source (background) questionnaire of PIAAC Cycle 1 can be downloaded here: http://www.oecd.org/skills/piaac/data/ (accessed 4 July 2021).

and Ethiopia, many sentences were partially replicated. One sentence, for instance, asked: 'Did the doctor voice suspicion that you may have tuberculosis?' The subsequent sentence read: 'Did one of these doctors voice suspicions that you may have tuberculosis?' The second question applied to situations where the respondent had several doctors taking care of him/her. The translations were supposed to be identical, except for the difference between 'the doctor' vs. 'one of the doctors' (and any language adjustments needed in Somali and Tigrinya because of this difference). The translators did not always translate sentences such as these consistently, which may have been due to the fact that the set-up of the questionnaire (e.g., who gets which question) was difficult to understand and general survey principles (e.g., standardization in surveys) not known. CAT tools with translation memories would have helped the translators to translate consistently in any case.

The software localization field, too, knows the challenges that come with fills, in particular if these are based on a rather simple source language, at least on the structural level. English, for instance, has a simple morphology, with word endings that do not undergo many changes from one sentence to another. This characteristic does not necessarily apply to other languages (Valli, 2019). De la Cova (2016) observes that English word order and lack of gender may not replicate well in other languages. Valli (2019) recommends that the number and nature of fills should be well considered in advance. The same applies to surveys. Moreover, those knowledgeable of translation and the different linguistic needs of the survey languages should have a say in the set-up of fills in a source questionnaire so that problems can be prevented. "Writing for translation" (De la Cova, 2016, p. 253) or even programming for translation could be the main message here. If ease of programming in the source language prevails, the resulting translations may be suboptimal, possibly even artificial, with detrimental effects on the validity and comparability of data.

Another known challenge with technical implication is that of text expansion. Compared to English, other languages are often longer (Dept et al., 2017). Microsoft (2018) states that text strings, when translated into German or Dutch, often expand by 40% (2018; see also Valli, 2019). This needs to be taken into account when designing buttons, menus or dialogue boxes in software. Transferred to the survey context, developers should ensure that buttons (e.g., 'forward' or 'backward') are sufficiently large to cater for different languages, or that pop-up windows contain all relevant text – without incorrect hyphenation or truncation. If the survey software offers default sizes for certain elements, the various testing scenarios should establish whether this is sufficient (see Sect. 5.7). The needed or required text length also plays a role when designing open-ended text boxes where the size of the text box is known to influence the response length (Dillman et al., 2014). Thus, these text boxes should fit the expected or desired response length – and they might possibly even be enlarged in general to cater to the response length in different languages (see also Meitinger et al., 2019, on response patterns for open-ended questions in different languages).

Directly related to open-ended text boxes, respondents should be able to type in characters from different scripts into open-ended text boxes. Thus, there should be no system restriction on the type of data that can be entered into these boxes. This

challenge is exacerbated when migrants are supposed to enter text into open-ended text boxes in self-administered surveys where laptops, tablets, etc. are handed out to respondents by interviewers. It needs to be ensured that the software and the keyboard support text entries in different scripts.

5.4.5 Content Differences with Technical Implications

Often, the default situation in multilingual computerized surveys is that a generic source questionnaire serves as a blueprint for all other language versions. Additional design solutions, however, should be possible to allow a cultural group to adapt content, such as adding relevant questions or response categories or changing routing based on culture-specific needs. How this can be achieved depends on the survey software and overall design decisions. However, in a purposefully designed comparative study, adaptations to the source questionnaire should all be signed off from a central organization and documented to ensure comparability. Asking about highest educational attainment based on country-specific response categories may serve as an example of an adaptation.

5.4.6 Preparing Source Questionnaires for Computer-Aided Translation

Sometimes, the output format of the survey software requires the use of dedicated translation tools (CAT tools) to read the file, but also the normal text processing formats Word or Excel are supported by CAT tools. CAT tools can only show their strength if the source text – here: the source questionnaire – is optimally prepared. This includes avoidance of manual hyphenation, of manual hard returns or of multiple blank spaces (instead of tabs); also key terminology or repetitive elements should be worded and spelled consistently. These simple style and formatting rules allow translation memories to correctly display identical or similar translations as stored in the translation memory, or term databases to reliably show pre-defined terminologies (Esselink, 2000; Valli, 2019).

5.4.7 Internationalization Testing

The localization industry calls for internationalization testing before a software product can be localized and passed on to the next step in the workflow (Esselink, 2000). This includes checking whether the software is ready for localizability. Essentially, the testing involves the issues and challenges addressed above. Key

questions to ask are: Does the software support all needed characters and scripts? Does it support different regional (date, time, etc.) formats? Does it allow for text expansion? Does it run on required operating systems? Is all text that needs to be translated or adapted accessible – likewise for icons, images or graphics?

To help internationalization testing, so-called pseudo translation (e.g., replacing text with more characters or with accented characters) can be carried out in an easy, low-cost way to identify issues in other languages, such as spacing issues, truncated text or issues with scripts. Based on this exercise and its analysis, recommendations can be made on how to proceed (Esselink, 2000; Lerum et al., 2014; Schäler, 2010).

These testing procedures from localization can equally be implemented for survey translation. Moreover, I want to stress that a source questionnaire itself should have been thoroughly tested in terms of wording, routing, and overall design, before proceeding to internationalization testing and then translation. Implementing changes on the source version – and consequently in all language versions – once the translation has started are cost-, time-, and work-intensive and there is the risk that not all source improvements are consistently implemented in all language versions.

5.5 Prior to Translation: Preparing Translation Teams

Once the source material is ready, translation can begin. In the localization industry, translators receive a so-called localization kit that does not only contain the source material to be translated but also reference materials, including translation memories, terminology databases, style guides, milestones, etc. (Esselink, 2000; Schäler, 2010; Valli, 2019). The importance of additional project information has already been discussed in Sect. 5.2, under the notion of briefing. For questionnaires, especially if the normal flow of text is interrupted through fills, or if text strings (e.g., for buttons or error messages) are not understandable without context information, annotations for translators will be helpful.

5.6 Translation, Including Adaptation

Having received the localization kit, translators in the localization industry start translating. Their task will always involve the use of specialized localization software. For survey translation, depending on output formats of source questionnaires from the survey software, translation files could be XLIFF files, which require the use of dedicated CAT tools for the translation, or Excel files, which can be translated with CAT tools but also with normal text processing programs. Translation may also take place within the survey software itself in specifically designed language editors. CAT tools can aid translation by supporting consistency of terms or of reoccurring text elements through the use of term databases and translation memories, or by

allowing the use of several automated checking routines (e.g., on spelling, punctuation, figures or formatting).

A number of translation decisions must be taken in view of the survey interface as well as with respondent activities in mind (see also Pan et al., 2020). If interviewer or respondent instructions refer to buttons on the screen (e.g., to the 'Help' button or the 'Next' button), the same translation of key terms should be used for the buttons themselves so that ease of navigation on the screen is ensured. This essentially means that the translation of the survey interface and the translation of the questionnaire should be coordinated in one way or the other.

Interviewer or respondent instructions such as 'Mark all that apply' or 'Tick only one box' should be translated with the ultimate layout and the concrete interviewer or respondent activity on the screen in mind. The translation of 'mark', 'tick' or 'box' could vary depending on these features or activities.

For questions that are asked in an open-ended fashion (e.g., the number of hours that a respondent spends on a given activity), it is important to consider the design of the survey and the position of the open-ended text box. Depending on whether it comes before the unit (here: 'hours') or after, the translation may need to be linguistically adapted to this position.

Words or phrases helping to structure a questionnaire or interview (such as: 'In the *following* . . .' or 'To what extent do you agree or disagree with the *following* statements?') always need to be translated in view of the visual survey design. The "following" could be translated in the sense of 'as follows below' if, and only if, the respondents themselves see the questions below. Otherwise, 'following' will need to be translated in a temporal sense.

If fills are used in a questionnaire, translators should be trained on how to understand these features and what they need to consider during translation. Wang et al. (2017) provide an example based on Chinese where strict adherence to the English fill structure resulted in a defective text in Chinese. Possibly, the translator (s) was not sufficiently informed on the use of fills. If fills do not work in a target language, this should be openly communicated to research teams.

5.7 Technical Pretesting

In the localization industry, testing is an integral part of software localization. It can start once the software is compiled in the target language. Testing is always based on the real application. The localization field differentiates between (a) linguistic testing, (b) interface testing,[8] and (c) functionality testing (Esselink, 2000).

The linguistic test targets all aspects related to language. Key questions that also apply to survey translation are: Has all text been translated, including error messages? Do the different scripts display correctly? Does the text hyphenate correctly?

[8] This is called 'cosmetic' testing in Esselink (2000).

Do fills display correctly (e.g., do they appear in the correct position in the sentence, or is capitalization or a small letter of a fill appropriate at the given position in the sentence)? Is all text translated in the intended sense, including interface elements, button labels, etc. (Pan et al., 2020)?

The interface test focuses on visual aspects. Questions that should be addressed here are: Is the text in dialog boxes or error messages displayed completely, i.e. without truncation? Are dialog boxes or error messages adequately (re)sized? Does the text fit on the screen in different resolutions? Is the localized version aesthetically acceptable? Do drop box designs display all response options? Are the different format conventions (e.g., date or time formats) correctly implemented?

Eventually, the actual functionality of the software is focused on. Esselink (2000) – speaking for the software localization sector – holds that functionality testing usually mirrors the processes that have been implemented on the source product. Moreover, the more thoroughly the source product has been prepared and tested prior to localization, the fewer problems will be found during testing of the localized product. Transferred to the survey context, the key question to ask during this final testing is whether the entire questionnaire works as intended, or whether problems were introduced through translation. For survey translation, functionality testing should involve the use of various testing scenarios that cover respondent groups receiving different parts of the questionnaire. Such a full-blown test can also identify whether the translation works in the context of more extensive routing and in different types of "paths". During the actual translation, questions are translated in a linear fashion one after the other. In a concrete survey context, however, this linear fashion may not be applicable since routing based on a given answer may send a respondent to questions much later in the questionnaire. Hence, this real-life testing is extremely important for ensuring that the questionnaire is intelligible in the different "paths" that a respondent could take through the survey.

In the localization industry, localized software often undergoes compatibility testing that checks how compatible a new product is with other products that are available in the target language (e.g., platforms, devices, web browsers). This type of testing can be crucial for translated surveys as well. For instance, Dillman et al. (2014, p. 345) discovered in one of their web studies that toggling back and forth between an English and Spanish questionnaire only worked in some versions of browsers, but not in others.

Last but not least, a remark seems appropriate on how this technical pretesting step relates to usability testing, which has gained momentum in survey research due to the rise of computer-assisted surveys. The main goals of usability testing are to improve data quality by the reduction of errors, and to prevent item or unit non-response by the reduction of respondent burden. Usability testing should not be confused with best practice questionnaire design; rather it should build upon best practice design and provide the ultimate test that ensures that interviewers and respondents can record answers easily and accurately (Geisen & Romano Bergstrom, 2017). Transferred to multilingual surveys, usability testing should first and foremost target the questionnaire in general, and as such it should be part of the aforementioned testing of the source questionnaire. However, surveys that undergo

larger cultural adjustments (e.g., a change in the writing direction) may be in need of additional usability testing once the translated version is available. This testing may still be similar to the interface and functionality testing described above.

5.8 Discussion and Recommendations

The aforementioned observations have hopefully shed some light on the complexities that can come with implementing a multilingual and multicultural survey questionnaire in a computerized survey environment. The good news is that the complexities can be mastered. What is important, however, is that technical aspects are considered early on in the development process in a multilingual survey. Just as early cross-cultural collaboration, translatability assessments or advance translations are important to ensure ease of translation and cultural relevance, so are early checks in ensuring that the software and programming work in a multilingual context. That is: When developing and programming a source questionnaire for a multilingual study, the diversity of study languages and their respective needs should always be considered. Ultimately, a well-designed survey on the linguistic, cultural as well as technical level is the pre-condition for sound data.

We require respondents to invest time and effort into replying to our questions, even though their (intrinsic) motivation may be low. We should, on our side, invest time and effort into providing questionnaires that are linguistically and culturally appropriate and function as intended. The best possible way to achieve this is early cooperation between survey developers, linguists, and translation technologists (see Lupsa, 2018, cited in Behr & Zabal, 2019). Prior to finalizing the design in the source language, issues or challenges for other languages can thus be identified and remedies found. In the same vein, Valli (2019) states for the field of software localization: "At minimum, the localization team should be *involved* in those product development phases in order to raise awareness about the future linguistic pitfalls." Checklists as the one provided in the appendix – or the work by Esselink (2000) or Microsoft (2018) – may additionally help to inform multilingual technical survey design. Furthermore, planning and workflows will need to cater to these additional layers of cooperation.

Finally, testing of translations in the computerized survey environment will need to be factored in both time- and budget-wise, and so does another loop of adjustments following the feedback from a first round of multilingual linguistic, interface, and functionality testing. In particular in languages that a research team does not speak (e.g., languages of refugees in a country) additional resources are required for this testing.[9]

[9] Some general information on scheduling and budgeting translation activities can be found here: https://ccsg.isr.umich.edu/chapters/translation/ (accessed 4 July 2021).

To conclude, the translation approach is rarely described (in detail) in multilingual migration research. It would be helpful if researchers document their procedures and include challenges encountered or lessons learned both in translation and in technical survey implementation. This way, surveys can raise awareness on challenges, they can learn from each other, and build on each other's experiences. We already see this transfer of lessons learned with other types of study challenges in migration research, for instance on sampling, field work or item understanding (Formea et al., 2014; Haug et al., 2019; Röder, 2018).

Appendix: Checklist – With a Focus on the Interplay Between Translation and Survey Technology

In General

- Specify the language(s) for a study and check whether it/they can be implemented in the chosen or planned survey software (see Sect. 5.4.1).
  - This includes checking to what extent text in different scripts can be entered, particularly in bilingual response situations (see Sect. 5.4.4).
- Check if multilingual recording can be added, if desired.

During Questionnaire Development and Implementation

- Identify textual elements that may require cultural adjustments (e.g., response formats in terms of date, time, name, address, telephone number or decimal separators) and decide on ways how this can be handled in the software (see Sect. 5.4.2).
- Identify non-textual elements that may require cultural adjustments (e.g., icons, images, links or colors) and decide on ways how this can be handled in the software (see Sect. 5.4.3).
- Identify fills in the questionnaire (if any are planned) and check to what extent these can be replicated in other languages (see Sect. 5.4.4).
- Consider text expansion in translation compared to the source questionnaire and check to what extent buttons, dialog boxes, text boxes, etc. cater to this (see Sect. 5.4.4).
- Plan for ways to allow cultural groups to adapt content, e.g. to suppress an irrelevant response category or to add a question (see Sect. 5.4.5).
- Prepare the source questionnaire in line with format and style requirements that pave the way for efficient use of computer-aided translation (CAT) tools (see Sect. 5.4.6).
- Once implemented, test the source questionnaire thoroughly.

Prior to Translation

- Provide instructions and further pertinent information to translation teams (see Sect. 5.5).
  - This should include information on particular features of the interface or technical questionnaire design (e.g., on the use of fills) (see Sect. 5.6).

Translation, Including Adaptation

- Consider having translators translate with CAT tools to render the translation process more efficient and less error-prone (see Sect. 5.6).
- Translate with the interface and the actual survey implementation in mind.
- Conduct linguistic, interface, and functionality testing of the translated versions (see Sect. 5.7).

References

Acquadro, C., Conway, K., Hareendran, A., Aaronson, N., & European Regulatory Issues and Quality of Life Assessment (ERIQA) Group. (2008). Literature review of methods to translate health-related quality of life questionnaires for use in multinational clinical trials. *Value in Health, 11*(3), 509–521. https://doi.org/10.1111/j.1524-4733.2007.00292.x

Acquadro, C., Patrick, D. L., Eremenco, S., Martin, M. L., Kuliś, D., Correia, H., Conway, K., & International Society for Quality of Life Research. (2018). Emerging good practices for translatability assessment (TA) of patient-reported outcome (PRO) measures. *Journal of Patient-Reported Outcomes, 2*(8), 1–11. https://doi.org/10.1186/s41687-018-0035-8

AIFS – Australian Government, Department of Social Services, & Australian Institute of Family Studies. (2018). Building a new life in Australia: The longitudinal study of humanitarian migrants. *Data Users Guide: Release 4.0*. https://aifs.gov.au/sites/default/files/bnla_data_users_guide_release_4.pdf. Accessed 6 October 2019. Accessed 18 Oct 2019

Behr, D. (2018a). Translating questionnaires for cross-national surveys: A description of a genre and its particularities based on the ISO 17100 categorization of translator competences. *Translation & Interpreting, 10*(2), 5–20. https://doi.org/10.12807/ti.110202.2018.a02

Behr, D. (Ed.) (2018b). Surveying the migrant population: Consideration of cultural and linguistic issues. GESIS-Schriftenreihe 19. http://nbn-resolving.de/urn:nbn:de:0168-ssoar-58074-2

Behr, D., & Scholz, E. (2011). Questionnaire translation in cross-national survey research: On the types and value of annotations. *mda: Methoden, Daten, Analysen, 5*(2), 157–179.

Behr, D., & Shishido, K. (2016). The translation of measurement instruments for cross-cultural surveys. In C. Wolf, D. Joye, T. W. Smith, & Y. Fu (Eds.), *The SAGE handbook of survey methodology* (pp. 269–287). Sage.

Behr, D., & Zabal, A. (2019). A meeting report: OECD-GESIS seminar on translating and adapting instruments in large-scale assessments (2018). *Measurement Instruments for the Social Sciences, 1*, 10. https://doi.org/10.1186/s42409-019-0011-y

Behr, D., Dept, S., & Krajčeva, E. (2018). Documenting the survey translation and monitoring process. In T. P. Johnson, B.-E. Pennell, I. Stoop, & B. Dorer (Eds.), *Advances in comparative survey methods: Multinational, multiregional and multicultural contexts (3MC)* (pp. 341–356). John Wiley & Sons.

Behr, D., Meitinger, K., Braun, K., & Kaczmirek, L. (2019). Cross-national web probing: An overview of its methodology and its use in cross-national studies. In P. C. Beatty, D. Collins, L. Kaye, J. L. Padilla, G. Willis, & A. Wilmot (Eds.), *Advances in questionnaire design, development, evaluation and testing* (pp. 521–544). Wiley.

Das, M., Vis, C., & Weerman, B. (2005). Developing the survey instruments for SHARE. In A. Börsch-Supan & H. Jürges (Eds.), *The survey of health, aging, and retirement in Europe–methodology*. MEA. http://www.share-project.org/uploads/tx_sharepublications/SHARE_BOOK_METHODOLOGY_Wave1.pdf. Accessed 6 Oct 2019

De la Cova, E. (2016). Translation challenges in the localization of web applications. *Sendebar: Revista de la Facultad de Traducción e Interpretación, 27*, 235–266.

Dept, S., Ferrari, A., & Halleux, B. (2017). Translation and cultural appropriateness of survey material in large-scale assessments. In P. Lietz, J. C. Cresswell, K. F. Rust, & R. J. Adams (Eds.), *Implementation of large-scale education assessments* (pp. 168–192). John Wiley & Sons.

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2014). *Internet, phone, mail, and mixed-mode surveys: The tailored design method*. John Wiley & Sons.

Dorer, B. (2011). *Advance translation in the 5th round of the European Social urvey (ESS). FORS working paper series 2011-4*. FORS. https://doi.org/10.24440/FWP-2011-00004

Dorer, B. (2020). *Advance translation as a means of improving source questionnaire translatability?: Findings from a think-aloud study for French and German*. Frank & Timme.

Eremenco, S., Pease, S., Mann, S., & Berry, P. (2018). Patient-reported outcome (PRO) consortium translation process: Consensus development of updated best practices. *Journal of Patient-Reported Outcomes, 2*(12), 1–12. https://doi.org/10.1186/s41687-018-0037-6

Esselink, B. (2000). *A practical guide to localization*. John Benjamins Publishing.

Formea, C. M., Mohamed, A. A., Hassan, A., Osman, A., Weis, J. A., Sia, I. G., & Wieland, M. L. (2014). Lessons learned: Cultural and linguistic enhancement of surveys through community-based participatory research. *Progress in Community Health Partnerships: Research, Education, and Action, 8*(3), 331–336.

Gambier, Y. (2016). Rapid and radical changes in translation and translation studies. *International Journal of Communication, 10*, 887–906.

Geisen, E., & Bergstrom, J. R. (2017). *Usability testing for survey research*. Morgan Kaufmann.

Goerman, P. L., & Caspar, R. A. (2010). Managing the cognitive pretesting of multilingual survey instruments: A case study of pretesting of the US Census Bureau bilingual Spanish/English questionnaire. In J. A. Harkness, M. Braun, B. Edwards, T. P. Johnson, L. Lyberg, P. P. Mohler, B.-E. Pennell, & T. W. Smith (Eds.), *Survey methods in multinational, multiregional, and multicultural contexts* (pp. 75–90). Wiley.

Goerman, P., Meyers, M., & García Trejo, Y. (2018). The place of expert review in translation and questionnaire evaluation for hard-to-count populations in national surveys. In D. Behr (Ed.), *Surveying the migrant population: Consideration of cultural and linguistic issues* (pp. 29–41). GESIS. http://nbn-resolving.de/urn:nbn:de:0168-ssoar-58074-2

Hansen, S. E., Jung Lee, H., Lin, Y., & McMillan, A. (2016). Instrument technical design. In *Guidelines for best practice in cross-cultural surveys*. Survey Research Center, Institute for Social Research, University of Michigan. https://ccsg.isr.umich.edu/chapters/instrument-technical-design/. Accessed 4 July 2021

Harkness, J. (2003). Questionnaire translation. In J. Harkness, F. J. R. van de Vijver, & P. P. Mohler (Eds.), *Cross-cultural survey methods* (pp. 35–56). Wiley.

Harkness, J. A., van de Vijver, F. J. R., & Mohler, P. P. (Eds.). (2003). *Cross-cultural survey methods*. Wiley.

Harkness, J. A., Braun, M., Edwards, B., Johnson, T. P., Lyberg, L., Mohler, P. P., et al. (Eds.). (2010a). *Survey methods in multinational, multiregional, and multicultural contexts*. Wiley.

Harkness, J. A., Villar, A., & Edwards, B. (2010b). Translation, adaptation, and design. In J. A. Harkness, M. Braun, B. Edwards, T. P. Johnson, L. Lyberg, P. P. Mohler, et al. (Eds.), *Survey methods in multinational, multiregional, and multicultural contexts* (pp. 117–140). Wiley.

Haug, S., Lochner, S., & Huber, D. (2019). Methodological aspects of a quantitative and qualitative survey of asylum seekers in Germany – A field report. *Methods, Data, Analyses, 13*(2), 321–340.

International Test Commission. (2017). *The ITC guidelines for translating and adapting tests* (2nd ed.). https://www.intestcom.org/page/14. Accessed 4 July 2021.

Jacobsen, J. (2018). Language barriers during the fieldwork of the IAB-BAMF-SOEP survey of refugees in Germany. In D. Behr (Ed.), *Surveying the migrant population: Consideration of cultural and linguistic issues* (pp. 75–84). GESIS. http://nbn-resolving.de/urn:nbn:de:0168-ssoar-58074-2

Johnson, T. P., Pennell, B. E., Stoop, I. A., & Dorer, B. (Eds.). (2018). *Advances in comparative survey methods: Multinational, multiregional, and multicultural contexts (3MC)*. John Wiley & Sons.

Latour, T., Jadoul, R., & Wagner, M. (2013). *Development of the CAPI questionnaire system*. Technical report of the Survey for Adult Skills (PIAAC). https://www.oecd.org/skills/piaac/_Technical%20Report_17OCT13.pdf. Accessed 18 Oct 2019

Lerum, C. B., Nelson, J. A., de Matos Capistrano, A. Integrated application localization. U.S. Patent No. 8,789,015. 22 Jul. 2014. https://patentimages.storage.googleapis.com/cd/4d/51/bb3e2a7a524a78/US8789015.pdf. Accessed 18 Oct 2019.

Maroto, J., & De Bortoli, M. (2001). *Web site localization*. http://pure.au.dk/portal/files/11487/Appendix_3.pdf. Accessed 18 Oct 2019.

Martinez, M., Marin, V., & Schoua-Glusberg, A. (2006). Translating from English to Spanish: The 2002 National Survey of Family Growth. *Hispanic Journal of Behavioral Sciences, 28*(4), 531–545.

Meitinger, K., Behr, D., & Braun, M. (2019). Using apples and oranges to judge quality? Selection of appropriate cross-national indicators of response quality in open-ended questions. *Social Science Computer Review*. https://doi.org/10.1177/0894439319859848

Microsoft. (2018). Internationalization checklist. https://docs.microsoft.com/en-us/windows/win32/intl/internationalization-checklist. Accessed 18 Oct 2019.

Mohler, P., Dorer, B., de Jong, J., & Hu, M. (2016). Translation: Overview. In *Guidelines for best practice in cross-cultural surveys*. Survey Research Center, Institute for Social Research, University of Michigan. http://www.ccsg.isr.umich.edu/

Pan, Y., Landreth, A., Park, H., Hinsdale-Shouse, M., & Schoua-Glusberg, A. (2010). Cognitive interviewing in non-English languages: A cross-cultural perspective. In J. A. Harkness, M. Braun, B. Edwards, T. P. Johnson, L. Lyberg, P. P. Mohler, et al. (Eds.), *Survey methods in multinational, multiregional, and multicultural contexts* (pp. 91–113). Wiley.

Pan, Y., Sha, M., & Park, H. (2020). *The sociolinguistics of survey translation*. Routledge.

Pötzschke, S., & Braun, M. (2017). Migrant sampling using Facebook advertisements: A case study of polish migrants in four European countries. *Social Science Computer Review, 35*(5), 633–653. https://doi.org/10.1177/0894439316666262

Pym, A. (2011). Website localization. In K. Malmkjaer & K. Windle (Eds.), *The Oxford handbook of translation studies* (pp. 410–423). Oxford University Press.

Röder, A. (2018). Methodische Herausforderungen quantitativer Befragungen von Geflüchteten am Beispiel einer Vorstudie in Sachsen. *Z'Flucht Zeitschrift für Flüchtlingsforschung, 2*(2), 313–329.

Sandrini, P. (2008). Localization and translation. *MuTra Journal, 2*, 167–191. http://translationconcepts.org/pdf/MuTra_Journal2_2008.pdf#page=167. Accessed 18 Oct 2019

Schäler, R. (2010). Localization and translation. In Y. Gambier & L. van Doorslaer (Eds.), *Handbook of translation studies* (Vol. 1, pp. 209–214). John Benjamins B.V.

Sha, M., Hsieh, Y., & Goerman, P. (2018). Translation and visual cues: Towards creating road map for limited English speakers to access translated internet surveys in the United States. *Translation & Interpreting, 10*(2), 142–158. https://doi.org/10.12807/ti.110202.2018.a10

Sin-wai, C. (2016). *The future of translation technology: Towards a world without Babel*. Routledge.

Smith, T. (2004). Developing and evaluating cross-national survey instruments. In S. Presser, J. M. Rothgeb, M. P. Couper, J. T. Lessler, E. Martin, J. Martin, & E. Singer (Eds.), *Methods for testing and evaluating survey questionnaires* (pp. 431–452). Wiley.

Upsing, B., Gissler, G., Goldhammer, F., Rölke, H., & Ferrari, A. (2011). Localisation in international large-scale assessments of competencies: Challenges and solutions. *Localisation Focus, 10*(1), 44–57.

Valli, P. (2019). Fundamentals of localization for non-localizers. In B. Maylath & K. St. Amant (Eds.). Translation and localization: A guide for technical and professional communicators. : Routledge.

van de Vijver, F. J. R., & Leung, K. (2011). Equivalence and bias: A review of concepts, models, and data analytic procedures. In D. Matsumoto & F. J. R. van de Vijver (Eds.), *Cross-cultural research methods in psychology* (pp. 17–45). Cambridge University Press.

Wang, L., Sha, M., & Yuan, M. (2017). Cultural fitness in the usability of US census internet survey in Chinese language. *Survey Practice, 10*(3). https://doi.org/10.29115/SP-2017-0018

Wild, D., Grove, A., Martin, M., Eremenco, S., McElroy, S., Verjee-Lorenz, A., & Erikson, P. (2005). Principles of good practice for the translation and cultural adaptation process for patient-reported outcomes (PRO) measures: Report of the ISPOR Task Force for Translation and Cultural Adaptation. *Value in Health, 8*(2), 94–104.

Wild, D., Eremenco, S., Mear, I., Martin, M., Houchin, C., Gawlicki, M., et al. (2009). Multinational trials—Recommendations on the translations required, approaches to using the same language in different countries, and the approaches to support pooling the data: The ISPOR patient-reported outcomes translation and linguistic validation good research practices task force report. *Value in Health, 12*(4), 430–440.

Willis, G. B. (2005). *Cognitive interviewing: A tool for improving questionnaire design*. Sage.

Willis, G. B. (2015). Research synthesis. The practice of cross-cultural cognitive interviewing. *Public Opinion Quarterly, 79*, 359–395.

Willis, G. B. (2016). Questionnaire pretesting. In C. Wolf, D. Joye, T. W. Smith, & Y. Fu (Eds.), *The SAGE handbook of survey methodology* (pp. 359–381). Sage.

Chapter 6
Surveying Illiterate Individuals: Are Audio Files in Computer-Assisted Self-Interviews a Useful Supportive Tool?

Florian Heinritz, Gisela Will, and Raffaela Gentile

6.1 Introduction

Due to the strong increase in the number of asylum seekers coming to Europe in the mid-2010s, several countries have started to survey new immigrants. Since new immigrants usually have insufficient competencies in the language of a receiving country, translations of questionnaires are essential if they are to participate in such surveys. With respect to refugees,[1] the available information on the general level of education in the main countries of origin shows that a considerable portion have poor or no reading skills in their native language. To avoid a systematic exclusion of this population, it is necessary to apply further strategies. This chapter presents strategies that enable the inclusion of illiterate immigrants in a survey: common personal interviews with native-speaking interviewers and computer-assisted self-interviews with audio files. First, we discuss the pros and cons of both approaches (Sect. 6.2). Since audio files are a relatively new opportunity made possible by modern technology, the subsequent section summarizes the methodological challenges of and first experiences with using audio files in surveys (Sect. 6.3). After this overview of the current state of research, we focus on the ReGES study (for "Refugees in the

[1] In this chapter, the term *refugee* is not used in the strict sense of the Geneva Convention, but rather it includes all asylum seekers looking for protection.

F. Heinritz (✉)
Universität Hamburg, Hamburg, Germany

Leibniz Institute for Educational Trajectories, Bamberg, Germany
e-mail: florian.heinritz@uni-hamburg.de

G. Will · R. Gentile
Leibniz Institute for Educational Trajectories, Bamberg, Germany
e-mail: gisela.will@lifbi.de; raffaela.gentile@lifbi.de

S. Pötzschke, S. Rinken (eds.), *Migration Research in a Digitized World*, IMISCOE Research Series, https://doi.org/10.1007/978-3-031-01319-5_6

Table 6.1 Literacy rates of different countries of origin compared to the EU (persons over 15 years, numbers in percentages)

Country	Total	Men	Women	Year of data
Afghanistan	31.74	45.42	17.61	2011
Eritrea	64.66	75.07	54.80	2008
Iraq	43.68	53.01	37.96	2013
Nigeria	51.08	61.25	41.39	2008
Pakistan	56.98	69.07	44.28	2014
Syria	80.84	87.76	73.63	2004
EU	99.13	99.33	98.93	2016

Source: UNESCO dataset (data.uis.unesco.org); Accessed 10 Sept 2019

German Educational System"),[2] in which both types of interview strategies to include illiterate participants were offered as a combined approach to optimize response rates and minimize response errors, including social desirability bias. Using the data of the first wave, the effectiveness and practicability of these strategies are considered in more detail (Sect. 6.4). In the discussion (Sect. 6.5), based on the experience from this study, we make recommendations for future research. In this context, we also highlight limitations, and identify additional methodological research desiderata.

6.2 Problem Statement and Status Quo

Generally, attempts are made to interview new immigrants as soon as possible after they arrive in their destination country. This fact makes it essential to translate questionnaires into the immigrants' languages of origin because they rarely have sufficient language skills to complete a questionnaire in the language of the destination country (on the challenge of translating survey instruments, see Behr, this volume). In addition, when interviewing a specific group of refugees, it is important to take into account the presumed proportion of illiterate individuals within this group. An inspection of the available data on literacy rates in the top six countries of origin of refugees in the EU in the mid-2010s shows that despite country-specific differences, the percentage of illiterate people is considerable (see Table 6.1).[3]

The literacy rate in Syria is relatively high at 80.8%, whereas in countries such as Afghanistan or Iraq, over half of the population over the age of 15 cannot participate in all activities of their community that require reading, writing, and calculating.

[2]The ReGES project is funded by the German Federal Ministry of Education and Research under grant number FLUCHT03. However, the authors have sole responsibility for the content of this publication.

[3]UNESCO generally refers to functional illiteracy in its data. According to this data, persons are *functionally illiterate* if they cannot participate in all the activities of their community that require reading, writing, and calculating (see UNESCO, 2006). For our subsequent analyses with the ReGES data, we defined people who claim to be able to read very poorly or not at all in their native language as *illiterate*.

Clear differences also exist between the genders, since all countries have lower literacy rates among women. Thus, an assumption can be made that a relatively high proportion of illiterate people can be found among those coming to the EU from these countries of origin. Initial findings on the education of refugees support this assumption, at least regarding the example of immigration of refugees to Germany. Surveys on the formal education of asylum seekers in Germany (Rich, 2016; Neske & Rich, 2016) have shown that in the first half of 2016, 27% of Afghans seeking asylum in Germany did not have any formal education (Neske & Rich, 2016). If the distortion of educational selectivity is to be avoided when doing research on new immigrants, those who have minimal or no reading and writing skills must also be able to participate in surveys.

One way to include illiterate persons and those with low reading and writing skills is to interview them personally (e.g., Grotlüschen & Riekmann, 2012; Nienkemper, 2015); a second, more modern strategy is to carry out computer-assisted interviews with supplemental audio files (e.g., Jacobsen, 2018; Kühne et al., 2019; Schroder et al., 2003; Turner et al., 1998a). Such audio files sometimes are used in personal interviews (see Jacobsen, 2018; Kühne et al., 2019)—so interviewers do not have to speak the relevant language—but predominantly, they are used in computer-assisted self-interviewing (CASI).

When surveying illiterate individuals, both strategies can be beneficial, but their disadvantages also must be considered, especially if these strategies are to be used to complement each other. Thus, in the next two subsections, we discuss the pros and cons of these two interview modes, in particular regarding illiterate individuals, before we present empirical experiences with audio files in Sect. 6.3, and then focus on our empirical experiences as researchers offering a study with audio files (Sect. 6.4).

6.2.1 Personal Interviews as a Solution to Survey Illiterate Individuals

The most common way to include illiterate individuals and those with low reading and writing skills in a survey is to interview them personally. When surveying newly arrived immigrants, personal face-to-face interviews require the employment of native-speaking interviewers or interviewers who have at least a sufficient command of the immigrants' language of origin.[4] Face-to-face interviews are especially

[4] In addition, previous research has shown that addressing migrants in their language of origin has positive effects, regardless of whether they are illiterate or not and whether the migrants also speak the language of the host country or not (e.g., Baykara-Krumme, 2010). Moreover, if an interviewer and a respondent speak the same language, the interviewee's possible questions or concerns can be answered adequately (see Allerbeck & Hoag, 1985), and a problem-free communication can be ensured. In addition, a less direct effect is that social proximity can be established by matching the mother tongue as an important ethnic characteristic (see Haarmann, 1983; Siegel, 2018), which can

beneficial because they avoid placing illiterate individuals in a situation in which written language plays a large role and which they might incorrectly perceive as a test (Nienkemper, 2015).

In addition, another advantage of interviewers being present is that they can be generally helpful, since they can answer all respondents' questions about the interview, regardless of whether they can read or not. Moreover, the presence of an interviewer is even more important to illiterate individuals who cannot use survey-relevant documents (e.g., cover letters, information on data protection, study information) to gain relevant information. If the written documents of a survey cannot be read, only an interviewer can convince potential illiterate respondents to participate, e.g., by explaining the objectives of the study. Therefore, in certain situations, the presence of interviewers is especially important, particularly when respondents are first contacted during the recruitment process for a study. While interviewers always can affect an interviewee's willingness to participate (see Groves & Couper, 1998; Hox & de Leeuw, 2002), an assumption can be made that this effect is even more pronounced for illiterate individuals. The same assumption seems to hold in situations when interviewees have had little prior experience with scientific surveys. Especially in the case of people who cannot read well, it is extremely important that interviewers personally explain the aspects of data protection and anonymity. Some specific groups of refugees often are only familiar with an interview in the context of their asylum procedure, which in many countries entails interviews conducted by official migration authorities. Survey staff can be specially trained to explain the differences between a scientific survey, which is characterized by voluntariness and anonymity, and interviews conducted by official migration authorities.

Thus, the use of interviewers has great advantages when surveying illiterate individuals. However, conducting face-to-face interviews also has several disadvantages. The greatest drawback of this strategy is certainly the influence that an interviewer can have on an interviewee's response behavior, which is true not only for illiterate individuals and immigrants, but generally for all respondents. In all the steps of the interview process, interviewer effects may appear, starting with increased coverage errors and unit nonresponse, and ending with outright measurement error (see West & Blom, 2017). The mere presence of interviewers and their characteristics such as gender, ethnicity, or age can impact response behavior (e.g., Glantz & Michael, 2014; Groves et al., 2009; Loosveldt, 2008).

Especially with respect to newly arrived refugees, an assumption can be made that in the presence of interviewers, they will show a response behavior adapted to the standards of their host country so to demonstrate their willingness to integrate, among other things (see Haug et al., 2017). This effect may be stronger if the interviewer visibly belongs to the host country's main ethnic group. However, even when an interviewer has the same ethnic origin as the interviewee (e.g.,

positively impact the migrant's willingness to participate (see Feskens et al., 2006). The willingness to cooperate also can be influenced strongly by mechanisms such as the motive of helping or liking (see Groves et al., 1992), a more pleasant atmosphere and a stronger basis of trust (see Dotinga et al., 2005), and through an awareness of solidarity based on common origin and collective identity (see Heckmann, 1992).

Kappelhof, 2014; van Heelsum, 2013), it is possible to observe different response behaviors.

Depending on the refugees' cultural background, an interviewer's gender may have a very strong influence on response behavior and the willingness to cooperate (see Baykara-Krumme, 2012; Blohm & Diehl, 2001). Furthermore, experts have argued that especially in countries with many dialects, the dialect of a person—the interviewer's or the interviewee's—could be used to draw conclusions about a person's political orientation (see Feskens et al., 2006).

Even if these theoretical considerations should be decisive when choosing a strategy for interviewing illiterate people, it also is important to consider that personal interviews with illiterate people who speak a foreign language can be successful only if it is possible to recruit foreign-language-speaking interviewers. Furthermore, it should not be underestimated that personal interviews usually are associated with higher costs.

6.2.2 Computer-Assisted Self-Interviewing with Audio Files as a Solution to Survey Illiterate Individuals

CASI with audio files is an alternative to the personal interview of illiterate individuals, and they can be performed in two ways. On the one hand, audio files can be used as an additional option to help these respondents understand the CASI questions. The additional audio files support the wording of the survey question, which is always displayed as text on the computer screen. On the other hand, a specific version of CASI with audio files—audio computer-assisted self-interviewing (ACASI)—can be used, which focuses more on the audio files than the written text. When using ACASI, sometimes the text is not even displayed. In research with ethnic minorities, this no-text strategy can be especially useful if they do not have a written language (see Cooley et al., 2001; Falb et al., 2016), since this avoids ambiguous transcriptions. Another technical feature in most multilingual ACASI approaches is that interviewees can switch to any other survey language at any time. This is useful, for example, if an interviewee speaks two native languages or understands some terms better in another language (e.g., in the language of the host country).

Compared to a face-to-face interview, CASI with audio files reduces interviewer effects to a minimum and still enables illiterate individuals to participate in a survey. Especially with regard to sensitive questions such as those about religion or health, CASI does not, in principle, produce any distortions (e.g., Couper et al., 2002). This is all the more important given that the presence of other people (e.g., family members) also can influence response behavior (e.g., Aquilino et al., 2000; Chadi, 2013). Since studies with refugees often include interviews in collective accommodations or in constricted living conditions, CASI offers interviewees the possibility of answering questions without other residents or family members influencing their

response behavior. CASI provides a significantly greater anonymity and privacy than personal interviews in which questions and answers are spoken aloud. Also, if headphones are used to listen to the audio files, a feeling of isolation from the surroundings and more privacy are created (see Tourangeau & Smith, 1998). Therefore, when interviewing illiterate individuals, CASI with additional audio files is a less reactive alternative to personal interviews with native-speaking interviewers.

Furthermore, audio files are intended not only to help functionally illiterate people, but also to provide all respondents with an additional aid to understanding that helps to facilitate their completion of a questionnaire (see Tourangeau & Smith, 1996). Previous studies have assumed that the use of audio files could increase the motivation to complete even long self-administered questionnaires (see Edwards et al., 2007). With respect to the practical implementation of a survey, one advantage is that an interviewer does not have to be present, which reduces the costs enormously.

However, as shown in Sect. 6.2.1, the willingness to participate in CASI surveys is lower than in personal interviews. Especially when questioning newcomers and illiterate people for the first time, personal contact seems indispensable. Furthermore, in a self-administered interview, it is usually not possible to query complex facts, especially for illiterate people if they rely only on audio files. The presence of an experienced interviewer can be very helpful, particularly in regard to dealing with difficult parts of a questionnaire. In summary, although a CASI survey can be a good way to interview illiterate people, especially on sensitive issues, its questions must not be too complex, and interviewees must be able to understand them by listening to audio files.

6.3 General Experience from the Field: Implementation and Usage of Computer-Assisted Self-Interviewing with Audio Files

To date, the number of migration studies that use CASI with audio files is rather limited (e.g., Mierzwa et al., 2013; Falb et al., 2016; Turner et al., 1996; Wong et al., 2007). Nevertheless, some fieldwork experiences are available, which we review in the following subsections. The implementation of audio files in a CASI survey depends strongly on the sample, research design, and software used, so it is not always possible to extract general points. First, we discuss the practical experiences of implementing audio files (Sect. 6.3.1) and then devote more attention to the results of extant methodological research on the effects of audio files (Sect. 6.3.2).

6.3.1 Implementation of Audio Files

Since using ACASI is a relatively new technical solution, many studies had to develop their own software for conducting ACASI (e.g., Beier et al., 2014; Mierzwa et al., 2013; Morina et al., 2017). Each of these studies implemented their ACASI, depending on whether the questions were to be answered via audio files only or via audio files supported by written text. In its simplest form, ACASI using only audio files should be kept very simple, with colored buttons or symbols as labels to indicate the available options (see Falb et al., 2016; Mierzwa et al., 2013).

Regarding CASI with written text and supporting audio files, several issues need to be considered when implementing the additional audio files. First, to ensure that the questions also are comprehensible for persons with reading difficulties, all the question texts, completion notes, and answer options—e.g., all the texts that the respondents see—should also be provided as audio files. Ideally, the currently read text would be highlighted, and then the program would automatically scroll to the next text to be read (see Beier et al., 2014).

The possibility exists that the audio files could play automatically on each new questionnaire page or that a respondent could actively start the audio file. An audio file can be read out loud when an interviewee presses the corresponding icon; and each text module (e.g., each answer category, question text, and completion note) can have its own audio file, or several text modules can be combined into one audio file. Each of these approaches has its own advantages and disadvantages. If each text module is provided with an audio file, the interviewee can have each text module (e.g., each answer category) individually read aloud. If interviewees do not understand one answer category, they can have that answer category read to them, rather than all the answer categories or even the entire item read again. For interviewees, this saves time and can have a positive effect on their motivation. However, providing a single track for each text module is technically complex, since considerably more audio files have to be recorded and implemented. If several text modules are combined in one audio file (e.g., question texts with completion instructions as one audio file, and all answer options as one audio file), all the modules are read to the interviewee in the same order as they appear on the screen. The most important argument for including questions and hint texts in one file is that this strategy guarantees that all the interviewees receive all the necessary information. Thus, this procedure contributes significantly to the standardization of the questionnaire and is especially important for answer categories. The recording of all answer categories in one audio file ensures that respondents listen to all the answer categories and do not simply choose the first somewhat suitable one. However, to ensure that the interviewee who does not have reading and writing skills knows which answer to click, all the answer options must also be marked in a way they can understand.

Regardless of how the audio files are implemented, one important requirement is that interviewees have the possibility to repeat the audio files and to hear the text

again, in case they did not understand it the first time (see Beier & Schulz, 2015; Gatward, 2002).

Finally, depending on the target group, it sometimes may be necessary to give an interviewee precise instructions on how to use a tablet or computer in general, and audio files in particular (see Falb et al., 2016). Some experience with regard to the technical requirements of the computers used in the surveys is available, for example, the memory size of the audio files used can slow down the interview (see Couper et al., 2009) or lead to problems with using certain computers in general (see Beier & Schulz, 2015).

6.3.2 Research on Effects of Using Audio Files

Several studies that compared CASI questionnaires with audio files to face-to-face interviews have been consistent in their theoretical assumptions that CASI questionnaires with audio files should be preferred when measuring sensitive topics (see Hewett et al., 2004; Le et al., 2006; Newman et al., 2002), which is in line with prior methodological research on surveying sensitive topics (for an overview, see Couper et al., 2002). Van de Wijgert et al. (2000) conducted a survey in Zimbabwe with a combination of face-to-face interviews and ACASI. They found that more than four-fifths of the surveyed women preferred the ACASI questionnaire. Better privacy was mentioned as a reason. Evidence also exists that ACASI leads to a greater openness to sensitive questions, when compared to other self-administered interviews without audio files (see Turner et al., 1995, 1998b), even if mode effects cannot be clearly distinguished from the effect of additionally offering audio files (for this criticism, see Couper et al., 2009). In theory, ACASI without additional text can be particularly relevant when other people are present who are believed to be able to read the text presented on the screen (see Couper et al., 2003). For situations in which other people are not present, an assumption can be made that CASI without additional audio files is more suitable for querying sensitive information, since ACASI "introduces the presence of a 'virtual' interviewer into the situation" (Couper et al., 2003: 386; for a review of different studies, see Couper et al., 2009).

Nass et al. (1997) found evidence that virtual interviewers also can have an effect on respondents. Their experimental study with students found that answering questions about gender roles is affected by whether the questions in the audio files are read by a male or a female voice. However, the use of genderless computer voices does not seem to be an optimal solution because it extremely increases the rate of item nonresponse (Nass et al., 2003).

Regarding the necessary support from interviewers, previous studies have shown that, compared to paper-based self-filled-in questionnaires, ACASI could be completed more often without the help of the interviewer (see Lessler et al., 2000; Turner et al., 1998a). Also, an important finding is that respondents rated audio files to be more and more helpful to the degree that their reading skills were poor (Lessler et al., 2000). However, when respondents had a choice, the percentage of their usage of

audio files appeared not to be very high, and was significantly reduced by their having more education (e.g., Couper et al., 2009).

6.4 The ReGES Study—Practical Experiences

One study that used both audio files and face-to-face interviews to survey immigrants, and which included illiterate people, is the Refugees in the German Educational System (ReGES) study. Therefore, the focus of the following section is on the experiences of this study with respect to conducting CASI with audio files and computer-assisted personal interviewing (CAPI) with native-speaking interviewers.

6.4.1 Basic Information on the ReGES Study

The ReGES study aimed at describing the situation of young refugees in the German educational system and at presenting their educational pathways in more detail. For that purpose, adolescents (aged 14–16) and parents of young children (from the age of 4 who had not yet entered elementary school) were interviewed at different points from 2018 to 2020. The results discussed here relate to the first wave in 2018 (for more details on the study, see Will et al., 2021).

One of the prerequisites for participation in the study was that only refugees who came to Germany after January 1, 2014 were interviewed. Therefore, an assumption was made that many of the interviewees would not speak sufficient German to conduct a survey in German. As a consequence, all the documents and survey tools were offered in the following languages: Arabic, English, Farsi (Persian), French, German, Kurmanji, Pashto, and Tigrinya (for more details on the translation process and the selection of languages, see Gentile et al., 2019).[5]

6.4.2 Benefits of CASI with Audio Files and Personal Interviews Conducted by Native-Speaking Interviewers

For various reasons, it was not clear in advance whether respondents belonged to the target group of interest. This precondition of the contacting and recruiting strategy, the little-studied target group, and the expected number of illiterate people made the use of native-speaking interviewers mandatory. However, to gain the benefits of both the strategies of CAPI with native speaker interviewers and CASI with audio

[5] Due to time and financial reasons, unfortunately, it was not possible to offer more languages or different dialects.

files, while reducing their drawbacks, the ReGES study originally employed not only CAPI but also a combination of CAPI and CASI with audio files. To ensure that the respondents were part of the target population (see Steinhauer et al., 2019), a screening interview was initially conducted as CAPI. In addition to the screening questions, other factual information, such as the current educational situation of the children and adolescents, was collected using CAPI. The expectation was that all respondents—not only those who were illiterate—would need to depend on the help of an interviewer when answering questions about the complex educational system, and it also was expected that this would lead to only a slight distortion of the interviewee's responses to such factual questions. Ideally, each family should have been contacted by an interviewer who spoke the native language of the interviewee. Since the spoken language of the families could not be determined clearly in the run-up to the contact, teams of interviewers were formed to deliver all offered languages.

After the screening interview, the subsequent CASI with audio files was offered to reduce nonresponse and response errors, including social desirability. Within the CASI questionnaire, which was conducted on a tablet, more sensitive questions were asked on topics such as migration biography, origin, or attitudes.

CASI should keep the influence of interviewers as low as possible and guarantee a certain anonymity, which also seemed particularly relevant regarding young interviewees who might be influenced by the presence of their parents, as well as regarding interviews in collective accommodations. To avoid excluding illiterate people from using CASI and to enable them to participate, it was essential to provide them with audio files. These audio files had the added benefit of preventing illiterate people from having to identify themselves. In addition, the use of these files aimed to reduce the risk of interviewees with poor reading skills—who might not inform the interviewer of this limitation beforehand—having difficulty understanding questions and then guessing at the answers.

To ensure their privacy, all interviewees received headphones, so no one else could hear the interview questions. Ideally, audio files should substitute for the interviewer and lower the effects of socially desirable response behavior. Although an interviewer needed to be present at all times, his/her role had to be passive during the entire CASI process. Mainly, interviewers only were allowed to react to questions of understanding or technical problems, and had to refrain from intervening on their own initiative to maintain the interview situation as anonymous as possible. Moreover, the interviewer was not able to see the point at which the respondents were in the CASI questionnaire or how they answered the questions. In addition, the interviewers were not allowed to look at the tablet unless the respondents asked for support.

Implementing audio files, especially in combination with various foreign interview languages, poses several challenges, which are explained in more detail in the following subsections.

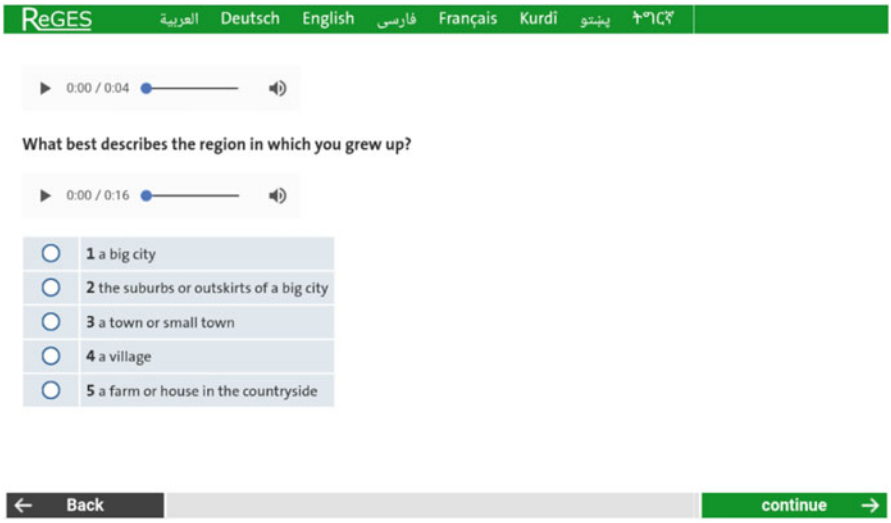


Fig. 6.1 Screenshot of an item from the CASI questionnaire with audio files (item adapted from the study Causes and Consequences of Socio-Cultural Integration Processes among New Immigrants in Europe [SCIP]; layout and programming of the CASI questionnaire provided by the infas Institute for Applied Social Sciences GmbH, Bonn, Germany)

6.4.3 Implementation of Audio Files in the ReGES Study

The ReGES study used an existing software solution (Gess Q) into which audio files were incorporated. The interviewee had to press a corresponding icon to listen to the audio files, and all questions and reference texts per item were recorded as one audio file. Following the same procedure, one audio file per item was recorded for all the response options of an item, and implemented so that the respondent usually had two audio files per item and per language available.

Indo-Arabic numerals were chosen to indicate the response options. They were all typed boldly and separated by a blank from the response options (e.g., "**1** yes; **2** no"; see Fig. 6.1 for an example of a screenshot from the CASI questionnaire). The corresponding numbers and answer options were then read out by the audio files, so the respondent could link the answer options they heard with the numbers preceding the options. In light of the respondents' daily use of mobile phones and computers, it was assumed that all respondents of different origins, and illiterate respondents as well, could easily read Indo-Arabic numbers. Prior to the study, this assumption was vetted in detail in expert discussions.

Recording one audio file for each question text and one audio file for all response options per item also reduced the effort involved with, and complexity of, making audio files by reducing the number of audio files that had to be recorded and implemented. Nevertheless, a total of approximately 40,000 audio files were integrated into the instrument. All the audio files were recorded by professional

speakers; five foreign languages were set to audio by female speakers; and Tigrinya, Pashto, and Kurmanji were spoken by male speakers.

6.4.4 Adaption of the Originally Planned Design

In the run-up to the field phase, the audio files were tested intensively after their implementation in the quality inspection phase. In a few cases, the audio files were assigned incorrectly to the items, and some audio files were completely missing. Even though the error rate was very low at less than 0.5% of all audio files, due to the complexity of the instrument and the upcoming field start, the faulty and missing audio files could not be recorded and implemented again. Instead, all the faulty and missing audio files were replaced with a standard audio file, prompting the respondent to pass the tablet to the interviewer for this question. Then, the interviewer would read the questions with the missing or faulty audio files, so the respondents who could not read sufficiently would not skip any relevant questions or randomly choose their answers.

In addition, due to this faulty implementation of some audio files, it was discussed that the interviewer could possibly conduct the CASI questionnaire together with an illiterate interviewee, who, otherwise, would have had to rely only on the incomplete CASI audio files. This measure would also ensure that all question texts and answers were read out correctly to illiterate interviewees. Therefore, the interviewers were instructed to actively offer the option of reading out loud the CASI text in whole or in part, and conducting the CASI questionnaire together with the interviewee.[6] Interviewers should offer this option to interviewees before they provide them with a tablet. This offer should be made especially if an interviewer noticed signs that an interviewee had reading and comprehension difficulties. Thus, the modes described in Fig. 6.2 were used.

6.4.5 Description of the Practical Experience of the ReGES Study

At the beginning of this section, we must emphasize that we cannot make any statements based on data concerning how illiterate interviewees would have behaved if native speakers were not available. Also, respondents were not assigned randomly to the above-mentioned two strategies for completing the CASI questionnaire: both self-selection and the interviewer's influence may have played a role when

[6] Even if, from a technical point of view, the CASI is no longer a self-interview when an interviewer reads questions aloud to an interviewee, it must be taken into account in the following analyses that we continued to speak of CASI even when parts of the interviews were read out loud.

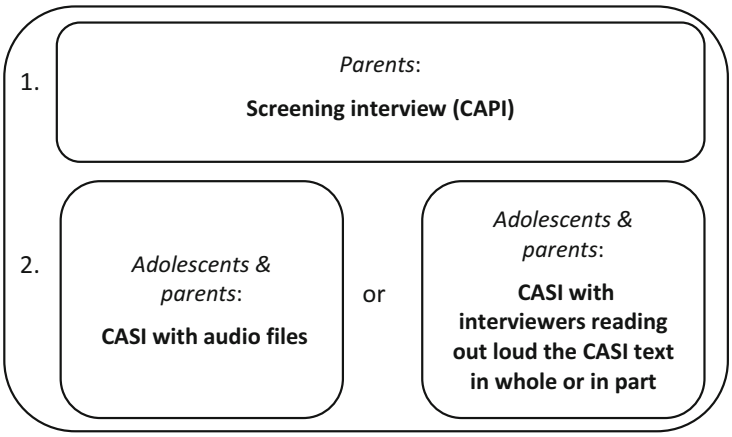


Fig. 6.2 Modes used in the ReGES study in wave 1

Table 6.2 Literacy rate of the ReGES sample's four most frequent countries of origin and total, numbers in percentages

Literacy rate	Syria	Iraq	Afghanistan	Iran	Total
Parents ($n = 3292$)	92.29	84.40	85.77	92.45	90.86
Adolescents ($n = 2406$)	96.17	96.23	95.45	94.44	95.18

Source: ReGES data, parents- and adolescents-CASI, Wave 1

distributing interviewees across these two options. While these limitations must be taken into account, we believe that our results are useful for researchers who focus on immigrants with high rates of illiteracy.

We divide our description into three parts. First, we describe the extent to which illiterate people participated in the survey (Sect. 6.4.5.1). In addition, we discuss the influence of native-speaking interviewers on different aspects of the study (Sect. 6.4.5.2). Finally, we present practical experiences regarding audio files (Sect. 6.4.5.3).

6.4.5.1 Self-Rated Reading Skills of Interviewees

The data show that a significant proportion of interviewees said that they could not, or only poorly, read in their mother tongue. For example, almost one-tenth of the parents said that they could read their mother tongue either very poorly or not at all (see Table 6.2). As expected, the proportion of illiteracy among young people was lower, but in their case, 5% would have had considerable difficulty understanding the questions in a survey mode that required literacy skills.

Nevertheless, the proportion of presumable illiteracy in the ReGES sample was significantly lower compared to UNESCO data on their countries of origin. This difference suggests that participants in the study tended to be more educated than the

population in the country of origin (for the first indications that refugees in Germany are a positively selected group, see Spörlein et al., 2020). In addition, it cannot be ruled out that illiterate people participated less frequently or did not state their reading skills truthfully in the survey.

6.4.5.2 Native-Speaking Interviewers

To reach the specific target group of refugees and include illiterate people, two strategies were discussed in this contribution: native-speaking interviewers conducting CAPI, on one hand, and CASI with audio files, on the other. To evaluate the advantages and disadvantages of both strategies, we started by analyzing the use of native-speaking interviewers in the ReGES face-to-face screening interview.

Our evaluation of the data showed that 83% of the screening interviews succeeded in establishing a language match between the interviewee and the interviewer. The high cooperation rate (80%) and relatively low refusal rate (11%) showed that the use of native speaker interviewers appears to have been beneficial at first glance. Furthermore, the data showed that more than three-quarters of the interviewees asked the interviewer for help during the subsequent CASI questionnaire. This indicates that, in many cases, an interview could not have been carried out successfully without the presence of an interviewer.

When interviewing ethnic minorities in their mother tongues, the effect of social desirability—which is relevant in any survey, but especially when conducting CAPI—may vary according to language match and ethnic background. We studied this theoretical assumption by focusing on the response behavior regarding questions that addressed a sensitive topic—gender roles. Whereas traditional gender roles are still prevalent in many of the refugees' countries of origin, most may be well aware that gender equity is a high priority in their host country Germany. Thus, social desirability could play a significant role in how interviewees answer these questions. We expected that they would tend to give more liberal answers when the screening interview was carried out in the German language and when the interviewer was from Germany.

On a scale ranging from 1 (*liberal*) to 4 (*conservative*), we found that interviewees gave more liberal answers to questions about gender roles when they performed the screening interview in the German language[7] (Table 6.3, Model 1). Controlling for interviewees' characteristics (gender, educational background of their family, age, country of origin), which also may influence gender role attitudes, exacerbates this effect (see Model 2). The characteristics of the interviewer (gender, age, migration background) also can have an effect on the interview language and the respondent's answering behavior. By keeping these characteristics constant, we

[7] We chose the screening interview language because we assumed that conducting the interview in German would frame the interview situation differently than conducting it in the interviewee's mother tongue.

Table 6.3 Attitude toward gender roles by language of screening interview, linear regression

	Model 1	Model 2	Model 3	Model 4
German in screening interview	−0.10** (0.03)	−0.11** (0.04)	−0.13*** (0.04)	−0.07+ (0.04)
Controls:				
Interviewee's gender, education, country of origin		✓	✓	✓
Interviewer's gender, age and migrant background			✓	✓
Interaction term: German in the screening-interview and reading out (= 1)				−0.43*** (0.10)
Constant	1.91*** (0.02)	2.08*** (0.07)	2.20*** (0.11)	2.21*** (0.12)
R^2	0.003	0.016	0.019	0.026
N	3175	3175	3175	3175

Source: ReGES data, own calculations, Wave 1
***Significant at $p < 0.001$, **significant at $p < 0.01$, *significant at $p < 0.05$, +significant at $p < 0.1$.
Standard deviations in parentheses

found that the effect of German as an interview language continues to increase both in terms of effect size and significance level (see Model 3).

Assuming that this effect would be enhanced when an interviewer read all the questions aloud (and therefore also the questions about gender roles) to the interviewee, we also looked at this interaction effect in our analysis. The significant interaction effect showed that interviewees gave even more liberal answers when an interviewer not only gave the screening interview in German, but also read aloud all questions to the interviewee (Table 6.3, Model 4).[8]

Interestingly, the migration background of the interviewer had no significant effects on the interviewee's response behavior. A plausible explanation for this finding may be that some interviewers in the ReGES study were born in Germany, even if they had a migration background, which made it more difficult for interviewees to identify their cultural affiliation, and so they may have seen them as representative of the majority, regardless of the interviewer's migration background.

In summary, using native-speaking interviewers may help to increase cooperation, and often during an interview, interviewers had to help interviewees. On the other hand, interviewers seemed to significantly influence interviewees' response behavior. Therefore, the additional possibility of using CASI with audio files to interview illiterate interviewees appears, at least, to be an alternative worth testing.

[8] However, it cannot be ruled out that interviewees who performed the interview in German might be a selective group (especially liberal, especially willing to integrate). Thus, the causal direction of our argument is by no means clear. Furthermore, as the low value of the R squared indicates, other factors exist that should be considered when explaining gender roles.

Table 6.4 Share of the sample that used audio files (in percentages) by the number of audio files used

Number audio files used	Illiterate	Not illiterate	Total
No audio file used	77.67	85.25	84.71
One audio file used	6.41	8.33	8.18
2–5 audio files used	5.70	3.83	3.96
6–20 audio files used	3.09	1.33	1.45
21–50 audio files used	3.33	0.53	0.74
More than 50 audio files used	3.80	0.74	0.96
N	*421*	*5281*	*5711*

Source: ReGES data, parents- and adolescents-CASI, Wave 1

6.4.5.3 Using Audio Files—Experiences from the Field

Only 873 ReGES interviewees used the opportunity to listen to the audio files on the CASI questionnaire. On average, the interviewees who used at least one audio file listened to 9.3 items. The rather high percentage of persons who used just one audio file, or listened to audio files just for the first few questions, may be explained by curiosity.

Almost 85% of interviewees did not use any audio files. Even more important, no interviewee answered the complete CASI solely by using audio files. Regarding a total length of 250–300 items per CASI, the maximum usage of audio files was 133 items. Our finding is in contrast to results of another German refugee study, the IAB-BAMF-SOEP study, where 7.5% of the respondents completed the entire questionnaire using audio files (see Kühne et al., 2019). However, the first wave of that study did not use any native-speaking interviewers. Therefore, if they wanted to participate in the survey, illiterate interviewees in the IAB-BAMF-SOEP had no alternative but to use the audio files. With regard to the ReGES study, Table 6.4 shows that illiterate interviewees used the audio files more often than those interviewees without reading difficulties.

The low usage of audio files in the ReGES study coincides with the feedback from the ReGES-interviewers involved during the fieldwork that indicated answering a CASI questionnaire of this length with audio files was cognitively demanding for illiterate people. Unlike in a personal interview, they had to choose the right answer and remember its corresponding option number when selecting it.

Nevertheless, even though the use of audio files did not seem to be as frequent as expected, the total number was less important than the fact that the interviewees with insufficient reading skills could use (some of) the audio files if they had problems with understanding. Examining the use of audio files in relation to self-assessed reading skills revealed a clear increase in audio file use when reading skills were low (see Fig. 6.3). However, by providing the interviewees with the opportunity to have the interviewers read the questions aloud, illiterate respondents did not have to rely solely on the audio files. The interviewees used this option far more often than they used the audio files. A total of 76.2% of the interviewees with no reading skills used

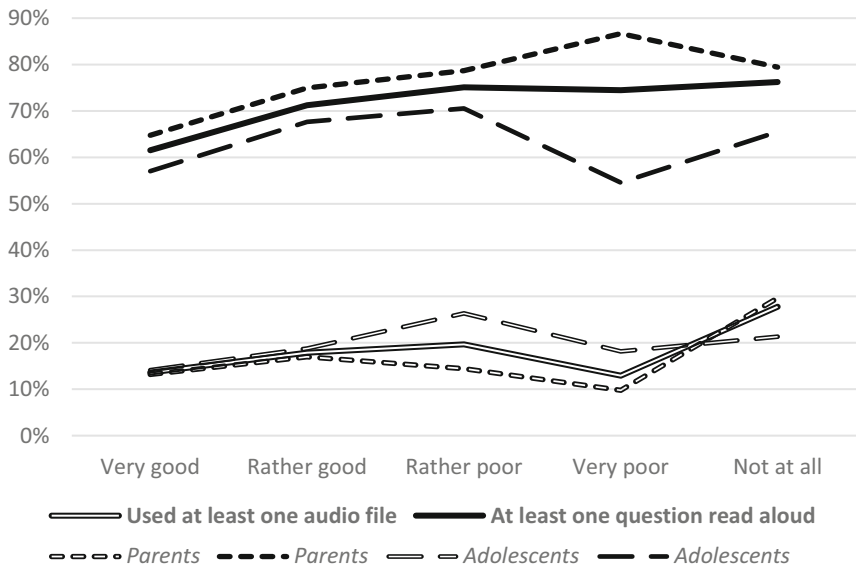


Fig. 6.3 Use of audio files and reading aloud by self-rated reading skills. (Source: ReGES data, wave 1; n = 5707 for the analysis of the audio files; n = 5665 for the analysis of whether the interviewer read out questions aloud; the lower number of cases in the second analysis was due to the fact that not all interviewers reported whether they read the questions out aloud)

the option of having an interviewer read the question aloud at least once. A comparison of adolescents and parents showed that parents used the interviewers' help of reading the questions aloud more often than adolescents. There may be different explanations for this finding. Especially in languages with many dialects, such as Arabic, the older generation often had not learned the written and standard language. Therefore, parents tended to have more problems understanding the standard language. In this case, the interviewers could explain question wording that the interviewees didn't understand in their own dialect, so the meaning of the question would be clearer. In addition, adolescents often are more familiar with using modern technologies.

Nevertheless, 84 participants with very poor or no reading skills used neither the audio files nor the opportunity to conduct the CASI questionnaire together with an interviewer. A more differentiated analysis showed that one-half of these interviewees received at least some help from bystanders present or an interviewer. The remaining cases need to be examined further.

The data indicate that, in principle, the use of audio files can be a functioning method to enable the participation of illiterate people. However, in our opinion, for a CASI survey of this length, offering audio files would not have been enough. First, the data show that the CASI duration increased enormously when an interviewee used the audio files. Model 1 in Table 6.5 shows that each question for which audio files were used prolonged the interview by 0.33 min. To clearly identify the effect of

Table 6.5 CASI duration in minutes by items using audio files, linear regression

	Model 1	Model 2
Number of items for which an audio file was used	0.33*** (0.05)	0.36*** (0.02)
Controls: (*age, starting language CASI, illiteracy*)		✓
Constant	34.84*** (0.43)	32.23*** (1.01)
R^2	0.021	0.036
N	1982	1982

Source: ReGES data, own calculations, Wave 1
***Significant at $p < 0.001$, **significant at $p < 0.01$, *significant at $p < 0.05$, +significant at $p < 0.1$.
Standard deviations in parentheses

Table 6.6 CASI duration in minutes depending on whether the interviewer read aloud some parts of the interview, linear regression

	Model 1	Model 2
Interviewer read aloud some parts	−3.31*** (0.56)	−3.49*** (0.56)
Controls: (*age, CASI starting language, illiteracy*)		✓
Constant	32.48*** (0.45)	29.72*** (1.01)
R^2	0.007	0.029
N	4794	4794

Source: ReGES data, own calculations, Wave 1
***Significant at $p < 0.001$, **significant at $p < 0.01$, *significant at $p < 0.05$, +significant at $p < 0.1$.
Standard deviations in parentheses

using audio files, Model 1 and Model 2 analyzed only the interviewees who did not use the read-aloud option. After controlling for age, the starting language of the CASI process, and reading skills, the effect of using audio files on the CASI duration increased and remained significant.

The increased CASI duration when using audio files became even clearer when analyzing only the cases in which the respondent used more than 10 audio files. For these cases, the CASI took on average 1 h to complete, almost twice as long as when audio files were not used.

In contrast to the use of audio files, CASI duration dropped by 3.31 min when an interviewer read aloud parts of the interview[9] (Table 6.6, Model 1). In this analysis, we included only persons who did not use any audio files. This effect also was significant after controlling for other variables (Model 2), which might also explain the high share of interviews in which the interviewers read aloud some questions (see Fig. 6.3). It may be possible that, to save time, the interviewers preferred to read the questions aloud. Unfortunately, we cannot test this presumption with the data.

Furthermore, during the fieldwork, interviewees evidently found it cognitively very challenging to answer such a long CASI questionnaire with just audio files, if

[9] By reading aloud parts of the interview, we mean, in all analyses, that the interviewer reported that he/she had read at least one question aloud to an interviewee. In 57.5% of the cases in which an interviewer read aloud parts of the CASI text, the interviewer reported having read aloud to the interviewee more than a quarter of all the questions.

they could not read sufficiently well. We observed that despite using the audio files, most of them still asked interviewers about their problems with comprehension. However, overall, when the interviewees used the audio files, interviewers had to read aloud fewer questions.

In the case of the ReGES study, the opportunity to ask interviewers about problems had one more practical benefit—when controlling the quality of the audio files, we found some bugs in their implementation (see Sect. 6.4.4). According to the interviewers, these problems appeared in 7.1% of the interviews. Even if this finding might not apply to other studies, surveying migrants in many different languages requires a vast quantity of audio files, and so a strong risk exists for such problems to occur. Therefore, the native-speaking interviewer who helped illiterate individuals by reading these questions aloud was also important in these cases.

Another explanation for the high percentage of interviewees who preferred to have an interviewer read the questions aloud could be that most interviewees thought the audio files were unnecessary given the interviewer's presence. Respondents might even have found it disrespectful to the interviewer to use the audio files instead of asking them for help.

However, even if most interviewees preferred to listen to the interviewer rather than using the audio files, a decisive reason for offering the latter was to minimize the interviewer's influence on response behavior and to ensure a high degree of anonymity. Therefore, we tested whether this goal had been achieved in the cases in which the audio files were used.

Since many interviewees avoided answering the CASI questionnaire on their own by asking an interviewer to read the questions aloud, we could compare whether a difference existed in the amount of item nonresponse to sensitive questions. As an indicator of sensitive questions, we took the 10 questions with the highest item nonresponse and built an index. These selected items included questions about religion, the asylum procedure, and migration history. This analysis showed that interviewees who used the opportunity to have an interviewer read aloud part of the CASI text refused fewer answers than the respondents who answered the whole CASI questionnaire on their own (Table 6.7, Model 1). After controlling for the match between the interviewer's and interviewee's language, and whether the interviewee's mother tongue was offered, the effect became slightly stronger and remained significant (Model 2). Of course, other aspects (e.g., more pressure when an interviewer read the question aloud) could explain this effect. Nevertheless, these results rebut the notion that the use of interviewers leads to higher item nonresponse.

Thus far, all these findings seem to show that no major disadvantages exist when using interviewers instead of implementing audio files as a strategy to include illiterate respondents. However, the previous analyses do not show whether the interviewers affected the interviewees' tendency to answer questions in a more socially desirable way. Thus, we checked for such an effect by reanalyzing the attitude toward gender roles, but this time, contingent on whether an interviewer read all the questions aloud or the interviewees used the audio files for these gender role questions. Model 1 in Table 6.8 shows that the interviewees who used audio files to

Table 6.7 Impact of reading aloud on item nonresponse (10 items with the highest nonresponse rate), linear regression

	Model 1	Model 2
Interviewer read some parts aloud	−0.06** (0.02)	−0.07** (0.02)
Controls: (*language match between the mother tongue of the interviewee and the language used by interviewer, mother tongue of the interviewee offered in the survey*)		✓
Constant	0.25*** (0.01)	0.33*** (0.04)
R^2	0.002	0.003
N	5224	5224

Source: ReGES data, own calculations, Wave 1
***Significant at $p < 0.001$, **significant at $p < 0.01$, *significant at $p < 0.05$, +significant at $p < 0.1$.
Standard deviations in parentheses

Table 6.8 Impact of audio file use versus reading questions aloud on gender roles, linear regression

	Model 1	Model 2
Interviewee used audio files for all questions on gender roles (vs. interviewer reading aloud all items)	0.30+ (0.16)	0.27+ (0.16)
Controls: (*country of origin, education, age, language match with interviewer*)		✓
Constant	1.96*** (0.03)	2.12*** (0.12)
R^2	0.005	0.023
N	766	766

Source: ReGES data, own calculations, Wave 1
***Significant at $p < 0.001$, **significant at $p < 0.01$, *significant at $p < 0.05$, +significant at $p < 0.1$.
Standard deviations in parentheses

answer these questions tended to answer the questions about gender roles more conservatively than those who asked the interviewer to read all the questions aloud (including those about gender roles). Even though this effect was less significant, it was much stronger (see Table 6.8) than the effect of the influence of language on the response behavior, which we analyzed in Sect. 6.4.5.2 (see Table 6.3). This effect retained its significance after controlling for the country of origin, education, age, and the match between the interviewer's and interviewee's language. This result can be interpreted as an indication that using audio files provides a higher anonymity, so the interviewee is less susceptible to a social desirability effect.

These results indicate that it may be worth using CASI for sensitive content, even when this means having to implement additional audio files to include illiterate people.

6.5 Discussion

Surveys in the social sciences should ensure that illiterate individuals are not excluded from participating in these studies. This is even more true when studies focus on education. In this case, it would be devastating to lose the undereducated in a systematic way. The present study examined two options, with their advantages and disadvantages, for including illiterate migrants in social science surveys: personal interviewing by interviewers and the use of computer-assisted self-interviewing with additional audio files. Based on data from the ReGES study, practical experience showed the extent to which these two strategies work when interviewing newly immigrated refugees. In addition to the conventional challenges associated with surveying illiterate people, we also discussed issues concerning the integration of different languages of origin in the two interview modes.

The main results of our study are as follows. In many cases, the possibility that interviewers could help with technical and comprehension difficulties appeared to be an important prerequisite for interviewees to successfully complete the computer-assisted self-interview. Furthermore, the overall usage of audio files was less than expected, and none of our respondents conducted the CASI survey solely on the basis of using the audio files. The use of these files increased the survey duration, which seemed to be an issue for both interviewers and interviewees. This finding might help to explain why respondents preferred to have an interviewer read the questions aloud to them, as in a face-to-face interview. Our analyses of answers to sensitive questions suggest that interviewees in a face-to-face mode respond even more frequently than when using CASI, but also tend to give slightly more socially desirable responses.

Regarding the technical implementation of the audio files, we have to state that despite intensive testing of the implemented audio files, technical problems could not be ruled out. This may be partly due to the fact that the questionnaire was not primarily designed to interview illiterate people; therefore, certain features (e.g., short questionnaires, no complex content) that were helpful in the design of CASI with audio files were neglected.

We can offer the following recommendations for researchers who might want to interview newly arrived refugees and want to include illiterate people as well. In our experience, no alternative exists to using native-speaking interviewers to recruit participants in the first wave of a survey with immigrants. Interviewers who speak the mother tongue of respondents are essential with regard to contacting and motivating newly arrived refugees, including those who are illiterate, and to ensuring that they are able to complete their interview successfully. Nevertheless, the known problems of social desirability effects while carrying out face-to-face interviews must be taken into account. It can be assumed that using CASI with audio files is especially helpful when investigating sensitive topics because it seems to reduce social desirability. However, it cannot be determined assuredly whether all illiterate individuals could successfully complete a CASI survey without the support of an interviewer.

Thus, if a high proportion of illiterate people are suspected in a target group, a personal interview should always be preferred. CASI with additional audio files should only be carried out if sensitive content is recorded, and the interview should be as short and simple as possible. The selected software should be able to implement additional features (e.g., so the currently read text is highlighted, and symbols are used instead of numbers) to further support illiterate interviewees. Moreover, it is extremely important that CASI with additional audio files is cognitively pretested with illiterate people to ensure that the procedure is not too demanding for the target group. In addition, the implementation of the audio files in different foreign languages should be checked carefully, and time for any corrections should be scheduled.

Even if the interview by native-speaking interviewers is the preferred variant, the CASI with audio files enables a multilingual, standardized survey of people with reading difficulties, even without native-speaking interviewers. For logistical or financial reasons, it is not always possible to employ enough interviewers in all the required languages. The use of audio files is particularly worthwhile the more difficult it is to recruit interviewees for a language, the more interviews are to be carried out, the larger the spatial area that is to be covered by fieldwork, and also in cases in which the interview is to be conducted in languages that do not have a written form.

In view of the methodological limitations mentioned in the present study, all discussed questions need to be reinvestigated, ideally in a methodological experiment. Moreover, the effect of interviewers on the response behavior of refugees should be investigated in more detail: on the one hand, with regard to interviewer and respondent characteristics, as well as the characteristics of the audio files voice; and on the other hand, with regard to the different contents of questionnaires.

Finally, our results clearly show that the presence of interviewers is very important (at least for this specific target group). Therefore, it seems worthwhile to perform further research not only on the implementation and use of audio files, but also on how to minimize interviewer effects.

References

Allerbeck, K. R., & Hoag, W. J. (1985). Wenn Deutsche Ausländer befragen. Ein Bericht über methodische Probleme und praktische Erfahrungen. *Zeitschrift für Soziologie, 14*(3), 241–246.

Aquilino, W. S., Wright, D. L., & Supple, A. J. (2000). Response effects due to Bystander presence in CASI and paper-and-pencil surveys of drug use and alcohol use. *Substance Use & Misuse, 35*, 845–867.

Baykara-Krumme, H. (2010). *Interviewereffekte in Bevölkerungsumfragen: ein Beitrag zur Erklärung des Teilnahme- und Antwortverhaltens von Migranten* (Band 19). pairfam—Das Beziehungs- und Familienpanel.

Baykara-Krumme, H. (2012). Sind bilinguale Interviewer erfolgreicher? Interviewereffekte in Migrantenbefragungen. In H. G. Soeffner (Ed.), *Transnationale Vergesellschaftungen* (pp. 259–273). Springer Fachmedien.

Behr, D. (this volume). Chapter 3: Computer-assisted migration research: What can we learn about source questionnaire design and translation from the software localization field. In S. Pötzschke & S. Rinken (Eds.), *Migration research in a digitized world: Using innovative technology to tackle methodological challenges*. Springer.

Beier H., & Schulz S. (2015). A free audio-CASI module for LimeSurvey. *Survey Methods: Insights from the Field.* http://surveyinsights.org/?p=5889. Accessed 14 Jan 2020.

Beier, H., Schulz, S., & Kroneberg, C. (2014). *Freundschaft und Gewalt im Jugendalter: Feldbericht der ersten Erhebungswelle (Technical Report)* (Working Papers Nr. 158). Mannheimer Zentrum für Europäische Sozialforschung.

Blohm, M., & Diehl, C. (2001). Wenn Migranten Migranten befragen: Zum Teilnahmeverhalten von Einwanderern bei Bevölkerungsbefragungen. *Zeitschrift für Soziologie, 30*(3), 223–242.

Chadi, A. (2013). Third person effects in interview responses on life satisfaction. *Schmollers Jahrbuch, 133*(2), 323–333.

Cooley, P. C., Rogers, S. M., Turner, C. F., Al-Tayyib, A. A., Willis, G., & Ganapathi, L. (2001). Using touch screen audio-CASI to obtain data on sensitive topics. *Computers in Human Behavior, 17*(2001), 285–293.

Couper, M., Singer, E., & Tourangeau, R. (2002). *Social desirability effects on self-reports of behavior: Understanding the effects of audio-CASI*. https://pdfs.semanticscholar.org/2a05/b818518ec4eb025b6585caea0d458382086d.pdf. Accessed 06 Jan 2020.

Couper, M. P., Singer, E., & Tourangeau, R. (2003). Understanding the effects of audio-CASI on self-reports of sensitive behavior. *The Public Opinion Quarterly, 67*(3), 385–395.

Couper, M. P., Tourangeau, R., & Marvin, T. (2009). Taking the audio out of audio-CASI. *Public Opinion Quarterly, 73*(2), 281–303.

Dotinga, A., van den Eijnden, R. J. J. M., Bosveld, W., & Garretsen, H. F. L. (2005). The effect of data collection mode and ethnicity of interviewer on response rates and self-reported alcohol use among Turks and Moroccans in the Netherlands: An experimental study. *Alcohol and Alcoholism, 40*(3), 242–248.

Edwards, S. L., Slattery, M. L., Murtaugh, M. A., Edwards, R. L., Bryner, J., Pearson, M., Rogers, A., Edwards, A. M., & Tom-Orme, L. (2007). Development and use of touch-screen audio computer-assisted self-interviewing in a study of American Indians. *American Journal of Epidemiology, 165*(11), 1336–1342.

Falb, K., Tanner, S., Asghar, K., Souidi, S., Mierzwa, S., Assazenew, A., Bakomere, T., Mallinga, P., Robinette, K., Tibebu, W., & Stark, L. (2016). Implementation of Audio-Computer Assisted Self-Interview (ACASI) among adolescent girls in humanitarian settings: Feasibility, acceptability, and lessons learned. *Conflict and Health, 10*(32).

Feskens, R., Hox, J., Lensvelt-Mulders, G., & Schmeets, H. (2006). Collecting data among ethnic minorities in an international perspective. *Field Methods, 18*(3), 284–304.

Gatward, R. (2002). Interviewing children using audio-CASI. *Social Survey Methodology Bulletin, 50*, 16–26.

Gentile, R., Heinritz, F., & Will, G. (2019). *Übersetzung von Instrumenten für die Befragung von Neuzugewanderten und Implementation einer audiobasierten Interviewdurchführung* (LIfBi Working Paper No. 86). Leibniz-Institut für Bildungsverläufe.

Glantz, A., & Michael, T. (2014). Interviewereffekte. In N. Baur & J. Blasius (Eds.), *Handbuch Methoden der empirischen Sozialforschung* (pp. 313–322). Springer Fachmedien.

Grotlüschen, A., & Riekmann, W. (2012). *Funktionaler Analphabetismus in Deutschland.—Ergebnisse der ersten leo.—Level-One Studie. Waxmann Verlag, Münster*. https://leo.blogs.uni-hamburg.de/wp-content/uploads/2014/01/9783830927754-openaccess.pdf. Accessed 26 Apr 2022.

Groves, R. M., & Couper, M. (1998). *Nonresponse in household interview surveys*. Wiley.
Groves, R. M., Cialdini, R. B., & Couper, M. P. (1992). Understanding the decision to participate in a survey. *Public Opinion Quarterly, 56*, 475–495.
Groves, R. M., Fowler, F. J., Couper, M., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2009). *Survey methodology*. Wiley.
Haarmann, H. (1983). Kriterien der ethnischen Identität. *Language Problems and Language Planning, 7*(1), 21–42.
Haug, S., Lochner, S., & Huber, D. (2017). Methodische Herausforderungen der quantitativen und qualitativen Datenerhebung bei Geflüchteten. In S. Lessenich (Ed.), *Geschlossene Gesellschaften. Verhandlungen des 38. Kongresses der Deutschen Gesellschaft für Soziologie in Bamberg 2016*.
Heckmann, F. (1992). *Ethnische Minderheiten, Volk und Nation: Soziologie inter-ethnischer Beziehungen*. F. Enke.
Hewett, P., Mensch, B., & Erlukar, A. (2004). Consistency in the reporting of sexual behaviour by adolescent girls in Kenya: A comparison of interviewing methods. *Sexually Transmitted Infections, 80*(2), ii43-ii48.
Hox, J., & de Leeuw, E. (2002). The influence of interviewers' attitude and behavior on household survey nonresponse: An international comparison. In R. M. Groves, D. A. Dillman, J. L. Eltinge, & R. J. A. Little (Eds.), *Survey nonresponse* (pp. 103–120). Wiley.
Jacobsen, J. (2018). Language barriers during the fieldwork of the IAB-BAMF of refugees in Germany. In D. Behr (Ed.), *Surveying the migrant population: Consideration linguistic and cultural issues* (pp. 75–84). GESIS-Schriftenreihe (19).
Kappelhof, J. W. S. (2014). The impact of method bias on the cross-cultural comparability in face-to-face surveys among ethnic minorities. *Methods, Data, Analyses, 8*(1), 79–118.
Kühne, S., Jacobsen, J., & Kroh, M. (2019). *Sampling in times of high immigration: The survey process of the IAB-BAMF-SOEP survey of refugees. Survey methods: Insights from the field.* https://surveyinsights.org/?p=11416. Accessed 14 Jan 2020.
Le, C. L., Blum, R. W., Magnani, R., Hewett, P. C., & Do, H. M. (2006). A pilot of audio computer-assisted self-interview for youth reproductive health research in Vietnam. *Journal of Adolescent Health, 38*(6), 740–747.
Lessler, J. T., Caspar, R. A., Penne, M. A., & Barker, P. R. (2000). Developing computer assisted interviewing (CAI) for the National Household Survey on Drug Abuse. *Journal of Drug Issues, 30*(1), 9–34.
Loosveldt, G. (2008). Face-to-face interviews. In E. de Leeuw, J. J. Hox, & D. A. Dillman (Eds.), *International handbook of survey methodology* (European Association of Methodology Series, pp. 201–220). Routledge.
Mierzwa, S., Souidi, S., Friedland, I., Katzen, L., & Littlefield, S. (2013). Effective approaches to user interface design with ACASI in the developing world. *Interactions, 20*(3), 58–61.
Morina, N., Ewers, S., Passardi, S., Schnyder, U., Knaevelsrud, C., Müller, J., Bryant, R., Nickerson, A., & Schick, M. (2017). Mental health assessments in refugees and asylum seekers. *Conflict and Health, 11*, 18.
Nass, C., Moon, Y., & Green, N. (1997). Are machines gender neutral? Gender-stereotypic responses to computers with voices. *Journal of Applied Social Psychology, 27*(10), 864–876.
Nass, C., Robles, E., Bienenstock, H., Treinen, M., & Heenan, C. (2003). Speech-based disclosure systems: Effects of modality, gender of prompt, and gender of user. *International Journal of Speech Technology, 6*(2), 113–121.
Neske, M., & Rich, A.-K. (2016). Asylerstantragssteller in Deutschland im ersten Halbjahr 2016: Sozialstruktur, Qualifikationsniveau und Berufstätigkeit. *BAMF-Kurzanalyse, 4*.
Newman, J. C., Des Jarlais, D. C., Turner, C. F., Gribble, J., Cooley, P., & Paone, D. (2002). The differential effects of face-to-face and computer interview modes. *American Journal of Public Health, 92*(2), 294–297.

Nienkemper, B. (2015). *Lernstandsdiagnostik bei funktionalem Analphabetismus. Akzeptanz und Handlungsstrategien.* Bertelsmann. https://www.die-bonn.de/doks/2015-analphabetismus-01.pdf. Accessed 14 Jan 2020.

Rich, A.-K. (2016). Asylerstantragsteller in Deutschland im Jahr 2015 Sozialstruktur, Qualifikationsniveau und Berufstätigkeit. In *Kurzanalysen des Forschungszentrums Migration, Integration und Asyl des Bundesamtes für Migration und Flüchtlinge*. BAMF.

Schroder, K. E. E., Carey, M. P., & Vanable, P. A. (2003). Methodological challenges in research on sexual risk behavior: II. Accuracy of self-reports. *Annals of Behavioral Medicine, 26*, 104–123.

Siegel, J. S. (2018). *Demographic and socioeconomic basis of ethnolinguistics*. Springer.

Spörlein, C., Kristen, C., Schmidt, R., & Welker, J. (2020). Selectivity profiles of recently arrived refugees and labor migrants in Germany. *Soziale Welt, 71*(1–2), 54–89.

Steinhauer, H. W., Zinn, S., & Will, G. (2019). Sampling refugees for an educational longitudinal survey. *Survey Methods: Insights from the Field*. https://surveyinsights.org/?p=10741. Accessed 27 June 2021.

Tourangeau, R., & Smith, T. W. (1996). Asking sensitive questions: The impact of data collection mode, question format, and question context. *Public Opinion Quarterly, 60*(2), 275–304.

Tourangeau, R., & Smith, T. W. (1998). Collecting sensitive information with different modes of data collection. In M. P. Couper, R. P. Baker, J. Bethlehem, C. Z. F. Clark, J. Martin, W. L. Nicholls II, & J. M. O'Reilly (Eds.), *Computer assisted survey information collection* (pp. 431–453). Wiley.

Turner, C. F., Danella, R. F., & Rogers, S. M. (1995). Sexual behavior in the United States: 1930–1990: Trends and methodological problems. *Sexually Transmitted Diseases, 22*, 173–190.

Turner, C., Rogers, S., Hendershot, T., Miller, H., & Thornberry, J. (1996). Improving representation of linguistic minorities in health surveys. *Public Health Reports, 111*(3), 276–279.

Turner, C. F., Forsyth, B. H., O'Reilly, J. M., Cooley, P. C., Smith, T. K., Rogers, S. M., & Miller, H. G. (1998a). Automated self-interviewing and the survey measurement of sensitive behaviors. In M. P. Couper, R. P. Baker, J. Bethlehem, C. Z. F. Clark, J. Martin, W. L. Nicholls II, & J. M. O'Reilly (Eds.), *Computer assisted survey information collection* (pp. 455–473). Wiley.

Turner, C. F., Ku, L., Rogers, S. M., Lindberg, L. D., Pleck, J. H., & Sonenstein, F. L. (1998b). Adolescent sexual behavior, drug use and violence: Increased reporting with computer survey technology. *Science, 280*(May 8), 867–873.

UNESCO. (2006). Education for all global monitoring report 2006: Literacy for life. .

van de Wijgert, J., Padian, N., Shiboski, S., & Turner, C. (2000). Is audio-assisted self-interviewing a feasible method for surveying in Zimbabwe? *International Journal of Epidemiology, 29*(5), 885–890.

van Heelsum, A. (2013). The influence of interviewers' ethnic background in a survey among Surinamese in the Netherlands. In J. Font & M. Méndez (Eds.), *Surveying ethnic minorities and immigrant populations: Methodological challenges and research strategies* (pp. 111–130). Amsterdam University Press.

West, B. T., & Blom, A. G. (2017). Explaining interviewer effects: A research synthesis. *Journal of Survey Statistics and Methodology, 5*(2), 175–211.

Will, G., Homuth, C., von Maurice, J., & Roßbach, H.-G. (2021). Integration of recently arrived underage refugees: Research potential of the study ReGES—Refugees in the German Educational System. *European Sociological Review*. https://doi.org/10.1093/esr/jcab033

Wong, F., Huang, Z., Thompson, E., De Leon, J., Shah, M., & Park, R. (2007). Substance use among a sample of foreign- and U.S.-born Southeast Asians in an urban setting. *Journal of Ethnicity in Substance Abuse, 6*(1), 45–66.

Part II
New Data Sources and Their Potential

Chapter 7
Leveraging the Web for Migration Studies: Data Sources and Data Extraction

Sebastian Rinken and José Luis Ortega

7.1 Introduction

In recent years, the amount of information accessible through the World Wide Web (in the following "the Web") has been growing exponentially. More and more people across the globe are turning to the Web to obtain the information they require for personal, social, or professional reasons. Thus, access to information via the Web has become a prerequisite for conducting many kinds of transactions, and it seems only slightly exaggerated to state that "everything is on the Web."

However, the Web is not a structured medium in which extant information is labelled or classified for easy retrieval. It is not a database in which each piece of data is set in a field and then defined with attributes and properties. Instead, the Web is a heterogeneous assortment of webpages that present manifold kinds of information in very diverse ways. Its enormous wealth of content requires navigation by means of hyperlinks. This digital environment gives precedence to style and appearance over the structure of data, and often displays content without context or hierarchy. Nevertheless, the inherent decontextualization of the Web is partially remedied by its increasing use as a dissemination channel for data that are suitable for research purposes.

The concept of *open data* refers to the fact that many organizations are making a vast range of big and complex datasets freely available with a view to spurring the development of new applications and services, as well as to fostering research on

S. Rinken (✉)
Spanish Research Council (CSIC), Institute for Advanced Social Studies (IESA), Córdoba, Spain
e-mail: srinken@iesa.csic.es

J. L. Ortega
Institute for Advanced Social Studies (IESA-CSIC) & Joint Research Unit Knowledge Transfer and Innovation, University of Córdoba and IESA-CSIC, Córdoba, Spain
e-mail: jortega@iesa.csic.es

S. Pötzschke, S. Rinken (eds.), *Migration Research in a Digitized World*, IMISCOE Research Series, https://doi.org/10.1007/978-3-031-01319-5_7

globally-pressing issues such as climate change, mobility patterns, etc. As a side-effect of this unprecedented possibility of sharing and reusing data, the scholarly community is demanding more transparency in scientific research and a higher degree of availability of research data that permit the checking, and eventually the reproduction, of scientific advances more easily. Ultimately, these developments have contributed to a more general call for "open science" (Bartling & Friesike, 2014).

Access to this enormous volume of data requires specific skills that enable the location, extraction, processing, and analyzing of large datasets of different typologies from different sources. Specific data management capacities are necessary for putting structured datasets to use for research purposes and for initiating new research approaches to unstructured data, for example, the data embedded in websites. To highlight two crucial examples, *web scraping* is a data extraction technique that harvests plain text from patterns in webpage source code for storage in relational databases, whereas Application Programming Interfaces (APIs) enable the development of scripts that can automatically retrieve data from a website.

Such data management competencies are especially beneficial for migration studies, a field of research that deals with complex, large-scale patterns of human mobility and interaction, which oftentimes elude timely, encompassing, and rigorous analysis by the traditional instruments of empirical social research (Dodge, 2019; Giannotti et al., 2016; Sirbu et al., 2020). The varied assortment of administrative records, censuses, and surveys that has dominated the field for decades does not guarantee the availability of reliable, complete, comparable, timely, and disaggregated data on migrant stocks and flows, let alone the patterns of interactions with autochthonous populations and the manifold processes of economic, social, and cultural integration. Notorious issues include diverging definitions of key concepts, insufficient periodicity of sources, insufficient size of (sub-)samples, coverage and/or participation biases, and event-based administrative records that impede cross-referencing, among others (Kupiszewska et al., 2010; Font & Méndez, 2013).

Such difficulties and shortcomings suggest that an enormous potential exists for innovative, large-scale datasets to generate relevant knowledge in the migration studies field. To mention some examples, mobile phones' Call Detail Records (CDRs) can track seasonal mobility between and within countries (Ahas et al., 2007; Deville et al., 2014) as well as post-disaster and war-induced displacements (Wilson et al., 2016; Salah et al., 2019); search engine terms can contribute to forecasting migration trends (Wladyka, 2017; Lin et al., 2019); social network messages can help with the analysis of the perceptions of, and opinions about, foreign-born populations (Greco et al., 2017); and geo-located social media data can provide information about personal (Twitter, Facebook) or professional movements (LinkedIn) (Barslund & Busse, 2016; Spyratos et al., 2019; State et al., 2014; Zagheni et al., 2014). Often, the insights emerging from such unconventional data sources can be validated with traditional sources, albeit typically with minor granularity and timeliness.

These examples illustrate how the data revolution is being seized by relevant parts of the migration research community. However, many migration scholars lack the

necessary skills for locating, extracting, and managing the large volumes of raw data available on the Web. This chapter provides a glimpse into the new universe of the opportunities granted migration researchers by the exponential increase of web-based data on human mobility, behavior patterns, relationships, preferences, and opinions. While obviously not a substitute for proper data-management training, this chapter seeks to encourage migration researchers to build the necessary skills, individually or at a team level. The chapter has been co-authored by a data scientist who is a novice regarding migration studies (JLO) and a migration scholar who is a novice regarding data science (SR). We hope that the result of this trans-disciplinary co-authorship, a process that taught the two of us a lot, proves useful to the migration research community.[1]

We begin by introducing several key concepts regarding the extraction and management of web information ("Key concepts" Section). Next, we describe selected services and repositories specializing in demography and migration studies that supply open data for research projects, as well as more generalist data sources ("Data sources" Section). The "Data extraction" Section presents various data extraction techniques that enable the obtaining of data from structured and unstructured Web sources, transform textual information in tabulated data, and carry out queries of linked data repositories. We conclude with an outlook on the promise the Web offers migration scholars, as well as related challenges.

7.2 Key Concepts

In this section, in addition to the primordial concept of *big data*, we also introduce the related concepts of *open data* and *linked data*.

7.2.1 Big Data

The term *big data*, which began to emerge at the turn of the twenty-first century, reflects the exponential growth of digital data due to the increasingly ubiquitous deployment of data-sensor equipment such as mobile devices, the generation of ever more copious metadata such as software logs or internet clicks and, more recently, the inter-connection of all kinds of gadgets, appliances, and procedures ("Internet of Things"). Rather than refer to any particular size of datasets, *big data* refers to amounts of data so massive that they transcend the capacity of customary processing facilities and techniques. Thus, in terms of actual data volume, the notion of what counts as "big" is highly dynamic, given that computing systems also have evolved

[1] We wish to acknowledge Steffen Pötzschke's insightful comments on draft versions of this chapter.

at a notoriously fast pace over these past decades, albeit with an increasing difficulty of living up to "Moore's Law" that processing power will double every 18 months or so. Therefore, any meaningful definition of the term has to refer to advanced techniques of data management, processing, and analysis (Mayer-Schönberger & Cukier, 2013), rather than data volume *per se*.

In addition to the astounding growth of data generation and data processing capacities, big data has been favored decisively by the unprecedented interoperability provided by the Internet. Access to a multiple and varied range of data types through the Web enables the integration and analysis of data dispersed across diverse sources. For example, to get to know their clients, a bank does not rely only on monitoring their account activity, but also can combine this information with streams of data relative to their preferences in social networks, shopping patterns, localization, etc. Apart from their innumerable commercial applications, such data offer new analytical possibilities for understanding all kinds of social behavior and phenomena, such as transportation needs, climate change, and the risk of exposure to pathogens, among many others.

The initial characterizations of the *big data* concept (Laney, 2001) used to rely on the "three Vs", i.e., volume (huge), velocity (near "real time"), and variety (comprising non-structured data with temporary and spatial references). Additional traits of fundamental importance include exhaustivity (data captured from entire populations rather than samples), flexibility (favoring scalability and the addition of new sources), high resolution (descending to fine-grained detail to permit deep analysis), and relationality (defining universal IDs that enable cross-reference data from different sources). The focal point of all these characteristics is flagged by two further "Vs": veracity and value, i.e., the overall quality of the data and hence its utility.

7.2.2 Open Data

Digital data are collected for all kinds of purposes by an ever-increasing range of organizations and institutions—commercial, governmental, and not-for-profit. Depending on the nature of these data and the objectives pursued in gathering them, they may or may not be made publicly available by default. For obvious reasons, data entailing information on any particular individual will not be eligible for public release on privacy grounds, nor will the data that affect national security. Private enterprises often prefer to restrict access to data that may confer a competitive advantage, unless potential higher-order gains can be obtained by granting free access.

Similar to the movements in the domains of software (Open Source) or scientific publishing (Open Access), the Open Data movement demands that data be freely available and re-usable without prior permission (Charalabidis et al., 2018). This demand refers primarily to the data produced by public administrations and governmental organizations, due to the absence of commercial interests and the intrinsic

value of transparency in democratic systems of government. Since the benefits of publishing specific kinds of data may be obvious (as is the case with epidemiological information or the spending details of public budgets) or largely unpredictable (for example, when information on specific resources is converted into mobile apps targeting people with special needs), the baseline demand of the Open Data movement is that all non-personal data be freely accessible.

The Open Data concept is usually associated with three interrelated characteristics, the first of which regards the full availability of, and access to, complete datasets (rather than samples) preferably via download from the Internet. The idea is to make the data available free of charge, so any fees must be limited and well justified. Second, the data should be re-usable and redistributable both from a legal viewpoint (cf. user agreement) and in terms of technical specifications; ideally, this reuse and redistribution should include the possibility of incorporation in other datasets. Third, the notion of universal participation alludes to the absence of restrictions regarding the kinds of data (re-)users and their fields of endeavor, be they commercial, educational, scientific, or otherwise.

The combination of these traits basically amounts to conceiving digital data as a public good. To unfold their virtues, the data need to be ready for processing by different hard- and soft-ware environments (interoperability) and also favor aggregation with other data (*linked data*).

7.2.3 Linked Data

The third crucial concept concerning data management, *linked data*, refers to the possibility of interconnecting and sharing data between different sources in an open and transparent way (Heath & Bizer, 2011). This concept was first coined by the British computer scientist Tim Berners-Lee—a decisive contributor to the creation of the Internet—to allude to a World Wide Web Consortium (W3C) project about developing a technology for linking data from different sources (Berners-Lee, 2006). Faced with the problem of different web sites offering information about organizations, events, objects, and persons in ways that could not be easily located mutually, W3C proposed to establish a huge and dynamic data network that could be navigated with hyperlinks. Since then, the successful implementation of that proposal has enabled data repositories across the globe to group and connect information (for an illustrative example, see https://lod-cloud.net/clouds/geography-lod.svg, a network graph that shows the connections between data repositories).

Linked data technology defines each object through a uniform resource identifier (URI), a unique tag that, similar to a web address, enables the identification of an object in a data network. For example, the URI http://www.wikidata.org/entity/Q30 corresponds to the entity *United States* in Wikidata. While Web content is written in hypertext markup language (HTML), the linked-data Web operates according to specific protocols of its own for publishing and querying the data. We will postpone an introduction to querying to a later section but briefly introduce the resource

description framework (RDF), a convention for publishing machine-readable data and linking them to other datasets (Curé & Blin, 2014).

RDF is not a computing language, but rather a framework in which other languages are written (Carroll & Stickler, 2004) by a process called *serialization*, which is a scheme for defining sets of basic information about the nature of the relation of an entity to other entities. The RDF structure is based on triples that include three fundamental elements: subject, predicate, and object. For example, the triple

<http://www.wikidata.org/entity/Q12418><http://purl.org/dc/terms/creator>
<http://dbpedia.org/resource/Leonardo_da_Vinci>.

expresses that entity Q12418 "Mona Lisa" has a relationship of the type creator with the resource "Leonardo da Vinci."

The RDF data model epitomizes the mutual advantages of cooperation via common and open standards (Bergman, 2009). Since this model enables linking data from different platforms, it improves their interoperability; for example, the above triplet connects objects deposited in three different locations. Also, the RDF syntax is unambiguous, and hence efficient, given that only one definition of each entity or relationship is provided. A third advantage of the RDF data model is scalability: there is no limitation regarding the volume of data that can be stored as linked data, or the number of repositories that can be interconnected in various ways.

7.3 Data Sources

This section reviews and describes some outstanding data sources for migration studies. Rather than personal preferences, our selection reflects objective criteria such as the volume of web traffic generated by these platforms and the data's use in the scientific literature. However, it seems prudent to remark that these and other platforms' data portfolios are subject to more or less continuous innovations and improvements. Also, we do not wish to suggest that the platforms presented here are the only ones useful for migration scholars, since our selection is just a glimpse of the wealth of the extant web resources, rather than an exhaustive compilation.

We distinguish between four categories of sources, which require different degrees of data management skills (ranging from basic to advanced) to make them useful for substantive research. First, the subsection "Specialized sites" presents specific portals that compile information on migration and related phenomena. These portals have the advantage of providing contextual analyses and graphs that make the data more understandable. However, in some cases, the data is not very fine-grained, heterogeneous, and of diverse origin, which may decrease their rigor and reliability. By contrast, the "Generalist data portals" subsection provides data on a wide range of topics, including but not limited to migration-related facts and figures. The offerings of inter-governmental organizations such as the European Union, OECD, and the UN feature prominently in this category. Disaggregation usually is

provided at the country and/or regional level, in line with each institution's membership. Third, "data repositories" are platforms for distributing research data, mostly survey files uploaded by researchers to guarantee transparency, quality control by peers, and reproducibility, as well as to facilitate re-utilization for additional studies. Such data are appreciated because they are original, disaggregated to high geographical detail (counties, municipalities, etc.), and often dedicated to aspects (habits, beliefs, etc.) not easily found in official sources. However, these data sometimes are limited to small samples, narrow purposes, and incompatible formats. Fourth and finally, we also draw attention to a dedicated dataset search engine.

7.3.1 Specialized Sites

The three specialized migration-data platforms we describe here are characterized by providing open access, without registration, to their data in an interchangeable format. The platform World Pop facilitates geospatial information on demographic issues in low- and middle-income countries across three continents; the Migration Data Portal provides a starting point for understanding human movements at a global scale; and the Migration Policy Institute's Migration Data Hub provides a general outlook on immigration to the United States.

World Pop (http://www.worldpop.org/) is the leading open-access platform for spatial demography, providing finely granulated geospatial datasets on population growth, distribution, and characteristics in Central and South America, Africa, and Asia. Launched in 2013 as a platform integrating various specialized portals, World Pop basically pursues the mission of making top-of-the-line geospatial mapping available regarding the global South where such data would otherwise be missing or insufficient. The ultimate aim of this portal is to foster scientific research and better-informed policy interventions regarding economic development, ecological sustainability, health-care, and other issues. Its high-resolution spatial distributions, rigorous methodologies, and open-source documentation have been recognized as vital input for development projects. The provision of up-to-date information is an important goal, yet achieving it is occasionally hindered by time-lagged source data.

The World Pop portal integrates a wide range of sources, including censuses, surveys, satellite data, administrative statistics, mobile phone data, and others to produce fine-gridded maps of population distributions. It employs advanced data management techniques such as machine learning to disaggregate information regarding administrative units of varying and often excessive size to grid units of just 100 × 100 m. Building on the static snapshots of population distributions and characteristics at certain time points, the portal also elaborates high-resolution maps of population dynamics. Its current data line-up covers 11 substantive areas, which include spatial population distributions by continent and country, internal migration, global settlement growth, and a host of development indicators.

World Pop's estimation and imputation procedures—documented extensively and made available as metadata—are in constant flux to seize on evolving

technological options. World Pop's data can be accessed in two ways: through a dedicated API application (see Sect. 7.4.1) or directly by discharging datasets from the website.

The **Migration Data Portal** (http://migrationdataportal.org/) caters to a broad audience (including policy makers, statistics officers, journalists, and the general public) with a view to showcasing how migration policy is evidence-based and to contributing to more facts-driven public debate on migration and its manifold effects. The portal was established with backing from the German government in the wake of the 2015 surge in refugee flows, and it is hosted at the Global Migration Data Analysis Centre (GMDAC), a Berlin-based research outfit belonging to the International Organization for Migration (IOM). It provides international migration data obtained from a range of sources with the stated aim to make international migration data and information more accessible, visible, and easier to understand.

Such emphasis on user-friendliness translates into a penchant for infographics, data sheets, and clickable maps with links to definitions and additional information. The portal's "Data" Section offers dynamic access to dozens of indicators pertaining to a vast variety of thematic groupings such as stock and flow statistics, integration processes, and public opinion, among others. When selecting any particular indicator and geographic area, related values such as the highest- and lowest-scoring countries are displayed, and a timeline ranging from 2000 through to the present adapts to the periodicity of the respective source data. The "Themes" and "Resources" Sections provide various kinds of background information on measurement, data sources, context, and analysis of migration data. The portal also offers data regarding the United Nation's Sustainable Development Goals.

While specialized scholars will find fault with some of the portal's details, anyone aware of the challenges of integrating data from so many diverse sources and diverse range of aspects cannot but admire the portal's accomplishments. Researchers may especially savor offerings such as a searchable database of innovations in data migration statistics.

Our third platform entry, the Migration Policy Institute (MPI), is a Washington-based think tank that aims to foster liberal ("pragmatic") migration policies. Focusing mainly on North America and Europe, with special emphasis on the United States, the MPI seeks to engage policy-makers, economic stakeholders, the media, and the general public. It conducts research on migration management and integration policies, and strives to make its publications accessible to non-specialist audiences. The MPI website includes a **Migration Data Hub** (https://www.migrationpolicy.org/programs/migration-data-hub) that supplies tables, graphs, and maps with recent and historical data on migration flows and stocks, residence status, integration markers, economic performance, employment, and remittances. For the international migration research community, this portal's data offerings are relevant mostly when fine-grained information on the US is needed, which can be provided at the state or even county level. Regarding a range of demographic, economic, and integration indicators, data output can be customized for user-defined comparisons between states.

7.3.2 Generalist Data Portals

In addition to specialist websites, an enormous wealth of migration-related information can be obtained from generalist data portals. The world's primary international or inter-governmental organizations predominate in this category, thanks largely to their capacity to leverage the vast statistical input provided by their respective member states. The flip-side of this advantage is that each organization's membership tends to define the geographical coverage of its data portfolio.

The **European Union** has merged all its data offerings in one platform (http://data.europa.eu) that provides both metadata from public sector portals throughout Europe at any geographical level (from international to local) and datasets collected and published by European institutions, prominently including Eurostat but also many other EU agencies and organizations. This portal merits extensive exploration, since it constitutes a veritable dataset library of tens of thousands of entries. Downloads are facilitated in a vast range of formats.

The Organization for Economic Co-operation and Development (OECD) maintains three databases relevant for migration studies (https://www.oecd.org/migration/mig/oecdmigrationdatabases.htm). Its **International Migration Database** provides recent data and historic series of migration flows and stocks of foreign-born people and foreign nationals in OECD countries as well as data on acquisitions of nationality. The **Database on Immigrants in OECD Countries (DIOC)** includes comparative information on demographic and labor market characteristics of immigrants living in OECD countries and a number of non-OECD countries (DIOC extended or DIOC-E). Finally, **Indicators of Immigrant Integration** are gathered on employment, education and skills, social inclusion, civic engagement, and social cohesion. Data are displayed in user-defined tables and charts and can be downloaded.

The United Nations' sprawling Internet presence includes tables, maps, and graphs based on estimates of international migration flows and migrant populations elaborated by the UN Department of Economic and Social Affairs (UN DESA). The bonus of global coverage comes at the price of varying definitions and data quality. Such issues have delayed the launch of the **United Nations' Global Migration Database (UNGMD)** (https://population.un.org/unmigration/index_sql.aspx), which draws on data from about 200 countries. At the time of writing, this database was still being tested.

Special mention is due to **Our World in Data** (https://ourworldindata.org), which is a website that, thanks to the collaborative effort of numerous academics, delivers open-access data and analyses on a vast range of issues, including the root causes of international migrations such as global income inequality or population growth. This portal features concise contextualization and impactful data visualizations.

7.3.3 Data Repositories

Apart from researchers' contributions to making the data generated by governmental organizations and public administrations widely accessible, the scientific community also has expressed an increasingly strong commitment to the public availability of research data. This concern is motivated, on one hand, by a quest for transparency, which demands that third parties be offered the opportunity to reproduce results to verify scientific discoveries or claims. On the other hand, a growing insistence has arisen that the very nature of scientific inquiry, being cumulative, demands open access to all the data used in research (Murray-Rust, 2008; Molloy, 2011). Both lines of reasoning are adopted increasingly by funding agencies and publication outlets, thus converting open data access into a requirement both for research projects to be financed, and their results to be published. This trend has led a number of different data repositories to flourish: scholars upload datasets and complementary information for other researchers, so they can freely reuse that data. Among the most important sites, by volume of uploaded datasets, are the Dryad Digital Repository, figshare, Harvard Dataverse, and Zenodo. The **Registry of Research Data Repositories** (http://www.re3data.org) database gathers the most extended list of data repositories arranged by type, language, country, and subject. At this time, the IOM's aforementioned Migration Data Portal is the only specialized data repository for migration studies found in the Registry of Research Data Repositories.

Dataverse, a repository management software designed at Harvard University, is one of two outstanding resources that should be highlighted. Initially conceived as a data repository of this particular institution, **Harvard Dataverse** (https://dataverse.harvard.edu/) has evolved into a platform that integrates repositories from all over the world that employ its software. Overall, at this time, Harvard Dataverse has gathered more than 114,000 datasets, about half of which belong to the Social Sciences, of which approximately 1600 datasets can be retrieved by a "migration" query. Note that all these numbers are increasing at an astonishing pace. **Zenodo** (https://zenodo.org/), another outstanding resource, was created in 2013, thanks to the collaboration between the research project OpenAIRE and the CERN (European Organization for Nuclear Research). This relatively novel, multidisciplinary repository is structured in some 7000 "communities," each of which represents an organization, research group, or subject-matter; however, few such communities are related to migration studies. This repository, which currently offers more than 115,000 data sets, also is growing fast.

To complete this roundup of web resources for locating research-based datasets on migration, two recent initiatives—geographically centered on Europe—that favor expert-defined taxonomies over algorithms are worth mentioning. The **Migration Research Hub** (http://www.migrationresearch.com), which is sponsored by the research network IMISCOE, aims to convert itself into a platform for identifying migration-related expertise across a wide range of topics. Although academic publications account for most of its content, the database also contains references to hundreds of datasets. For its part, the **EthMigSurveyDataHub** (https://ethmigsurveydatahub.eu/) focusses specifically on improving the access, usability,

and dissemination of survey data on the economic, social, and political integration of ethnic and migrant minorities. Currently, this project is developing online databases of such surveys, as well as their questionnaires' items.

7.3.4 Dataset Search Engine

To locate specific datasets across the growing and diverse range of extant data repositories, a specific search engine, Google Dataset Search (https://datasetsearch.research.google.com), is available. This search engine indexes datasets written in many different formats, on the one condition that the metadata is written according to Google's stated instructions (cf. schema.org). Query results are listed in the website's screen's left-side margin, while the right-side displays detailed information about a chosen item, such as title, contents, link to the repository, last update, author, license, format, etc. Searches can be customized with various parameters. This recently launched service will prove enormously useful to researchers in any thematic domain, including migration studies.

7.4 Data Extraction

Many repositories do not require any particular procedure: after downloading the data (in widely used formats such as .csv or .xlsx), a user just has to clean the file to retrieve the precise information required, or perhaps adapt the data to the processing system that they are using. However, in some cases, the platforms provide several endpoints that automatize data extraction and develop specific applications for data analysis. In the following discussion, we briefly present three different data extraction techniques: Application Programming Interfaces (APIs), Web Scraping techniques, and the SPARQL language.

Each of these options has specific advantages when employed with regard to particular types of data. APIs and SPARQL are suitable for obtaining and analyzing structured data, such as governmental statistics, because these endpoints are commonly created by the data providers for dissemination. Both techniques are appropriate for large and updated datasets. However, scraping techniques are recommended for extracting the non-structured data available on the Web, such as textual information (e.g., public opinion blogs or media reports) or links on social network sites.

While we anticipate that the following section might strike some readers as somewhat more arcane than the preceding sections, we would like to stress that the procedures sketched here do not necessarily require prolonged training to be put to use. Indeed, a key goal of this chapter is to encourage migration scholars to become acquainted with these techniques (for further reading, see Salah et al., forthcoming).

7.4.1 Application Programming Interfaces (APIs)

In general terms, an application programming interface (API) refers to a set of functions and procedures that enable software to interact with other applications (Blokdyk, 2018). A Web API is designed specifically to provide direct access to web servers with a view to using their data in other applications in a massive and automatic form. Web APIs are offered by the data provider, which means that it is only possible to access the data if the server has implemented a public API to operate with it; fortunately, this is the case with many of the data portals mentioned previously. Commonly, access is obtained through a HTTP protocol specifying a URL that pinpoints a route to the server where the data are stored. Normally, a data provider supplies a handbook or guide with information about the content, structure, and type of queries that their API supports.

An API can be accessed in two ways: by the representational state transfer (REST) procedure, on one hand, and the simple object access protocol (SOAP) method, on the other. REST uses different parameters in the data provider's URL to retrieve and filter the needed information.[2] This is the most extended method because it grants more freedom when designing a query to enable the selection of the exact elements required for a given research purpose. SOAP requests are not defined in terms of target URLs; rather they use SQL and SPARQL (see Sect. 7.4.3) to query the database.

APIs supply data in several formats that facilitate the understanding of the structure of the dataset and its subsequent processing and management. The most important data formats are JavaScript object notation (JSON), which is supported by the REST protocol, and extensible markup language (XML), which is employed by REST and SOAP.

JSON is an open data-interchange format that facilitates both reading by humans and parsing and generating by machines (Crockford, 2006). This format displays the information in a structured, user-friendly, and compressible way, which makes it possible to, for example, contract and expand items, depict hierarchical object structures, color objects by type, and filter the text. In addition, JSON is suitable even for non-structured data because it is not necessary to previously define fields and attributes. These advantages have made JSON the most extended format, and it is implemented in most extant Web APIs. However, the XML format remains common as well, and it displays information in a structured form using marks that define each object and their associated attributes (Abiteboul et al., 2000). Similar to JSON, XML enables the contraction or expansion of hierarchical data trees. Xpath is the language used for the extraction and processing of data in XML.

[2] For example, https://api.census.gov/data/2019/pep/population?get=STATE,NAME&POP&for=state:* retrieves data from the 2019 US Census dataset population (Population Estimates from the US and Puerto Rico), variables STATE (State code), NAME (State name) and POP (Resident Population total), grouped by state.

The use of APIs has important advantages for users and data providers. First and foremost, APIs permit access to huge data volumes automatically and rapidly. Second, they do so in ways that adapt to the requirements of the user's data processing system by enabling structured queries and filters; by capturing only the specific data needed for a given purpose; and by saving time, storage space, and system capacity. Third, APIs facilitate access to real-time data: retrieval is possible as soon as the data are generated. This advantage is a huge bonus when compared to closed data files, especially with regard to short-lived or high-frequency parameters such as online social networks, etc. Finally, APIs favor free-flowing web traffic, since they are much less taxing in terms of file size than the formats, images, and applications commonly disseminated on the Web, which reduces the risk of server saturation and failure.

However, APIs have some limitations that impede or restrict their use for research purposes. The principal problem is that some data providers, especially commercial ones that offer paid services on demand, do not allow API access to all the information displayed on their respective websites, since they consider some of that information to be sufficiently valuable to impede it from being massively processed by third parties. For example, Facebook and Twitter only make a fraction of all their data freely available. Another limitation is that APIs require some programming skills. Although many sites provide an endpoint that helps with the writing of queries, sometimes it is necessary to know some programming language (JavaScript, SQL, Python, R, etc.) when designing a query and storing the data.

7.4.2 Web Scraping

Web scraping is a technique for extracting data from Web pages. This is achieved by simulating the navigation of a web user and capturing the information displayed on the computer screen (Vanden Broucke & Baesens, 2018; Mitchell, 2018). Technically, this procedure extracts text patterns from the HTML structure of a web page and then creates a structured file (.csv, .xls, etc.) with the harvested data. Among the most popular applications for web scraping are Heritrix (Internet archive), Import.io (web solution) and Scrapy (Phyton). The main advantage of web scraping as compared to API access is that the number and type of data retrievable is almost unlimited. Thus, this tool is recommended when API access is rated as too restrictive with respect to the research objectives being pursued.

In addition to the possibility of seeking information from specific and pre-identified websites, the Web also can be tracked automatically by following the network of links that connect each web page; the applications employed to achieve this tracking are variously called *web crawlers*, *spiders*, or *bots*. Such applications explore all the links they identify according to a range of parameters (links from specific sites, depth, content, etc.) and extract basic information about the webpages accessed (Olston & Najork, 2010). These instruments have long been used by search engines (Googlebot, Bing bot) to index content from the Web, as well as

by webmasters for detecting broken links (link rot) and design errors. With respect to web scraping, the role of bots is to run the navigation, whereas the scrapper extracts information from each visited page. This joint activity enables a massive extraction of data from large websites and an integration of information from various web pages.

R language, which is used widely in the scientific community to develop statistical analyses and data processing, provides an easy option for creating a crawler, among other reasons because it is freely available and does not require advanced programming knowledge (Munzert et al., 2014; Aydin, 2018). The most important tool is Rcrawler, a package that permits the development of a crawler to extract information from the Web. This package includes a specific module for scraping data from patterns in HTML and Cascading Style Sheets (CSS) tags (Khalil & Fakir, 2017).

Rcrawler can be customized to define different values for a range of arguments. For example, the following script tracks the United Nation's site on Refugees and Migrants with depth 1 (only the main page of the site), 10 threads, a request delay of 2 seconds, and a timeout of 10 seconds. These parameters enable the definition of the behaviour of the crawler from a technical viewpoint.

```
Rcrawler(Website = "https://refugeesmigrants.un.org/", MaxDepth = 1, no_conn =
    10, RequestsDelay = 2, Timeout = 10, urlregexfilter = "/refugees-compact/",
    KeywordsFilter = c("Syria"))
```

Rcrawler also permits the filtering of the crawling process by addressing and retrieving only pages that fit specific criteria. For example, the previous script surfs only on pages in the "Refugees" Section and extracts only information from pages that include the word *Syria*. However, this process only harvests information about web pages (title, url, inlinks, outlinks, etc.), so another module is necessary to extract text patterns from the content of these pages. ContentScraper is a function of Rcrawler that extracts text from the HTML by using Xpath and CSS tags. In the following example, the script extracts the table "Number of first-time asylum applicants" from the Wikipedia entry "European migrant crisis."

```
data <-ContentScraper(Url = "https://en.wikipedia.org/wiki/European_migrant_
    crisis", XpathPatterns = c("//table[@class='wikitable toptextcells sortable']"),
    PatternsName = c("table"), ManyPerPattern = TRUE)
```

The argument "XpathPatterns" points out in which part of the HTML code the text to extract is located. Applications designed with R offer much more flexibility than any commercial bot. Rcrawler enables users to define each parameter with any value, thus focusing the process toward the needed data. However, it requires some programming and is limited to the functionalities of each package; also, the data obtained need to be cleaned and processed before being analyzed.

While advantageous to the proficient user, the extraction of massive data from the Web has important legal and technical implications that require careful consideration. In many countries, legal constraints tend to be concerned with the subsequent use of data, rather than the extraction process as such (Kienle et al., 2004; Marres &

Weltevrede, 2013). For example, United States case law considers data duplicity and re-publishing to be legal (*Feist Publications vs. Rural Telephone Service; Associated Press vs. Meltwater U.S. Holdings, Inc.*) (Hamilton, 1990; Schonwald, 2014), while considering massive data extraction that inflicts any commercial damage (*eBay vs. Bidder's Edge; Cvent, Inc. vs. Eventbrite, Inc.*) as illegal (Chang, 2001)—thus, copyrighted material is protected. In Europe, the General Data Protection Regulation (GDPR) limits the web scraping of EU residents' personal data (email and physical addresses, full names, birthdates, etc.), even if publicly displayed on web pages, by requiring the affected subjects' consent; importantly, it does not interfere with the extraction of non-personal data, and thus mostly affects the data extraction from online social networks. These rulings and regulations confirm that the automated extraction of non-personal data from websites is perfectly legal, especially when serving scholarly purposes.

However, web-scraping can inflict functional damage to websites. Automated high-frequency, large-volume requests of web pages can occupy an excessive amount of bandwidth, thereby slowing down the service, impeding access by other internauts, or even triggering server failure, thus causing potentially serious economic harm. Also, web-scraping may distort statistics about visits, downloads, likes, etc. Consequently, in their terms and conditions, some corporations and entities explicitly prohibit the scraping of their web resources.

Such problems can be largely prevented by following simple courtesy rules that permit data extraction in a respectful way. The most important rule is to identify any crawling/scraping process in the "user-agent" field of the application by stating that this is a robot and revealing its IP, which enables the tracing of a person responsible for web resources malfunctioning or misconduct. Second, appropriate time intervals (ideally several seconds) should be set between requests (many crawlers/scrapers include options to define the rate of petitions). Also, the number of threads (simultaneous requests launched to the server) should be chosen conservatively. Last, the extraction period and corresponding use of bandwidth can be reduced by the efficient design of crawlers/scrapers, with a view to visiting only the pages necessary for achieving well-specified objectives. These criteria should be modulated according to the size and importance (i.e., number of daily visits) of a site: while goliaths such as Google, LinkedIn, or YouTube maintain high-volume servers that are barely affected by the activity of a bot, less well-resourced sites can collapse due to invasive crawling.

7.4.3 SPARQL Language

SPARQL is the recursive acronym of *SPARQL Protocol and RDF Query Language*. As its name indicates, it is a language designed to query data repositories that contain RDF data (Feigenbaum, 2009). The result of a query can be displayed in several formats: XML, JSON, RDF, and HTML. Different interfaces (YASGUI, Virtuoso, Stardog, Sparql Playground, etc.) can be used to access any SPARQL endpoint to help write the queries.

The main advantage of SPARQL is that this language is simple and easy to use if RDF philosophy is understood. Another important benefit is that it enables the integration of different datasets in the same query by linking data from distant parts. However, SPARQL has important drawbacks: the RDF model is not widely extended, not all data repositories provide their data in RDF format, and it is necessary to be acquainted with the structure of a SPARQL data query.

The SPARQL structure is defined by five elements: (1) a prefix declaration (PREFIX) of the URIs that will be used in the query if we want to abbreviate them, (2) a dataset definition (FROM) stating what RDF graphs are being queried, (3) one or several result clauses (SELECT, ASK, CONSTRUCT, DESCRIBE) identifying what information to return from the query, (4) a query pattern (WHERE) specifying what to query for in the underlying dataset, and (5) query modifiers that slice, order, and otherwise rearrange the query results; for example, a user can limit the number of results (LIMIT), rank them by some criteria (ORDER BY), or define from what point they are visualized (OFFSET).

Since the repository commonly establishes the RDF graph by default, the dataset definition is optional. In contrast, the prefix declaration is mandatory. SPARQL accepts as many prefixes as the user wants, even if they are not used in the query. To know the correct prefix of any URI, the user can employ prefix.cc, which is a search engine that retrieves the full link of any prefix. To define query patterns, a user can place an interrogation mark before the corresponding word.

To illustrate, the following example selects two fields: Title and Title_Subject. First, it returns articles from the category Human migration; next, it takes the title from the articles; and then, the title of the subject.

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
Prefix category: <http://dbpedia.org/resource/Category:>
PREFIX dct: <http://purl.org/dc/terms/>
SELECT ?Title ?Title_Subject WHERE{
?Articles dct:subject category:Human_migration .
?Articles rdfs:label ?Title .
?Articles dct:subject ?Subjects .
?Subjects rdfs:label ?Title_Subject .
FILTER (LANG(?Title) = "en")
}
```

In this example, a filter is set to limit the results to the English language ("en").

7.5 Concluding Remarks

The terms *big data, open data,* and *linked data* allude to an ongoing transformation of scientific research (Crosas et al., 2015). Troves of previously unavailable data are opening up innovative research opportunities, improving the reproducibility of results, and interconnecting datasets across the globe. This data revolution

contributes to making scientific research more collaborative and transparent than seemed conceivable just a few years ago. The center of gravity of the research process is inexorably being displaced from (relatively small-scale) data production toward (markedly large-scale) data extraction.

Although this data revolution does not mean that empirically minded social scientists will cease to conduct fieldwork of their own, it does mean (at least in our opinion) that the value of primary research data will be benchmarked increasingly against data collected by third parties via the Internet. Even conceding, as we are inclined to do, that large volumes of data do not necessarily convey rigorous knowledge, not to mention wisdom, it would be foolish to ignore the huge opportunities awarded social researchers by the transformative leap of pervasive digitization with respect to the timeliness, cost, variety, versatility, comprehensiveness, and ubiquity of data. The Web provides access to real-time, free-of-charge, population-level, and finely-grained data on an ever-expanding array of facets of social reality, everywhere on the globe. The ensuing research opportunities are especially obvious in the field of migration studies, given that, rather often than not, migrant populations are "hard-to-reach" (Font & Méndez, 2013) with traditional research methods due to the combination of cultural diversity and geographic dispersion, on the one hand, and precarious administrative status and under-coverage by official sources, on the other.

In this chapter, we have reviewed some outstanding examples of the current line-up of web resources relevant for migration research, including specialized websites, generalist data portals, dataset repositories, and a dedicated dataset search engine. In addition to providing access to a continuously growing trove of data, these resources facilitate the sharing of our own primary data so to make them reviewable and reusable. This snapshot should not be mistaken for a permanent inventory: since the Internet is in constant evolution, additional web resources for migration scholars are sure to emerge rather sooner than later, potentially eclipsing some of the offerings mentioned here. Thus, the search for data sources relevant to a specific research project constitutes a vital (if obvious) precondition for benefiting from the ongoing data revolution.

However, to seize the opportunities granted by the process of ever-more ubiquitous digitization, source identification is only the first step, since scholars also require some technical skills (Light et al., 2014). In most research institutions, data processing is not (yet) a task for specialized programmers, but rather a set of abilities that researchers need to develop to effectively manage their projects. In this chapter, we have sketched three of the most important ways of extracting data from the Web: APIs, web scraping, and SPARQL. More and more organizations (governments at all levels, intergovernmental and international institutions, for-profit companies, etc.) are creating open endpoints for accessing the data they produce, thereby generating new research possibilities. Presently, APIs are the best and most common mode of data provision, although the use of linked data technologies and SPARQL endpoints is becoming more frequent mainly by governmental and statistics offices. Yet, in some cases, data are not easily accessible and web scraping techniques are necessary for obtaining the required information. This technique enables the harvesting of valuable non-structured data from different websites, and can adapt to any web structure, although it must be used responsibly to prevent disruption.

The risk of automated data requests overwhelming the capacity of target servers is one of the challenges originated by researchers' piggybacking on the ongoing process of global digitization. A second challenge of paramount importance is the preservation of privacy and the protection of personal data. This is a serious problem given that vast amounts of personal information, including highly sensitive data, are easily accessible on social networks and other web platforms. In spite of regulations (GDPR) and case laws that prohibit abusive scraping, the risk of security breaches is evident (Isaak & Hanna, 2018). The responsibility to protect non-public personal data by employing encryption and anonymization procedures falls primarily on platform owners and web managers. The GDPR forbids the scraping of publicly available personal data without subjects' consent, which contrasts with a more laissez-faire approach in the US (cf. *hiQ Labs, Inc. vs. LinkedIn Corp.*). Regarding both server saturation and privacy protection, it seems fair to say that scholars' web crawling poses much less of a challenge than web mining for commercial or political purposes. Still, the imperative of ethical conduct requires researchers to proactively prevent any harm that potentially derives from their data collections.

Third, the many advantages of web-based data extraction must not obscure the most basic of methodological precautions—not to confuse data coverage with truth. Any data category or research result contains some trace of its conditions of production, a context that inevitably shapes patterns of visibility, intelligibility, and knowability. Even with regard to population-level observational data, concepts such as *coverage bias* and *selection bias* continue to be pertinent: cases in point are the socio-demographically skewed distribution of body-sensor wearables and Internet penetration rates, as is the digital divide across the global North and South. Since the increasing pervasiveness of digitization seems prone to breed hubris, we pointedly recommend big-data analysts to be humble instead.

Based on the expectation that these challenges will prove manageable, we would like to conclude by insisting on the strategic importance of migration scholars building, either individually or collectively, the skills necessary for successfully navigating this emerging new world of big, open, and linked data.

References

Abiteboul, S., Buneman, P., & Suciu, D. (2000). *Data on the Web: From relations to semistructured data and XML*. Morgan Kaufmann.

Ahas, R., Aasa, A., Mark, Ü., Pae, T., & Kull, A. (2007). Seasonal tourism spaces in Estonia: Case study with mobile positioning data. *Tourism management, 28*(3), 898–910.

Aydin, O. (2018). *R Web Scraping Quick Start Guide: Techniques and tools to crawl and scrape data from websites*. Packt Publishing.

Barslund, M., & Busse, M. (2016). *How mobile is tech talent? A Case Study of It Professionals Based on Data from LinkedIn* (CEPS Special Report n° 140). Centre for European Policy Studies.

Bartling, S., & Friesike, S. (Eds.). (2014). *Opening science: The evolving guide on how the internet is changing research, collaboration and scholarly publishing*. Springer. https://doi.org/10.1007/978-3-319-00026-8

Bergman, M. (2009). Advantages and Myths of RDF. *AI3, April*.
Berners-Lee, T. (2006). *Design issues: Linked data*. https://www.w3.org/DesignIssues/LinkedData.html
Blokdyk, G. (2018). *API Application Programming Interface Brisbane* (288 pp). Emereo Pty Limited.
Carroll, J. J., & Stickler, P. (2004). RDF triples in XML. In *Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters* (pp. 412–413).
Chang, E. W. (2001). Bidding on trespass: eBay, Inc. v. Bidder's Edge, Inc. and the abuse of trespass theory in cyberspace-law. *AIPLA QJ, 29*, 445–468.
Charalabidis, Y., Zuiderwijk, A., Alexopoulos, C., Janssen, M., Lampoltshammer, T., & Ferro, E. (2018). *The world of open data – Concepts, methods, tools and experiences*. Springer.
Crockford, D. (2006). *The application/json media type for javascript object notation (json)* (No. RFC 4627).
Crosas, M., King, G., Honaker, J., & Sweeney, L. (2015). Automating open science for big data. *The ANNALS of the American Academy of Political and Social Science, 659*(1), 260–273.
Curé, O., & Blin, G. (2014). *RDF database systems: Triples storage and SPARQL query processing*. Morgan Kaufmann.
Deville, P., Linard, C., Martin, S., Gilbert, M., Stevens, F. R., Gaughan, A. E., ... Tatem, A. J. (2014). Dynamic population mapping using mobile phone data. *Proceedings of the National Academy of Sciences, 111*(45), 15888–15893.
Dodge, S. (2019). A data science framework for movement. *Geographical Analysis*. https://doi.org/10.1111/gean.12212
Feigenbaum, L. (2009). *SPARQL by example: A tutorial*. World Wide Web Consortium. https://www.w3.org/2009/Talks/0615-qbe/
Font, J., & Méndez, M. (Eds.). (2013). *Surveying ethnic minorities and immigrant populations*. Amsterdam University Press.
Giannotti, F., Gabrielli, L., Pedreschi, D., & Rinzivillo, S. (2016). *Understanding human mobility with big data* (In *Solving Large Scale Learning Tasks. Challenges and Algorithms* (pp. 208-220)). Springer.
Greco, F., Maschietti, D., & Polli, A. (2017). Emotional text mining of social networks: The French pre-electoral sentiment on migration. *Rivista Italiana di Economia Demografia e Statistica, 71*(2), 125–136.
Hamilton, M. A. (1990). Justice O'Connor's Opinion in Feist Publications, Inc. v. Rural Telephone Service Co.: An Uncommon Though Characteristic Approach. *Journal of Copyright Society USA, 38*, 83.
Heath, T., & Bizer, C. (2011). Linked data: Evolving the web into a global data space. *Synthesis lectures on the semantic web: theory and technology, 1*(1), 1–136.
Khalil, S., & Fakir, M. (2017). RCrawler: An R package for parallel web crawling and scraping. *SoftwareX, 6*, 98–106.
Kienle, H. M., German, D., & Muller, H. (2004, September). Legal concerns of web site reverse engineering. In *Proceedings. Sixth IEEE International Workshop on Web Site Evolution* (pp. 41–50). IEEE.
Kupiszewska, D., Kupiszewski, M., Martí, M., & Ródenas, C. (2010). *Possibilities and limitations of comparative quantitative research on international migration flows*. European Commission.
Isaak, J., & Hanna, M. J. (2018). User data privacy: Facebook, Cambridge Analytica, and privacy protection. *Computer, 51*(8), 56–59.
Laney, D. (2001). 3D data management: Controlling data volume, velocity and variety. *META group research note, 6*(70), 1.
Light, R. P., Polley, D. E., & Börner, K. (2014). Open data and open code for big science of science studies. *Scientometrics, 101*(2), 1535–1551.
Lin, A. Y., Cranshaw, J., & Counts, S. (2019). Forecasting US Domestic Migration Using Internet Search Queries. In *The World Wide Web Conference* (pp. 1061–1072). ACM.
Marres, N., & Weltevrede, E. (2013). Scraping the social? Issues in live social research. *Journal of cultural economy, 6*(3), 313–335.

Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt.
Mitchell, R. (2018). *Web Scraping with Python: Collecting More Data from the Modern Web*. O'Reilly Media.
Molloy, J. C. (2011). The open knowledge foundation: open data means better science. *PLoS biology, 9*(12), e1001195.
Murray-Rust, P. (2008). Open data in science. *Serials Review, 34*(1), 52–64.
Munzert, S., Rubba, C., Meißner, P., & Nyhuis, D. (2014). *Automated data collection with R: A practical guide to web scraping and text mining*. Wiley.
Olston, C., & Najork, M. (2010). Web crawling. *Foundations and Trends® in Information Retrieval, 4*(3), 175–246.
Salah, A. A., Korkmaz, E. E., & Bircan, T. (forthcoming). *Data science for migration and mobility research*. Oxford University Press.
Salah, A. A., Pentland, A., Lepri, B., & Letouze, E. (Eds.). (2019). *Guide to Mobile Data Analytics in Refugee Scenarios*. Springer.
Schonwald, R. J. (2014). Associated Press v. Meltwater US holdings, Inc.: fair use, a changing news industry, and the influence of judicial discretion and custom. *Berkeley Technology Law Journal, 29*, 799–833.
Sîrbu, A., Andrienko, G., Andrienko, N., Boldrini, C., Conti, M., Giannotti, F., … Pappalardo, L. (2020). Human migration: the big data perspective. *International Journal of Data Science and Analytics*, 1–20.
Spyratos, S., Vespe, M., Natale, F., Weber, I., Zagheni, E., & Rango, M. (2019). Quantifying international human mobility patterns using Facebook Network data. *PLOS ONE, 14*(10), e0224134.
State, B., Rodriguez, M., Helbing, D., & Zagheni, E. (2014). Migration of Professionals to the U.S. Evidence from LinkedIn Data. In L. M. Aiello & D. McFarland (Eds.), *Social Informatics. 6th International Conference, SocInfo 2014, Barcelona, Spain, November 11-13, 2014. Proceedings* (pp. 531–543). Springer International Publishing. https://doi.org/10.1007/978-3-319-13734-6_37
Vanden Broucke, S., & Baesens, B. (2018). *Practical web scraping for data science: Best practices and examples with Python*. Apress.
Wilson, R., Erbach-Schoenberg, E., Albert, M., Power, D., Tudge, S., Gonzalez, M., … Pitonakova, L. (2016). Rapid and near real-time assessments of population displacement using mobile phone data following disasters: The 2015 Nepal Earthquake. *PLoS currents, 8*.
Wladyka, D. K. (2017). Queries to Google Search as predictors of migration flows from Latin America to Spain. *Journal of Population and Social Studies, 25*(4), 312–327.
Zagheni, E., Garimella, V. R. K., Weber, I., & State, B. (2014). Inferring International and Internal Migration Patterns from Twitter Data. *WWW '14 Companion: Proceedings of the 23rd International Conference on World Wide Web,* 439–444. https://doi.org/10.1145/2567948.2576930

Chapter 8
How Canada's Data Ecosystem Offers Insights on the Options for Studying Migration in an Unprecedented Era of Information

Howard Ramos and Michael Haan

8.1 Introduction

The contemporary era has been marked by at least three unparalleled trends. The first is the unprecedented mobilities of people (Urry, 2016); The second is unprecedented migration (United Nations, 2019); and the third trend, which forms the focus of this chapter, is the never before seen amounts of data and information that can be used to track internal and international migration (Haan, 2019). This includes traditional data from censuses and national surveys, but also administrative records and other new data sources. All three trends have sparked academic and policy debate, as well as agreement on the need to use data to not only track movement, but to also generate evidence-based policy interventions to meet the needs of migrants and facilitate their integration into adopted host communities.

Despite these three trends, it remains surprisingly difficult to track migration. This is, in part, due to an overall lack of awareness among researchers and policy makers of available data options and the pitfalls that come with using them. Few sources, moreover, offer practical and straightforward overviews of the decisions involved in the processing of data and the analytical gaps that are found in them. For these reasons, in this chapter, we map out options for understanding and analyzing migration through data.

In doing so, we first consider traditional Census and survey-based options from statistics agencies. We then explore the increasingly available option of using governmental administrative records. Finally, we assess other emerging big data sources that can be harvested from new media or private sector records. Along the way, we address some of the technical and methodological issues that arise from using data from different sources to study migration. These include the need for

H. Ramos (✉) · M. Haan
Western University, London, ON, Canada
e-mail: howard.ramos@uwo.ca; michael.haan@uwo.ca

S. Pötzschke, S. Rinken (eds.), *Migration Research in a Digitized World*, IMISCOE Research Series, https://doi.org/10.1007/978-3-031-01319-5_8

consistent definitions, limits in the scope of what can be examined, and the need for an awareness of issues around access and balancing necessity and proportionality around privacy. We also highlight the work that needs to be done to extend data created for one purpose, in the case of administrative records and big data, to be used for others, as is the case when these data are translated for research purposes.

In discussing these issues, we focus on Canada, our home country, but we also extend our observations to other countries. We conclude by advocating for the creation of data spines across sources and working towards shared national and international standards that will truly leverage the full potential of data on migration. There has never been a time more in need of migration research, or with as much information on the topic, as now.

8.2 Why Focus on Canada's Immigration Data Ecosystem?

Global migration made international headlines around the world in 2015, with unseen numbers of people migrating into Europe as a result of the (largely Syrian) refugee crisis. The increase in the international flow of people became a focal point of attention for social scientists and policy makers. This is especially the case for immigrant-settler countries, such as Australia or Canada, where immigrants make up 28.2% and 22% percent of their populations, respectively (Australia Bureau of Statistics, 2017; Chavez, 2019). Increased migration has also been witnessed in non-traditional immigrant countries in Europe that opened their doors to refugees and newcomers at the peak of the refugee crisis.

With approximately 300,000 immigrants landing permanently in Canada each year, and over 500,000 temporary residents on the ground (Hussen, 2018) for most of the 2010s and an increase to 400,000 immigrants landing in the 2020s and about the same number of temporary residents, Canada has become a world leader in migration and immigration policy (Trebilcock, 2019). It also leads in terms of refugee resettlement, surpassing the United States in the total numbers in 2018, and settling far more refugees (as a proportion of the population) than other countries (Radford & Connor, 2019). Canada has developed cutting-edge immigration policy and has invested heavily in tracking migration and immigrant settlement to be able to understand how newcomers fare in the country.

Canada is also a leader in terms of data collection and innovative uses of this information to study migrants, as its data landscape is broad, and includes a wide range of sources. Specifically, the country's censuses and national surveys collected by its national statistical agency, Statistics Canada, include immigrant status variables. Many other countries capture migrants in their data, as can be seen in the case of Australia, the United Kingdom, and the United States, but Canada is somewhat unique in its approach to administrative data on immigrants. Unlike many European or Scandinavian countries with population registers, such as Denmark or Sweden (Careja & Bevelander, 2018), Canada has no single and comprehensive

administrative record to identify all people in its population. Instead, it has invested into processing and using a wealth of administrative data from Immigration, Refugees and Citizenship Canada (IRCC). These include using immigrant landing records and temporary resident records, as well as linkage of those records to other administrative databases, such as Canada Revenue Agency tax filing records or health records. Other countries without population registries have also invested in administrative data uses, especially to study income and employment issues, as can be seen in Australia or the United States, and/or migration flows, as seen in the United States or United Kingdom, and, like Canada, aim to link administrative records to other data to better capture migration (Ernsten et al., 2018; Rogers & McNally, 2018). Canada has also invested in linking its data across Censuses, multiple administrative records at the national level, and administrative records at subnational levels. Like other countries, researchers in Canada also collect their own data and are increasingly turning to Big Data from online or other sources. The goal of our chapter, however, is not to offer a comprehensive comparative scoping review of data available, nor be able to speak generally about practices globally, but rather, to share our experience with the Canadian migration data ecosystem and to highlight issues that arise within it.

Our insights and analysis are drawn from our nearly 50 years of collective experience working on migration and immigration issues. As such, we share some of the obstacles, solutions, and opportunities we see from navigating an increasingly complex data landscape. The chapter discusses issues encountered across three categories of data: (1) censuses and national surveys, (2) administrative data, and (3) emerging data sources. These groupings are used by the Migration Data Portal (2019), which was developed by the International Organization for Migration's (IOM's) Global Migration Data Analysis Centre, and offer a useful framework to frame our discussion.

8.3 Canada's Data Ecosystem

8.3.1 Censuses and National Surveys

Probably the most widely used data source for studying migration and immigration is the census. It is an official count, or survey, of a population that aims to describe it and are usually done by countries every five or 10 years (Ruotsalainen, 2011). In Canada, data are collected through a combination of mailout surveys, face-to-face enumerations, and online surveys. Since 2006, rather than entering income data manually (which is prone to error or crude approximations), tax records have been used to provide economic information on individuals.

The Canadian census has two components: the short form, which in 2021 consisted of the 17 core questions asked to all participants, and the long form, which in 2021 had an extended set of questions sent to 25% of households (Statistics Canada, 2021a, b). With respect to migration, the long form of the census has

information on mobility within the country over the last 5 years. It also captures mother tongue. Through census linkage to administrative records it also captures important factors, such as immigrant year of landing, admission category, the source country of immigrants. These were previously captured in the long form. Another strength of the census is that it also collects information on an individual's other demographic, social, and economic characteristics. For these reasons, the census has long been a dominant source of information about Canadian immigrants. It has been used to study ethnic origins of immigrants (Boyd, 1999), earning disparities between immigrants and native-born citizens (Li, 2000), and home ownership (Haan, 2007), to name but a few examples. With the addition of immigrant admissions category since 2016 census, it is likely to become even more widely used in the future for research on immigrant settlement. These categories offer insight into the pathways immigrants used to land in Canada and, for example, offer opportunities to study differences between economic immigrants versus those arriving through family or refugee intake categories.

The census is also heavily used in other immigrant-settler countries, such as Australia and the United States. In fact, it is used by 149 countries around the world. Almost all, 87%, collect information on country of birth, however, about a quarter lack detailed information on citizenship (Migration Data Portal, 2019). The most prominent case where citizenship information is not collected is the United States, as seen through the controversy sparked by the Trump administration's attempt to include it in the country's 2020 Census (Mervis, 2019). Even fewer countries (roughly 50%) collect information on immigrant period of arrival (Migration Data Portal, 2019). Countries are also limited in terms of how many questions are asked in their census.

Despite this obstacle, there have been some innovative analytical strategies used to look at immigrants longitudinally, such as the 'double cohort method' (see Myers, 1999; Myers & Lee, 1998).

For countries where censuses are an option, there are at least three considerations when using the data. The first is that the census is a cross-sectional dataset. This means that one cannot directly study trends of immigrants across time. As a result, any changes observed cannot definitively be linked to cause in a previous temporal period (Borjas, 1993). Failure to recognize this can lead to erroneous conclusions, such as the "cross-sectional integration fallacy" identified by Hum and Simpson (2004). They observed the fallacy with respect to immigrant earnings, however, the mechanisms they identify apply to any causal inference made. The problem is that a quasi-cohort is created, which essentially compares people from a given group, say immigrants, who theoretically are of the same age and are then described or compared against in a later cohort. Because there is no one-to-one match, the comparison may fail to account for new immigrants or migrants that fit the same profile of the later cohort or other factors that make the groups different.

At least for Canada, a second problem occurs when focusing on issues of internal migration. The Canadian Census contains questions only on place of residence one and 5 years ago. This means it cannot capture migration that happened earlier than that period, nor can it capture the timing of moves that occur within the period. This

means that the analysis of repeat migration within the periods is missed. Several researchers have commented on problems associated with this, and a good review of them can be found in Aydemir and Robinson (2006).

Censuses definitions also change over time, both in terms of specific questions as well as spatial units, which presents a third set of issues to wrestle with. Take, for example, the questions around ethnicity, visible minority, or occupation, which all have seen changes over the last 50 years. With respect to ethnicity, participants were discouraged from answering 'Canadian' for several years and the construction of ethnicity as a single versus multiple origin also changed (see Boyd, 1999). A direct measure of visible minority that allows for the capture of race only emerged in 1996 in response to being able to measure it as a part of the Canadian Multiculturalism Act and adoption of Employment Equity (see Boyd et al., 2000). The measurement of occupation has also changed across censuses, with the Standard Occupational Classification used before 2001 being replaced by the National Occupational Classification system, used from 2001 onward, and updated in 2006, 2011, and 2016.

In terms of spatial analysis, Census Tracts (CT), Dissemination Areas, and other geographic units also change over time. Such revisions are made due to new road construction, neighborhood growth, population growth within the CT or other units, and community development. In most cases, a CT, or other units, are split into multiple units over time, requiring researchers to recreate the original boundaries by aggregating the data if they want to study changes between multiple Censuses. In other cases, boundary revisions occur in ways that make the statistical 'reconstruction' of the original geographical boundaries laborious and/or impossible (Kaida et al., 2020). To address some of the challenges, the Canadian Longitudinal Census Tract Database has been developed to study neighborhood changes at the CT level using the 1971–2016 Canadian Censuses (see Allen & Taylor, 2018 for more details). Taken together, these issues also make longitudinal analysis more difficult and require a fairly high level of technical sophistication.

Methods for capturing information also change over time, which is a fourth problem to working with Census data. With respect to migration, immigration, and immigrant integration, two recent methodological changes affect how comparable data are over time. The first is the linking of administrative landing records to the census, which we will discuss below, to capture micro-categories of immigrants. While past censuses asked respondents to self-identify as immigrants, since 2016, immigrant variables are derived from landing records (Statistics Canada, 2016). Similarly, administrative tax records have been used in place of reporting income since the 2006 Census. Using census data over time was also complicated with the use of a National Household Survey in place of a census in 2011. During that year, participation was voluntary, leading some to argue that the data were worthless because of response bias (Hulchanski, 2014). It has since been discontinued as a survey and Censuses are again used. Nevertheless, this means that methods have changed through time, which makes comparison inconsistent over long period.

Another set of options for exploring issues of migration and immigration and immigrant settlement is looking at other surveys collected by national statistics agencies. With the Canadian case, these include the Ethnic Diversity Survey

(EDS), the General Social Surveys (GSS), Labour Force Survey (LFS) or Canadian Community Health Survey (CCHS). In Canada, the GSS program was established in 1985 and is a series of independent, annual, voluntary, cross-sectional surveys, each covering one topic in-depth (Statistics Canada, 2019a). At the time of writing there are 35 cycles of the GSS. These surveys cover a wide range of topics including civic participation (Fong & Shen, 2016; Wong & Tézli, 2013), sense of belonging to Canada (Hou et al., 2018; Wong & Tézli, 2013) as well as discrimination and health (Nakhaie & Wijesingha, 2015), among other issues. The cycles focusing on identity are particularly useful for those studying immigration, as they tend to contain similar identity questions as those in censuses but offer more detail and have a wider range of variables to compare against. Unlike censuses, they tend to capture a sample of the population. Until 1998, the target sample size was 10,000, increased in 1999 to a target of 25,000, and, in 2015, this was reduced to 22,000 before being further reduced to 20,000 from 2016 onward (*ibid*). Though rather large, shrinking sample sizes do pose challenges for statistical power, especially when immigrant subgroups become the focus.

The EDS, by contrast, aimed to dig into the ethnic, racial and immigrant experience in Canada and to explore it through economic, political, and social and cultural spheres. It has questions on year of immigrant arrival and immigrant status and measures across these dimensions. It was conducted just once, in 2002, and had a sample of 42,000 people. The survey has been used to look at a wide range of issues, such as economic performance and the link to ethnic ties (Li, 2008), perceptions and experience of discrimination (Reitz & Bannerjee, 2007), and religion and integration (Reitz et al., 2009), to offer but a few examples. Although an old survey, it has been linked to more recent administrative data, which we outline below. Such linkage shows how a data spine approach, where administrative data act as the spine for other data to connect to, can preserve the life-span of older cross-sectional surveys.

LFS is another option, which is designed to capture employment and unemployment trends and are done monthly with the aim of offering information regarding job creation, education and training, as well as income supports and pensions (Statistics Canada, 2019b). The LFS is quasi-longitudinal, in that people remain in a rotating panel for 6 months, but it suffers from fairly small sample sizes. Yet another option for studying migration in Canada is the Canadian Community Health Survey (CCHS). The CCHS was originally designed to get a better sense of the health statuses of Canadians, but its relatively large sample size (of 65,000 participants each year since 2007, down from 130,000 from 2001 until then), its regular collection schedule of every 2 years, and its ability to look at both regular and special topics within health (Statistics Canada, 2022) make it a popular option among researchers. As it pertains to migration, however, the CCHS has many of the same problems as the other cross-sectional surveys listed here.

Household surveys, conducted by national statistics agencies, are also used in many other countries (Migration Data Portal, 2019). Portugal and Ireland, for instance, rely on household surveys alone to track migration in their borders (de Beer et al., 2010). Countries such as Australia or the United States have similar general

surveys, as well as ones that focus on race or ethnicity. The United Kingdom has a Labour Force Survey that can capture those born outside the country and similar data are collected from the European Union Labour Force Survey carried out in 28 countries (Del Fava et al., 2019); many countries around the world conduct similar types of surveys.

Although GSS and other household surveys offer much potential to examine a wide range of topics, their sample size presents obstacles. This has a profound effect on the study of migration and immigrants in secondary or rural regions (see Ramos & Yoshida, 2011; Yoshida & Ramos, 2012; or Yoshida & Ramos, 2015). This is because the random sampling, and even the clustering strategies used by statistical agencies, often means that very few people from these regions are captured. A related problem is that the power of the models, and sophistication of analysis that can be done, is very limited because of the small sample. As a consequence, the areas that often most need analysis are missed, or are subject to very basic engagement through descriptive statistics. These issues add to many of the same obstacles faced with censuses.

Longitudinal surveys are yet other option available to researchers. They are useful because they contain repeated observations from the same individuals, allowing researchers to assess the longer-term migratory behaviours of people. Methodologically, this is advantageous because it allows researchers to compare the variation *within* individuals to variation *across* individuals. With such data researchers can asses how unique a particular individual is relative to her/his peer group, immigration cohort, visible minority group, and so forth.

In the Canadian context, The Longitudinal Survey of Immigrants to Canada (LSIC), which follows immigrants that landed in the country during the October 1, 2000, through September 30, 2001 period, is a somewhat dated but useful dataset (Statistic Canada, 2007). Like the EDS, it is linked to the data spine of more recent administrative data, which we discuss below. The strength of LSIC is that it contains comprehensive information about immigrants, tracking them 6 months after arrival (Wave 1) and then again at two (Wave 2) and 4 years after arrival (Wave 3). This overcomes the problems faced with Census panels or cross-sectional surveys by providing a one-to-one match of migrants and immigrants over time. It also allows researchers to examine a more detailed level of immigration categories compared to censuses before 2016 (the first year that admission category was included). The survey offers researchers the ability to assess education, racial and ethnic diversity, integration, labor market outcomes and population demography. It has been used creatively by researchers to examine the intersection of a number of immigrant experiences. For example, some have looked at immigrant language proficiency, gender and health (Pottie et al., 2008), while others have used it to analyze the experience of family pathway immigrants (VanderPlaat et al., 2013) and youth, as well as migration (Houle, 2007; Newbold, 2007; Yoshida & Ramos, 2013), among other uses.

Fewer countries have conducted longitudinal surveys off immigrants. Australia's Longitudinal Survey of Immigrants to Australia (LSIA) and New Zealand's Longitudinal Immigration Survey (LisNZ) are two comparable examples to the LSIC. The

last cohort of immigrants captured in the LSIA was in 2004/2005 (Australia Bureau of Statistics, 2011) and the last cohort of the LisNZ was surveyed in 2009 (Stats NZ, 2019). Germany's Socioeconomic Panel (SOEP), the Dutch LISS immigrant panel and the UK Household Longitudinal Survey are other examples that have longitudinal data on immigrants.

The LSIC, like other surveys, faces obstacles related to definition changes across waves as well as small sample size when it comes to studying recent immigrants in secondary regions. A more serious issue is that the survey, like other longitudinal surveys, suffered from high rates of attrition. The final sample across all waves of the LSIC was just under 8000 observations, down from the initial sample of 20,300 in Wave 1 (Statistics Canada, 2007). Careja and Bevelander (2018) note that the problem is particularly striking for immigrant populations with higher rates of attrition than other populations. Their review of longitudinal surveys found that about a third of immigrants, sometimes over half, leave panel surveys across countries.

Small samples in secondary regions and high rates of attrition mean that longitudinal surveys are very costly to run and are, therefore, often among the first datasets to be cut in times of austerity. In Canada, nearly every longitudinal survey that was once conducted by Statistics Canada has been cut. Thankfully, these surveys were cut at the same time as administrative data became more readily available.

8.3.2 Administrative Records

Administrative records are a rich and largely untapped source of data that policy makers and academic researchers can use to track migration. This is especially the case in countries that lack national registries. The decision to move to using administrative data in Canada was taken, in part, because of the elimination of the 2011 Census, which forced researchers and policy makers to look for alternatives. A positive consequence of the move was the creation of data linked to administrative records which have turned them into a data spine for other datasets in the statistical ecosystem. National registry's play that role in other countries. We believe that administrative data can act as valuable spines to link other data. Linkage to administrative data is an approach that increases the power of older surveys, cross-sectional data, or small set data created by researchers, communities, NGOs, or industry.

For the study of migration and immigrant settlement, the Permanent Resident Landing File (PRLF) is one such administrative option for creating a data spine. Every landed immigrant to Canada has a landing record, often completed by immigration officers. This administrative data allows the Canadian government to collect and maintain information on newcomers to the country. The file is both large (it captures all newcomers since 1980), detailed (languages spoken, citizenship, education at landing, intended destination, size of immigrating unit, and admission category are only some of the variables on the file), and widely used for learning

more about the country's newest residents. There are millions of unique records in the file, allowing for a detailed assessment of how Canada's efforts to recruit immigrants across different pathways have evolved over time. The disadvantage of the PRLF is that it only has information on immigrants at the time of landing, so it is not possible to learn about how immigrants are doing in Canada after that point without linking the data to other files, such as taxfiler data. This is possible because every newcomer has a unique identifier, which allows methodologists to find them in other datasets and link the files together.

Probably the best Canadian example of a successful linkage of the PRLF is the Longitudinal Immigration Database (IMDB). Linked to the T1 Family File (the main tax returns that taxfiling units submit to the Canada Revenue Agency on an annual basis), the IMDB contains all of the fields in the PRLF, as well as nearly every field that the Canada Revenue Agency requires taxfilers to submit. As with PRLF, every immigrant that has landed in Canada since 1980 is on the file, allowing for an analysis of economic outcomes from then onward. Since individuals are taxed differently if they're married or have children, the IMDB also enables researchers to look at the composition of tax filing units. The data has been used to study employment and earnings outcomes (Hou & Bonikowska, 2016; Kaida et al., 2019; Warman et al., 2015), inter-provincial mobility and retention of immigrants (Haan et al., 2017), migration and immigration in secondary regions (Ramos & Bennett, 2019; Yoshida & Ramos, 2017), as well as analysis looking at the range of immigrant and refugee pathways not available in other datasets (Kaida et al., 2019; Yoshida et al., 2016).

A number of countries lacking population registries have also considered using administrative data to explore migration issues. The Australia Bureau of Statistics, for example, has linked data on net overseas migrations with visa grants information administered by the country's then Department of Immigration and Citizenship (Temple & McDonald, 2018). The United States has assessed how Department of Homeland Security and United States State Department records can be used to study immigration (Grieco & Rytina, 2011) and has looked into how census data can be linked with Internal Revenue Service records (Akee & Jones, 2019). Researchers in the United Kingdom have also assessed how health data from the National Health Service can be linked to census-based longitudinal studies (Ernsten et al., 2018). Further, the United Kingdom aims to make administrative data the core of its immigrant and migration data infrastructure, while linking it to censuses and surveys (Rogers & McNally, 2018). Countries and researchers are increasingly moving toward administrative data to examine immigration and migration issues. Despite this trend, the Canadian case sheds light upon a number of obstacles with using administrative records.

Gaining access to administrative data is one of the biggest barriers researchers face in working with this form of information. Accessing the PRLF and the IMDB, in Canada, for instance, is especially difficult for those outside the Federal civil service or who are not affiliated with a Canadian university. Recently, Statistics Canada has created an interactive IMDB portal (see Statistics Canada, 2019c), but this does not allow researchers to access the microdata or conduct their own research.

Although access has improved in recent years, with the IMDB now housed in Research Data Centers located at most major universities across Canada, researchers still need to go through a security and screening process before gaining access, as well as being subject to vetting rules before data can be released. This slows the research process and limits who can use the data. The Research Data Centre approach is also one used by other countries, such as Germany, which has centres across the country (Bender et al., 2014). Because the data are complex and can identify individual immigrants, such protocols are not unreasonable.

Another obstacle in working with administrative data is that researchers and policy makers can only look at the issues that the data captures. More specifically, in the case of Canada's IMDB, this means a focus on economic integration as well as mobility, failing to examine other non-economic issues (Costigan et al., 2016). Another downfall is that the focus of the IMDB is on Principal Applicants. They are immigrants who drive the application to settle in Canada with sparse linkage to the family that may come with them. Although the database does have family records on the taxfiler side of the data, issues still arise over being able to distinguish between the landing family, what linkage would be when immigrants land, versus the perpetual family, or how families evolve over time (Ramos & Bennett, 2019). Recent versions of the database have included a new family marker, however, it is still early days in looking beyond the individual as the unit of analysis. The database is also being linked to a number of existing surveys and other databases, like the GSS, the CCHS, or LSIC, through partnership with IRCC.

At the same time as researchers are beginning to discover the IMDB, Statistics Canada has moved towards making even more impressive data environments though its Secure Data Linkage Environment (SDLE). One such environment is the Canadian Employer Employee Dynamics Database (CEEDD). The CEEDD contains IMDB records and a long list of other administrative files. Some of these files include corporate tax return and owner files, records of employment, employer-issued earnings statements, and exporter/importer information.

Two additional obstacles with administrative data include struggles with the size and complexity. In the case of Canada's IMDB files, when linked together, they exceed 30 gigabytes, which creates software and processing issues around analysis. This means using the dataset requires very advanced data processing and analytic skills, such as linking many different files, one-to-many merges, many-to-many merges, and the extensive use of lag variables. As such, it is not feasible for intermediate or non-specialist researchers. These are likely to be some of the longer-term issues surrounding the analysis of IMDB and many other administrative files for years to come.

Another source of administrative data that migration and immigration researchers can explore, and is also largely untapped, are sub-national administrative records. In Canada, these tend to come from provinces and territories. To date, some work has been done to explore how provincial health records can be used to assess immigrant retention and mobility within a province, as seen with recent work done in the provinces of Manitoba (Fransoo, In progress), New Brunswick (McDonald et al., 2018), and Ontario (Vigod et al., 2019). Most other provinces are also exploring how

health data can be better harnessed to understand the experiences of their populations, including immigrants. Many provinces have also looked into how criminal records can be used as well. To fully maximize the use of sub-national records, it is important to be able to link them to the PRLF or other national records, such as those on taxfilers, as done in the IMDB, to gain a full longitudinal portrait of immigrants. Key to being able to do that is creating common data protocols and standards. The greater the linkage to a common base, the stronger and more comprehensive a 'data spine,' the more complete a portrait of experience researchers and policy makers will get. These linkages have been done in British Columbia, Manitoba, Ontario, and New Brunswick, at the time of writing.

Similar to the situation with the PRLF and IMDB, accessing sub-national records requires security and sensitivity around privacy. Currently, the Canadian Institute of Health Research helps coordinate health data and acts as a hub for health data linkage. However, this has largely been done for epidemiological purposes, and considerations of how health data can be used to study migration and immigrant settlement has created many ethical and policy debates (McDonald et al., 2019). It has also led to much discussion on which is the best avenue for researchers to access the data and to work on procedures for access and vetting. Additionally, the use of health data for studying migration and immigration may require, in some jurisdictions, reform of privacy laws. Most provinces in Canada have separate legislation, in addition to regular privacy acts, to protect health information of individuals. For example, researchers in the Province of New Brunswick needed to create a 'Research Act' to allow Health records to be used for purposes other than their original intention. This is an issue that also affects other forms of new data derived from social media, apps, or information gathered from smartphones, which we discuss below. Linkage across administrative record also requires working out authority and ownership of data, given that it links provincial records to federal records and national and sub-national policies, practices, protocols and standards may not always align. Linkage of educational and other sub-national records is still in the early days and few provinces have developed means for linking that data to other sources yet, but we expect to see significant progress in these areas in the next 5 years and will continue to build off of the national administrative data that can act as spines for other data across the country.

Another issue with sub-national administrative data is the relative lack of development of some of these sources. They are largely raw records that were not created for the purpose of research and, thus, require significant cleaning and coordination before linkage to national data or cross province analysis can be done. This creates obstacles for sub-national governments, especially in smaller and secondary regions, where civil servants often do not have the mandate, skills, or time to turn administrative data into a research-ready platform. Although this is also true of national sources, the issue appears to be more prevalent at the lower levels of jurisdiction. This is also a problem encountered in other countries. Researchers and policymakers should, therefore, expect challenges when working with sub-national administrative data.

An area that has largely been unexamined, in terms of migration and immigrant settlement data, is information from municipalities, which are yet another level of sub-national data. As researchers and policy makers consider municipal records, one of the biggest obstacles is that, like other sub-national governments, most municipalities do not have the financial or human resources to process their administrative data. They in turn face challenges in linking it to provincial and national level data. Such data, like other administrative data, was not created with the purpose of studying migration and, as a result, also has obstacles in terms of access, shared standards, definitions, units of analysis, and providing adequate documentation that need to be smoothed if such data can be used for meaningful analysis. Despite these obstacles, according the United Nations (2018), over 55% of people worldwide live in cities, with major centers such as Toronto, London, New York or Berlin becoming global cities (Sassen, 2016), largely due to the flow of migration and immigrant settlement (Sanderson et al., 2015). Even smaller, secondary cities are increasingly linked to the world through immigration as well (Haan & Prokopenko, 2016). For these reasons, there is much opportunity in exploring municipal administrative records. It is on this front that the academic research community has much to offer, especially from those working in computer science and the social sciences.

8.3.3 Other Data Options

Researchers also have much to offer in terms of collecting their own surveys, as has been traditionally done, as well as using new sources of information to understand migration and immigrant settlement, such as mobile phone data or processing administrative records from service providers or other NGOs. These are far too numerous to fully enumerate in this chapter and the goal was not to offer a scoping review, but rather discuss how they fit in the data ecosystem and issues that are faced in different corners of it. So, here we offer some insight on additional options, rather than going into full detail on any specific data sources. We focus on surveys generated by the academic research community, Big Data, and NGO data.

Original surveys driven by researchers play an important role in filling gaps moved by those collected by national statistical agencies. They can explore topics that are not covered by their surveys and administrative data and are nimble in terms of development and delivery. One of the key obstacles with such work, however, is an inconsistency in how basic units of analysis, such as immigrant or refugee, as well as other parameters, are defined across studies (de Beer et al., 2010; Pritchard et al., 2019). This, in turn, makes comparison across studies difficult and, more importantly, as with inconsistency in government data sources, it makes longitudinal comparison difficult and also makes linking to national or international data sources next to impossible. Here again, investing in common data standards and protocols help leverage the data. Another obstacle with such initiatives is that many researcher-led surveys have small samples due to cost and other constraints, which make their power weak. It is, thus, important for researchers to consider how they can use

comparable questions to those of national and international instruments and how they can be linked to them through common geographies or other units. If they do so, they can tap into larger samples and generalize results beyond the confines of results harvested.

Policy makers and researchers are also turning to Big Data, or information that can be gathered from social media, apps, smartphones, and other technology (Jünger, 2019; Keusch et al., 2019). Such data have the potential for offering real-time analysis of migrants and offer a wide range of analyses. This all the while that such apps or technology can offer services to those who use them, such as hosting information or offering translation. Like administrative records, such harvested data will need to be seen in light of the original purpose of collection versus the use for studying migration and immigrant settlement. For example, an app that hosts immigrant settlement information may capture geo-location data to help provide relevant information to the user. Such data can also be used to study mobility patterns or be linked to other geo-spatial data, extending uses beyond the original intent.

To fully maximize such information, issues over data standards, definitions and protocol are important. They are all key to linking across data sources. Yet another consideration around using such data is ethics and how it complies with privacy laws in different jurisdictions (Scassa, 2019). Concerns, for instance, are already being raised in the European Union over governments using smartphone data to identify undocumented migrants and use the information as a weapon to deport refugees (Meaker, 2018). As with all data, in the wrong hands, such data could be used to the detriment of those most vulnerable. For this reason, national governments have an important role to play in creating common data practices and protocols that weigh the necessity for gathering private information and the proportionality in protecting privacy.

Settlement organizations and other NGOs also have access to information on those using their organization's services or their members. They also have data they must collect for their funders, as governments and other funders have demanded transparency and accountability of organizations. Such data has already been used by Immigration, Refugees and Citizenship Canada, through its Immigration Contribution Agreement Reporting Environment (iCARE), which accesses the data to evaluate and assess the settlement of Syrian refuges to the country (e.g. IRCC, 2019). These data have not yet been used extensively in Canada for research, but this will soon change now that iCARE has been linked to the IMDB. In addition to this, data organizations often collect their own information on programs and services of their clients and those data could be an important tool for researchers. However, like with other sources, such data suffer from a lack of consistency in how key concepts and units of analysis are defined, and most organizations lack the financial and human resources needed to process this information. This is an area where academic researchers can play a significant role alongside national governments and statistic agencies in helping coordinate the broader national and international ecosystems.

8.4 Lessons Learned on the Obstacles to Overcome and Opportunities to Pursue

Across each data source there is a common set of issues that can and need to be addressed by the policy and research community. They are issues that affect all data landscapes, not just the Canadian immigration data ecosystem. A key obstacle is the need to create common data standards that afford consistent measures that have common definitions (de Beer et al., 2010). As noted above, definitions of key terms, units of analysis, and geographies commonly change over time. Additionally, different datasets have their own definitions for core concepts like migrant, immigrant, refugee and so forth. A scoping review of literature showed that, in research on child and youth refugees, there were over 200 different groupings of age to define children and/or youth (Pritchard et al., 2019). Such inconsistencies limit the ability to link data and to offer robust comparative analysis, without requiring a considerable amount of background work prior to analysis.

Another obstacle to overcome is data access. This is particularly the case for confidential government microdata, administrative data and data built by individual researchers. In all cases, issues of privacy create barriers to access. Common protocols could be developed for accessing administrative data internationally, nationally and sub-nationally. For researchers collecting their own surveys or compiling their own data, using information repositories, or offering open access to replication datasets, needs to become the norm. This will enable using data across a wider range of users and potentially be linked to other sources, which can help bridge the gaps they may have because of small sample size. Creating greater access to data will foster greater engagement of migration and immigration issues.

A third obstacle to overcome is responsibility for developing new data. The various aspects of data development (creating and testing questionnaires, collecting and cleaning data, generating documentation and derived variables, among other considerations) has largely been the responsibility of national statistics agencies. Increasingly, with the advent of administrative data, other Federal or national departments are contributing to the data development process. This can include disclosing data, defraying the costs of development. It also means that national statistics agencies will increasingly play the role of providing technical assistance in addition to data collection. They are well positioned to generate and advocate for data standards across a data landscape. The sheer volume of potential new data sources is likely to require the hiring of many more data scientists across all levels of government. It will also require a greater engagement of academic institutions to assist with the Herculean task of shaping the new data landscape. Academic institutions can and should become more centrally involved in the data development process and help with the generation of common protocols. With greater access also comes the need for improved numeracy and development of the skills needed to use different levels and types of data, especially for those aiming to use administrative data.

Another challenge is that most datasets, especially administrative data, are created for one purpose in mind and applying them to issues of migration and immigration will stretch those purposes. To date, the focus of administrative data has been on economic issues (Costigan et al., 2016) and individuals (Ramos & Bennett, 2019). Such focus is largely because of constraints of the data, which do not measure a range of other issues. In other words, one can only answer the questions that are asked or measure things that have data collected on them. To overcome this obstacle, data linkage to censuses, surveys, and other administrative datasets has been key. We would argue that such linkage can also happen to yet other sources, if researchers, government departments and national statistical agencies adopt a data spine model that recognize the need for national data hubs and common practices.

8.5 A Call for Creating Spines and Data Standards

In place of a conclusion, we offer a call to policy makers and researchers to continue to work towards creating national data spines and common data standards and practices. We make this plea because we believe it is a huge opportunity to meet the challenges of the current migration and immigrant settlement data landscape and can lead to improved research and policy-outcomes. The notion of a national data spine is to create a core data infrastructure that can be used to link across national statistics, survey and administrative records, and sub-national statistics as well as research conducted by those in the academic, NGO, and private sectors. The main way forward in this regard is for national statistical agencies to work in partnership with stakeholders across the data ecosystem so that all data can be linked to administrative and census records. This will allow smaller and more nimble surveys as well as a series of administrative records across areas to be connected. This will mean that data standard will need to be developed as well as common procedures and protocols. Those will also help navigate the ethical, privacy, and other challenges that arise from unprecedented development and access to information (Scassa, 2019). This is underway, to some extent, in Canada through the Secure Data Linkage Environment. It is focused on creating a space where information from individuals across data sources can be combined. However, it still has much work to do in terms of creating common measures and practices. Moreover, it also can be seen in the rapid data collection during the COVID-19 outbreak. Statistics Canada generated crowdsourcing data, for the first time in its history, and worked to use questions similar to those being launched by NGOS, and in consultation with stakeholders, that could potentially be linked through geographic units to the agencies other data. We believe that coordination of such issues and data will form a national data spine that will allow for a more robust portrait of people, while also minimizing the cost of analysis. Doing so will take advantage of the nimbleness of non-government sources, as they will be able to link to it, allowing a fluid and dynamic data landscape. As well, if countries adopt a set of common protocols, it will allow for better international comparisons. This is already being done in the European Union

but could be extended to larger networks of countries for international or global standards. Perhaps one avenue would be to leverage international organizations, such as the United Nations, or trade pacts, such as the G7, or other unions, like NATO.

Similarly, it will be worth exploring common data management practices and legislative standards for ensuring ethical use of data and adequate and responsible protection of privacy. Doing so will allow for better linkage within countries as well as across them. Here, we believe national statistics agencies can play a key role in coordinating data ecosystems. The same role can be done transnationally through organizations like EUROSTAT or the OECD. In addition to focusing on creating, collecting and managing data, their mandates will also need to focus on helping promote the harmonization of goals and practices amongst researchers, and across all sectors. To do so will mean that such agencies, and the research community that works with them, will need to work to create new relationships, trust, and means of accessing data across the data system to promote fluid flow and exchange of information. This may seem lofty, however, the creation of common, or even international standards, in other sectors has sparked innovation and collaboration leading to better products and outcomes. We cannot see why doing the same with data construction and management would not have similar positive outcomes. Developing data spines and standards will truly tap into the extraordinary amount of data collected and offer stronger insights on the unprecedented migration and immigrant settlement the world is currently experiencing.

References

Akee, R., & Jones, M. R. (2019). *Immigrants' earnings growth and return migration from the US: Examining their determinants using linked survey and administrative data* (No. w25639). National Bureau of Economic Research. https://www.nber.org/papers/w25639

Allen, J., & Taylor, Z. (2018). A new tool for neighbourhood change research: The Canadian longitudinal census tract database, 1971–2016. *The Canadian Geographer, 62*(4), 575–588.

Australian Bureau of Statistics. (2011). *Guide to migrant statistical sources* (2nd ed.). Catalogue no. 3414.0. https://www.abs.gov.au/ausstats/abs@.nsf/Lookup/3414.0main+features22011%20(Edition%202)

Australian Bureau of Statistics. (2017). *Migration, Australia, 2015–16*. https://www.abs.gov.au/ausstats/abs@.nsf/previousproducts/3412.0main%20features32015-16

Aydemir, A., & Robinson, C. (2006). *Return and onward migration among working men* (Analytical studies branch research paper series). Statistics Canada: Catalogue no. 11F0019M1E(273).

Bender, S., Burghardt, A., & Schiller, D. (2014). *International access to administrative data for Germany and Europe*. Available at SSRN 2393357. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2393357

Borjas, G. J. (1993). Immigration policy, national origin, and immigrant skills. A comparison of Canada and the United States. In D. Card & R. Freeman (Eds.), *Small differences that matter* (pp. 21–43). University of Chicago Press.

Boyd, M. (1999). Canadian, eh? Ethnic origin shifts in the Canadian census. *Canadian Ethnic Studies Journal, 31*(3), 1–21.

Boyd, M., Goldmann, G., & White, P. (2000). Race in the Canadian census. In L. Driedger & S. Halli (Eds.), *Chap in race and racism: Canadian challenge 2000*. McGill-Queens University Press and Carleton University Press.

Careja, R., & Bevelander, P. (2018). Using population registers for migration and integration research: Examples from Denmark and Sweden. *Comparative Migration Studies, 6*(1), 1–27.

Chavez, B. (2019). *Immigration and language in Canada, 2011 and 2016* (Ethnicity, language and immigration thematic series). Statistics Canada, Ottawa, 89-657-X.

Costigan, C., Lehr, S., & Miao, S. (2016). Beyond economics: Broadening perspectives on immigration to Canada. *Canadian Ethnic Studies, 48*(1), 19–44.

de Beer, J., Raymer, J., Van der Erf, R., & Van Wissen, L. (2010). Overcoming the problems of inconsistent international migration data: A new method applied to flows in Europe. *European Journal of Population, 26*(4), 459–481.

Del Fava, E., Wiśniowski, A., & Zagheni, E. (2019). *Modelling international migration flows by integrating multiple data sources*. https://osf.io/preprints/socarxiv/cma5h/

Ernsten, A., McCollum, D., Feng, Z., Everington, D., & Huang, Z. (2018). Using linked administrative and census data for migration research. *Population Studies, 72*(3), 357–367.

Fong, E., & Shen, J. (2016). Participation in voluntary associations and social contact of immigrants in Canada. *American Behavioral Scientist, 60*(5–6), 617–636.

Fransoo, R. (In progress). *Profile of immigrant health status and health care use patterns*. Manitoba Centre for Health Policy.

Grieco, E. M., & Rytina, N. F. (2011). US. Data sources on the foreign born and immigration. *International Migration Review, 45*(4), 1001–1016.

Haan, M. (2007). The homeownership hierarchies of Canada and the United States: The housing patterns of white and non-white immigrants of the past thirty years. *International Migration Review, 41*(2), 433–465.

Haan, M. (2019). Surviving the next avalanche: Skills development and the brave new world of administrative data. *Canadian Issues Spring/Summer, 2019*, 7–10.

Haan, M., & Prokopenko, E. (2016). *Overview of secondary migration of immigrants to Canada. Pathways to Prosperity*. http://p2pcanada.ca/files/2016/02/Overview-of-Secondary-Migration-of-Immigrants-to-Canada.pdf

Haan, M., Arbuckle, J., & Prokopenko, E. (2017). Individual and community-level determinants of retention of Anglophone and Francophone immigrants across Canada. *Canadian Studies in Population, 44*(1–2), 59–76.

Hou, F., & Bonikowska, A. (2016). Selections before the selection: Earnings advantages of immigrants who were former skilled temporary foreign workers in Canada. *International Migration Review*.

Hou, F., Schellenberg, G., & Berry, J. (2018). Patterns and determinants of immigrants' sense of belonging to Canada and their source country. *Ethnic and Racial Studies, 41*(9), 1612–1631.

Houle, R. (2007). Secondary migration of new immigrants to Canada. *Our Diverse Cities Summer, 3*, 16–24.

Hulchanski, D. (2014). *It is better not to know than to know: The 2011 and 2016 National Household Survey (NHS)* http://neighbourhoodchange.ca/2011-and-2016-nhs/

Hum, D., & Simpson, W. (2004). Economic integration of immigrants to Canada: A short survey. *Canadian Journal of Urban Research, 13*(1), 46–61.

Hussen, H. A. (2018). *Annual report on immigration to parliament*, 2018. Immigration, Refugees, and Citizenship Canada, Ottawa. https://www.canada.ca/content/dam/ircc/migration/ircc/english/pdf/pub/annual-report-2018.pdf

IRCC (Immigration, Refugees and Citizenship Canada). (2019). *Syrian outcomes report*. https://www.canada.ca/en/immigration-refugees-citizenship/corporate/reports-statistics/evaluations/syrian-outcomes-report-2019.html

Jünger, S. (2019). *Using georeferenced data in social science survey research: The method of spatial linking and its application with the German general social survey and the GESIS panel*. PhD diss., DEU.

Kaida, L., Hou, F., & Stick, M. (2019). The long-term economic integration of resettled refugees in Canada: a comparison of Privately Sponsored Refugees and Government-Assisted Refugees. *Journal of Ethnic and Migration Studies*, 1–22.

Kaida, L., Ramos, H., Singh, D., & McLay, R. (2020). How to capture Neighborhood change in small cities. *Canadian Studies in Population*. https://doi.org/10.1007/s42650-020-00026-8

Keusch, F., Leonard, M. M., Sajons, C., & Steiner, S. (2019). Using smartphone technology for research on refugees: Evidence from Germany. *Sociological Methods & Research*, 0049124119852377.

Li, P. S. (2000). Earning disparities between immigrants and native-born Canadians. *Canadian Review of Sociology, 37*(3), 289–311.

Li, P. S. (2008). The role of foreign credentials and ethnic ties in immigrants' economic performance. *Canadian Journal of Sociology, 33*(2).

McDonald, J. T., Cruickshank, B., & Liu, Z. (2018). Immigrant retention in NB: An analysis using administrative Medicare Registry data. *Journal of Population Research, 35*(4), 325–341.

McDonald, T., Calhoun, A., & Haan, M. (2019). *Internal Mobility of Immigrants in Canada: A comparison of the strengths and weaknesses of alternative data sources*. Presented at the 2019 Annual General Meeting of the Canadian Economics Association, held in Banff, Alberta, May 31-June 2, 2019.

Meaker, M. (2018). *Europe is using smartphone data as a weapon to deport refugees*. Wired. https://www.wired.co.uk/article/europe-immigration-refugees-smartphone-metadata-deportations

Mervis, J. (2019). *Why the U.S. Census Bureau could have trouble complying with Trump's order to count citizens*. https://www.sciencemag.org/news/2019/09/why-us-census-bureau-could-have-trouble-complying-trump-s-order-count-citizens

Migration Data Portal. (2019). *International Organization for Migration, Global Migration Data Analysis Centre*. https://migrationdataportal.org/themes/migration-data-sources

Myers, D. (1999). Cohort longitudinal estimation of housing careers. *Housing Studies, 14*(4), 473–490.

Myers, D., & Lee, S. W. (1998). Immigrant trajectories into homeownership: A temporal analysis of residential assimilation. *International Migration Review, 32*(Fall), 593–625.

Nakhaie, R., & Wijesingha, R. (2015). Discrimination and health of male and female Canadian immigrant. *Journal of International Migration and Integration, 16*(4), 1255–1272.

Newbold, B. (2007). Secondary migration of immigrants to Canada: An analysis of LSIC wave 1 data. *The Canadian Geographer/Le Géographe canadien, 51*(1), 58–71.

Pottie, K., Ng, E., Spitzer, D., Mohammed, A., & Glazier, R. (2008). Language proficiency, gender and self-reported health. *Canadian Journal of Public Health, 99*(6), 505–510.

Pritchard, P., Maehler, D., Pötzschke, S., & Ramos, H. (2019). Integrating refugee children and youth: A scoping review of English and German literature. *Journal of Refugee Studies, 32*(SI1), i194–i208.

Radford, J., & Connor, P. (2019). *Canada now leads the word in refugee resettlement, surpassing the U.S*. Fact Tank: News in the Numbers. Pew Research Center. https://www.pewresearch.org/fact-tank/2019/06/19/canada-now-leads-the-world-in-refugee-resettlement-surpassing-the-u-s/

Ramos, H., & Bennett, M. (2019). Do immigrants who land in Atlantic Canada with Family stay? *Pathways to Prosperity*. http://p2pcanada.ca/wp-content/blogs.dir/1/files/2019/07/Do-Immigrants-with-Family-Stay-or-Leave.pdf

Ramos, H., & Yoshida, Y. (2011). *Why do recent immigrants leave Atlantic Canada*. Atlantic Metropolis Centre–Working Paper Series 20321.

Ramos, H., & Yoshida, Y. (2015). From away, but here to stay? Trends in why a cohort of recent immigrants left Atlantic Canada? In E. Tastsoglou, B. Cottrell, & A. Dobrowolsky (Eds.), *Warmth of the welcome: Is Atlantic Canada a home away from home for immigrants?* CBU press.

Reitz, J. G., & Banerjee, R. (2007). Belonging? Diversity, recognition and shared citizenship in Canada. In K. Banting, T. Courchene, & F. L. Siedle (Eds.), *Belonging? Diversity, recognition and shared citizenship in Canada* (pp. 1–57). McGill-Queens University Press.

Reitz, J. G., Banerjee, R., Phan, M., & Thompson, J. (2009). Race, religion, and the social integration of new immigrant minorities in Canada. *International Migration Review, 43*(4), 695–726.

Rogers, N., & McNally, J. (2018). *Using administrative data sources to improve our understanding of movements of international migrants within the UK. Conference of European Statisticians: Working Session on Migration Statistics* (Working paper 8). United Nations Economic Commission for Europe. Geneva Switzerland, October 24–26 https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.10/2018/mtg1/UK_Data_Integration_ENG.pdf

Ruotsalainen, K. (2011). *A census of the world population is taken every ten years*. Statistics Finland. https://www.stat.fi/tup/vl2010/art_2011-05-17_001_en.html

Sanderson, M. R., Derudder, B., Timberlake, M., & Witlox, F. (2015). Are world cities also world immigrant cities? An international, cross-city analysis of global centrality and immigration. *International Journal of Comparative Sociology, 56*(3–4), 173–197.

Sassen, S. (2016). *The global city: Strategic site, new frontier* (pp. 89–104). Managing Urban Futures, Routledge.

Scassa, T. (2019). As our economy becomes more data driven, Canadians need a national data strategy that encourages innovation and provides security and privacy. *Policy Options*. (January 15). https://policyoptions.irpp.org/magazines/january-2019/why-canada-needs-a-national-data-strategy/

Statistics Canada. (2007). *Longitudinal survey of immigrants to Canada: Detailed information for 2005 (Wave 3)*. https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=4422&lang=en&db=imdb&adm=8&dis=2#a4

Statistics Canada. (2016). *2016 census dictionary*. https://www12.statcan.gc.ca/census-recensement/2016/ref/dict/index-eng.cfm

Statistics Canada. (2019a). *General social survey: An overview, 2019*. https://www150.statcan.gc.ca/n1/pub/89f0115x/89f0115x2019001-eng.htm

Statistics Canada. (2019b) *Labour force survey*. https://www.statcan.gc.ca/eng/survey/household/3701

Statistics Canada. (2019c). *Longitudinal Immigration Database (IMDB): Interactive app*. https://www150.statcan.gc.ca/n1/pub/71-607-x/71-607-x2019003-eng.htm

Statistics Canada. (2021a). *2A questions and reasons why they are asked*. https://census.gc.ca/many-languages-nombreuses-langues/2a-questions-eng.htm

Statistics Canada. (2021b). *Topics covered by the 2021 Census*. https://www12.statcan.gc.ca/census-recensement/2021/ref/98-26-0001/2020001/004-eng.cfm

Statistics Canada. (2022). *Canadian community health survey – Annual component*. https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=3226

Stats, N. Z. (2019). *Longitudinal immigration survey: New Zealand*. http://archive.stats.govt.nz/browse_for_stats/population/Migration/lisnz.aspx

Temple, J. B., & McDonald, P. F. (2018). Australian migration propensities by visa class: An analysis of linked administrative data. *Journal of Population Research, 35*(4), 399–416.

Trebilcock, M. (2019). The puzzle of Canadian exceptionalism in contemporary immigration policy. *Journal of International Migration and Integration, 20*(3), 823–849.

United Nations. (2018). *2018 Revision of world urbanization prospects*. https://www.un.org/development/desa/publications/2018-revision-of-world-urbanization-prospects.html

United Nations. (2019). *A world on the move – UN DESA reveals latest data on international migration*. Department of Economic and Social Affairs. https://www.un.org/development/desa/en/news/population/international-migration-2.html

Urry, J. (2016). *Mobilities: New perspectives on transport and society*. Routledge.

VanderPlaat, M., Ramos, H., & Yoshida, Y. (2013). What do sponsored parents and grandparents contribute? *Canadian Ethnic Studies, 44*(3), 79–96.

Vigod, S. N., Arora, S., Urquia, M. L., Dennis, C. L., Fung, K., Grigoriadis, S., & Ray, J. G. (2019). Postpartum self-inflicted injury, suicide, assault and homicide in relation to immigrant status in Ontario: A retrospective population-based cohort study. *CMAJ Open, 7*(2), 227–235.

Warman, C., Sweetman, A., & Goldmann, G. (2015). The portability of new immigrants' human capital: Language, education, and occupational skills. *Canadian Public Policy, 41*(Supplement 1), S64–S79.

Wong, L. L., & Tézli, A. (2013). Measuring social, cultural, and civic integration in Canada: The creation of an index and some applications. *Canadian Ethnic Studies, 45*(3), 9–37.

Yoshida, Y., & Ramos, H. (2012). Destination rural Canada: An overview of recent immigrants to rural small towns. *Social Transformation in Rural Canada*, 67–87.

Yoshida, Y., & Ramos, H. (2013). Destination rural Canada: A basic overview of recent immigrants to rural small towns. In J. R. Parkins & M. G. Reed (Eds.), *The social transformation of rural Canada: New insights into culture, identity and collective action* (pp. 67–87). UBC press.

Yoshida, Y., & Ramos, H. (2017). *Demographic and economic profiles of immigrant Taxfilers to Atlantic Canada*. Perceptions of Change Project. http://perceptionsofchange.ca/demecoimmig.pdf

Yoshida, Y., Ramos, H., & VanderPlaat, M. (2016). How do the economic outcomes of economic versus family sponsored immigrants compare? *Canadian Diversity, 13*(1), 25–29.

Chapter 9
Assessing Transnational Human Mobility on a Global Scale

Emanuel Deutschmann, Ettore Recchi, and Michele Vespe

9.1 Introduction

Transnational mobility is the *sine qua non* of international migration. All international migrants have in common the basic fact of crossing (at least) one country border at some point of their migration trajectory. For a quantitative take on migration, thus, knowledge of the mobility flows of the world population amounts to a preliminary framing of global migration. Moreover, mobility data can contribute directly to understanding the scale of seasonal and other temporary forms of migration, which are hardly captured by official statistics (Gabrielli et al., 2019).

However, there is a surprising dearth of systematic information detailing the size of travel flows across countries worldwide. The Global Mobilities Project (GMP) at

An earlier version of this chapter was published as a working paper, which also contains an extensive appendix with additional information (Recchi et al., 2019a).

E. Deutschmann (✉)
University of Flensburg, Flensburg, Germany

Migration Policy Centre, Robert Schuman Centre for Advanced Studies, EUI, Florence, Italy
e-mail: emanuel.deutschmann@uni-flensburg.de

E. Recchi
Sciences Po, Observatoire Sociologique du Changement (OSC), CNRS, Paris, France

Migration Policy Centre, Robert Schuman Centre for Advanced Studies, EUI, Florence, Italy

Institut Convergences Migrations, Paris, France
e-mail: ettore.recchi@eui.eu

M. Vespe
European Commission, Joint Research Centre (JRC), Ispra, Italy
e-mail: michele.vespe@ec.europa.eu

S. Pötzschke, S. Rinken (eds.), *Migration Research in a Digitized World*, IMISCOE Research Series, https://doi.org/10.1007/978-3-031-01319-5_9

the European University Institute's Migration Policy Centre (MPC) aims to fill this gap by addressing different dimensions of transnational[1] mobilities (Recchi, 2017).

In this specific sub-project, we capitalize on two of the most comprehensive data sources on transnational human movements at a global scale:

1. Data on tourism, i.e., cross-border visits that include an overnight stay (*nota bene*: not necessarily for leisure), from the World Tourism Organization (UNWTO);
2. Data on cross-border air passenger traffic from Sabre, a private company that collects data directly from the airline industry.

Given that their data have been collected for different purposes, both sources, taken individually, have clear limitations when used in the attempt to provide insights into global human mobility. These limitations result in under-reporting of the scale of actual mobility across national borders. The data on tourism is incomplete in that people moving between countries for reasons other than tourism (in particular, returning residents) are not included. It is also distorted because visitors from some countries with few departures are not counted since their specific travel origin does not show up in the receiving country's tourism statistics. The data on air passenger traffic, in turn, does not factor in people who do not travel by airplane. In particular, journeys between neighboring countries, where cross-border mobility is particularly high (Deutschmann, 2016), are likely to be severely underestimated since people often use car, railway, or bus transportation rather than flights. We propose to remedy these systematic biases by combining and adjusting the two data sources, thereby producing more reliable estimates of cross-country human mobility globally. We describe the merging of these sources also as a possible precedent for similar endeavors for other types of country-to-country flows (like migration).

In the following sections, we firstly make general remarks about the composition of transnational mobility data in the two baseline sources and give an overview of the procedures followed to combine them (Sect. 9.2). We then describe these procedures in more detail in Sect. 9.3. Section 9.4 highlights some findings derived from the first explorations of the newly created dataset. In the conclusion (Sect. 9.5), we outline some pending limitations, advocate the use of this novel dataset to study transnational human mobility empirically in social science research and describe a set of general lessons from our project that might prove useful for other researchers embarking on similar endeavors.

[1] While we are aware that in the field of migration studies 'transnational' has a more demanding meaning that involves the regular movement of the *same* individuals across certain borders (Wimmer & Glick Schiller, 2002), we use the term 'transnational' in the meaning it has in the field of international relations, where it is employed to describe any movement by non-state actors that spans across national borders (Nye & Keohane, 1971).

9.2 Discerning the Composition of Transnational Mobility Flows

Our aim is to obtain robust estimates of the absolute number of yearly travels from and to every country worldwide. In formal terms, we set out to measure the volume of cross-border travels T across all pairs of sovereign states a, b, c, ... n on the planet. Such travels are carried out by both non-residents (NR) and residents (R) of receiving countries and take place by *air* (flights) or by *land/water* transportation (trains, buses, cars and other private road vehicles, boats, ferries and ships),[2] which we indicate by exponents ^{A} and ^{L}, respectively. Therefore:

$$T_{a\to b} = NR^{A}_{a\to b} + R^{A}_{a\to b} + NR^{L}_{a\to b} + R^{L}_{a\to b}$$

Unfortunately, no existing source contains information on all four components simultaneously. The original tourist files include only $NR^{A}_{a\to b} + NR^{L}_{a\to b}$, i.e., they register tourist *arrivals* in destination countries, but not tourists returning to their countries of origin.[3] Air traffic statistics include $NR^{A}_{a\to b} + R^{A}_{a\to b}$, i.e., air passengers only.[4] Thus, both datasets are suboptimal as they systematically exclude $R^{L}_{a\to b}$. Despite their differences, we expect the two datasets to be strongly correlated, because they share the same core component: $NR^{A}_{a\to b}$. They should diverge only when $R^{A}_{a\to b}$ and/or $NR^{L}_{a\to b}$ are large and/or not correlated.

The original UNWTO tourist files, however, also record residents of b going from b to *a with all transportation means*, that is $R^{A}_{b\to a}$ and $R^{L}_{b\to a}$. If we imagine that these people return to their country of residence in the same year of their outbound travel, we can count them as part of $R^{A}_{a\to b}$ and $R^{L}_{a\to b}$. We can thus assume that $R^{A}_{a\to b} + R^{L}_{a\to b} = R^{A}_{b\to a} + R^{L}_{b\to a}$. This assumption falls short of the travelers who: a) travel by the end of the year and come back in the following calendar year, or b) resettle abroad. As for a), we can maintain that these travelers are offset by similar travelers 12 months earlier. As for b), these travelers are migrants. A comparison of migration flows (in the most conservative estimate: Abel & Cohen, 2019, p. 8) and global tourist flows (in the conservative estimate of Deutschmann, 2016) shows a 1 to 98 relationship. That is, migrant travel corresponds to about 1% of tourist travels. Thus, 1% is the approximate maximum size of the error we introduce in our tourism

[2] Other statistically marginal forms of mobility (by foot or bike, for instance) are also included, provided they take place legally (i.e., they are registered). Unregistered or illegal border crossings are in fact left out by default from tourism and air traffic statistics, and, as a consequence, from our estimates.

[3] For a small number of receiving countries, returning residents are reported in the original UNWTO data, however without indication of where they are returning from. Thus, this information cannot be used in research interested in obtaining country-to-country flow estimates.

[4] Note that air traffic statistics do not allow us to distinguish between these two components since they are based on the location of the airport of origin and destination, not on the residence or nationality of travellers.

estimates through this assumption (see also Sect. 9.4). Conceptually, migration (be it voluntary or involuntary) is excluded from our estimates, even though we cannot rule out that some 'visitors' may overstay their travels and thus become migrants. More on this issue will be explored in the Conclusions (Sect. 9.5). We therefore revise the original UNWTO tourism data to build a yearly matrix of tourists/visitors travelling from a to b that also includes (returning) travellers from b who moved to a:

$$T_{a\to b}^{\text{revised}} = NR_{a\to b}^{A} + NR_{a\to b}^{L} + R_{b\to a}^{A} + R_{b\to a}^{L}$$

Hereafter, we will call this the GMP-revised tourism data [1]. Its creation is described in detail in Sect. 9.3.1.

As for the air passenger data, which we use in its KCMD-revised form [2] (see explanation below), we assume that they tend to be lower than the revised tourism data [1] because travelers also move by other means of transportation. However, [1] and [2] should converge progressively as the distance between origin and destination increases, given that air travel tends to become the exclusive means of transportation at long distances. This distance-mediated relationship between [1] and [2] leads us to transform the air passenger data. We compute an estimate of transnational mobility [3] that adjusts [2] by a factor that accounts for the distance between countries. The formal procedure to estimate [3] is described in Sect. 9.3.3.

In a final step, we combine the two revised sources, [1] and [3], to create an integrated dataset on global transnational mobility. As we hold that both [1] and [3] tend to underestimate actual mobility flows, our final estimate is always the largest of the two when we have both information—that is, either [1] or [3]. When we lack [3], we take [1], and vice versa.

Figure 9.1 provides an overview of this procedure. The individual steps are described in more detail in the following sections. The resulting final dataset covers 196 sender and receiving countries, generating a matrix of 38,220 cases (i.e., country pairs) per year. For the entire 2011–2016 period, about 9.5 billion trips (approx. 61%) are ultimately derived from [1] and 6 billion trips (approx. 38%) from [3]. Overall, 12.0% of cells are empty, which can mean either a total absence of transnational mobility between these countries (most likely in the case of pairs of small and distant nations) or missing data. The Global Transnational Mobility Dataset covers an estimated total of 15.7 billion trips.

9.3 Building the Dataset

In the following subsections, we outline in more detail how we handled the raw data and proceeded toward the production of the final Global Transnational Mobility Dataset. We first describe the creation of the GMP-revised tourism data (Sect. 9.3.1).

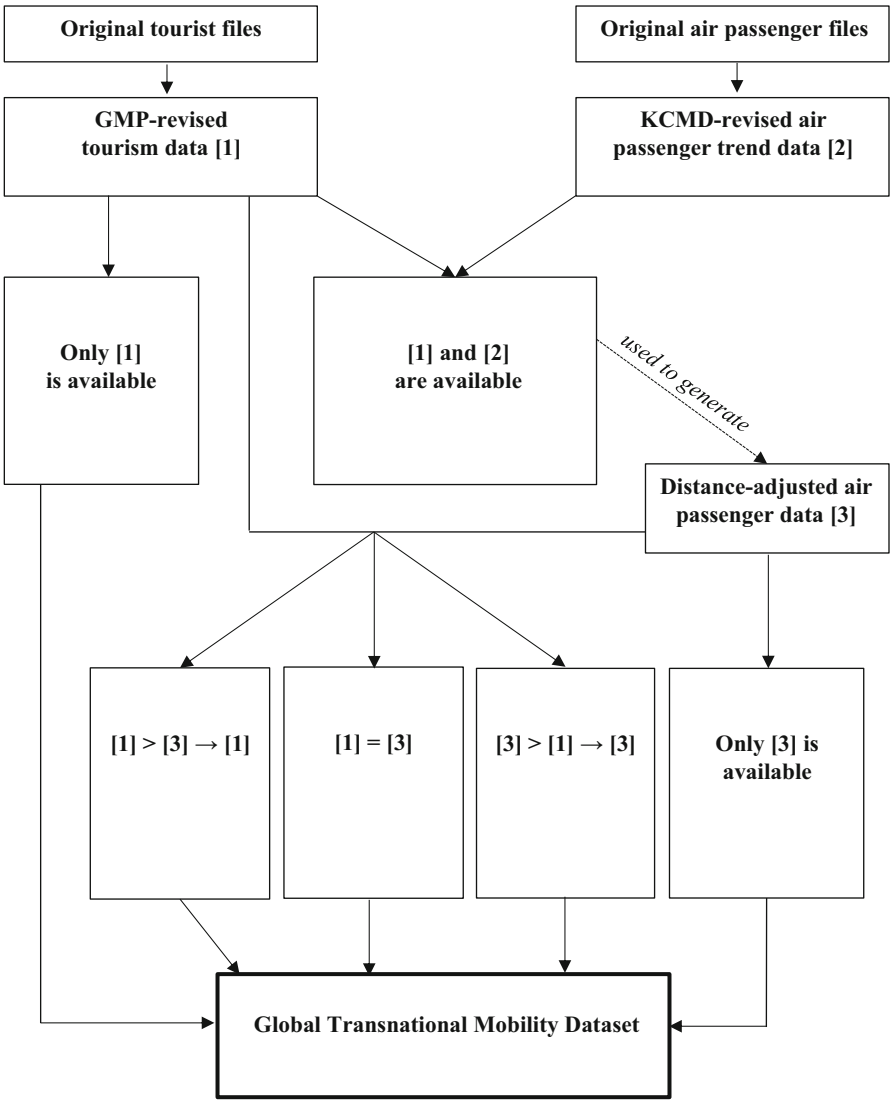


Fig. 9.1 Overview of the data composition

Second, we bring the KCMD-revised air passenger trend data in (Sect. 9.3.2). Third, we introduce the correction factor that adjusts the latter source, taking geographic distance into account (Sect. 9.3.3). Finally, we describe the merging and finalization of the dataset (Sect. 9.3.4).

9.3.1 Creating the GMP-Revised Tourism Data [1]

Our first source, the UNWTO tourism data, was obtained by the Global Mobilities Project (GMP) of the EUI's Migration Policy Centre (MPC) from the UNWTO as a set of files containing yearly flows from 1995 to 2016 for a global set of countries and territories worldwide (UNWTO, 2015).[5] While the harmonization and collection of national statistics on travels is part of the UNWTO mission, its online data are highly aggregated (see: https://www.unwto.org/unwto-tourism-dashboard, consulted December 18th, 2019). Therefore, we drew on the original country data kindly provided upon request by this organization. This dataset consists of 219 distinct files, one per receiving country/territory. To create a unified, standardized, and usable dataset (hereafter the GMP-revised tourism data), we took the following steps:

Step 1: Prioritizing the different UNWTO operationalizations of 'arrivals'

The country-to-country flow data on arrivals is reported in eight different categories in the UNWTO data (see Table 9.2 in the Appendix). The UNWTO defines arrivals—and describes its sources—as follows:

> *Arrivals* data measure the flows of international visitors to the country of reference: each arrival corresponds to one inbound tourism trip. If a person visits several countries during the course of a single trip, his/her arrival in each country is recorded separately. In an accounting period, arrivals are not necessarily equal to the number of persons travelling (when a person visits the same country several times a year, each trip by the same person is counted as a separate arrival).
>
> Arrivals data should correspond to *inbound visitors* by including both tourists and same-day non-resident visitors. All other types of travelers (such as border, seasonal and other short-term workers, long-term students and others) should be excluded, as they do not qualify as visitors. Data are obtained from different sources: administrative records (immigration, traffic counts, and other possible types of controls), border surveys or a mix of them. If data are obtained from accommodation surveys, the number of guests is used as estimate of arrival figures; consequently, in this case, breakdowns by regions, main purpose of the trip, modes of transport used or forms of organization of the trip are based on complementary visitor surveys. (UNWTO, 2015, p. 9).

To include as many cases as possible in the unified dataset, we use all eight 'arrivals' categories, in the order of preference shown in Table 9.2 in the Appendix.

Step 2: Creating a unified dataset

We then created a unified dataset that contains the relevant country-to-country flow data for all cases for which this information was available.[6] In doing so, we exclude

[5] At UNWTO, we thank Jacinta Mora for facilitating our access to these tourism statistics.

[6] There are 18 countries that are part of the UNWTO data collection that do not report arrivals by country of origin (which means they may be part of the full tourism dataset as senders of tourists but not as receivers): Afghanistan, Bonaire, Djibouti, Equatorial Guinea, Eritrea, Gabon, Ghana, Guinea-Bissau, Liberia, Libya, Mauritania, Saba, Sao Tome and Principe, Sint Eustatius, South Sudan, Syrian Arabic Republic, Turkmenistan, and United Arab Emirates.

several 'odd' sender categories, such as 'other countries of the world', which cannot readily be included in a country-to-country flow matrix. Details about this procedure and its consequences are described in Recchi et al. (2019a, Appendix).

Step 3: Adding returning residents

In line with the considerations made in Sect. 9.2, we add the returning residents $R^{A}_{b\to a} + R^{L}_{b\to a}$, to the incoming non-residents $NR^{A}_{a\to b} + NR^{L}_{a\to b}$ to obtain a more complete picture of human mobility across borders. In doing so, we effectively double the number of trips in the tourism dataset. Furthermore, the matrix becomes symmetric, i.e., mobility flows are now, by necessity, the same in both directions ($T^{\text{revised}}_{a\to b} = T^{\text{revised}}_{b\to a}$). Note that information is only added up if it was available in both directions. If one of the two values were missing (i.e., if information was available for the tie $a\to b$ but not for $b\to a$), the overall value was set to missing. This was done on the grounds that the overall information was considered unreliable when information in one direction was unavailable and that the other source (distance-adjusted air traffic data) is to be preferred.[7] After this step, we have obtained the GMP-revised tourism data [1].

9.3.2 Bringing in the KCMD-Revised Air Passenger Trend Data [2]

The second source is a dataset on global air passenger traffic collected by a private travel industry company, Sabre (2020). The dataset contains information on the total number of passengers flying between any two airports worldwide, regardless of whether the flights are direct or indirect. Here, we draw on a simplified and reduced version created by researchers at the European Commission's Knowledge Centre on Migration and Democracy (KCMD) that represents the yearly trend between countries (henceforth KCMD-revised air passenger trend data [2]). This version was generated through a time-series decomposition that dissects the raw overall air passenger flow between two countries into a trend component, a seasonal component, and a residual component (Gabrielli et al., 2019). In the KCMD-revised air passenger trend data [2] used here, the monthly trend data is aggregated to yearly averages. The data is available for the years 2011 to 2016.

We merge the two datasets [1] and [2] using ISO 3166-1 alpha-3 country codes. In line with the considerations made in Sect. 9.2, we hypothesize:

(a) [1] to be on average larger than [2], as it includes both air passengers and land/water travellers;

[7] We are indebted to Thomas Ginn for his close examination of our dataset, which led to the realization that this point was not clearly specified in a previous version of this manuscript.

(b) [1] and [2] to be highly correlated, since many travellers use flights to cross borders;
(c) [1] and [2] to be more strongly correlated as the distance between country pairs increases, since people are more likely to use air transportation at longer distances.

All three hypotheses hold empirically. As expected, tourism figures based on [1], where cross-border trips are reported with all transportation means, tend to be higher than air passenger figures based on [2], which report journeys that take place with flight transportation only. Table 9.1 shows the distribution of the deviations between the two data sources across cases (i.e., country pairs), by year. Negative values denote that there are more tourists than air passengers; positive values denote that there are more air passengers than tourists travelling between a pair of countries. The median (50th percentile) across years is −2410 trips, and even at the 75th percentile of cases, there are still more tourists than air passengers (−85 trips). Table 9.1 also reveals that, as the distribution is quite stable over time, the divergence between the two sources is no coincidence, but does indeed reflect the structural difference described above in hypothesis (a).

Figure 9.2 shows the relationship between the tourist-air passenger discrepancy and geographic distance (based on CEPII's GeoDist dataset [Mayer & Zignago, 2006]). A clear pattern emerges: there are only sizeable discrepancies at short geographic distances. The most extreme negative deviations (i.e., a lot more tourists than air passengers) are Hong Kong ↔ China (89–93 million, depending on year and direction), Macao ↔ China (37–43 million), United States ↔ Mexico (30–34 million), and Germany ↔ Poland (26–33 million). As Fig. 9.2 clearly shows, extreme cases consistently cluster together over time (different shapes represent different years). This suggests that these discrepancies are not random but systematic and meaningful. The inspection of specific cases with the highest negative[8] deviations helps to understand the rationales of the discrepancies, which can overlap and reinforce each other:

(a) Mobility between *nearby countries*: tourists exceed air passengers because many people move across borders with land (train, car, bus) or water (ferry, ship) transportation. Examples include the four extreme outlier country pairs tagged in Fig. 9.2.
(b) *Grand-tour tourism*: Here, people fly to one country (e.g., from the U.S. to the Netherlands), and then go by car or train to other countries (e.g., France). In these other countries, they are counted as tourists (e.g., through hotel registration data) but not as air passengers.

[8] In fact, there are few exceptional cases in which there are more air passengers than registered tourists. These are mostly distant country pairs with large contingents of migrants or returning nationals (who are not registered by tourism statistics) but relatively modest inflows of other visitors (e.g., India and Oman).

Table 9.1 Distribution of the difference between tourists and air passengers

Percentiles	2011	2012	2013	2014	2015	2016
Min	−89,300,000	−89,800,000	−89,400,000	−90,200,000	−92,400,000	−93,400,000
1%	−3,918,997	−4,064,395	−4,002,791	−4,361,469	−3,865,980	−4,136,718
5%	−514,371	−581,089	−661,828	−655,484	−569,920	−643,928
10%	−192,821	−212,287	−235,487	−232,265	−183,901	−218,354
25%	−22,009	−27,635	−30,651	−28,778	−24,436	−28,451
50%	−1997	−2536	−2924	−2493	−2189	−2323
75%	−63	−113	−126	−94	−56	−55
90%	1770	1220	998	1371	1480	4097
95%	11,824	10,775	8400	10,081	10,992	28,604
99%	131,253	140,405	109,720	113,494	140,005	257,340
Max	1,137,767	834,788	1,070,940	1,191,830	1,396,962	2,525,211
Obs.	5359	5771	5649	5653	5779	5262
Mean	−210,505	−219,209	−232,735	−232,250	−224,670	−243,573
Std. Dev.	2,175,910	2,132,248	2,178,926	2,221,919	2,251,131	2,393,686
Skewness	−30	−30	−28	−28	−29	−27
Kurtosis	1105	1131	1048	1020	1043	939

Note: Negative values indicate that there are more tourists than air passengers; positive values indicate that there are more air passengers than tourists travelling between a pair of countries

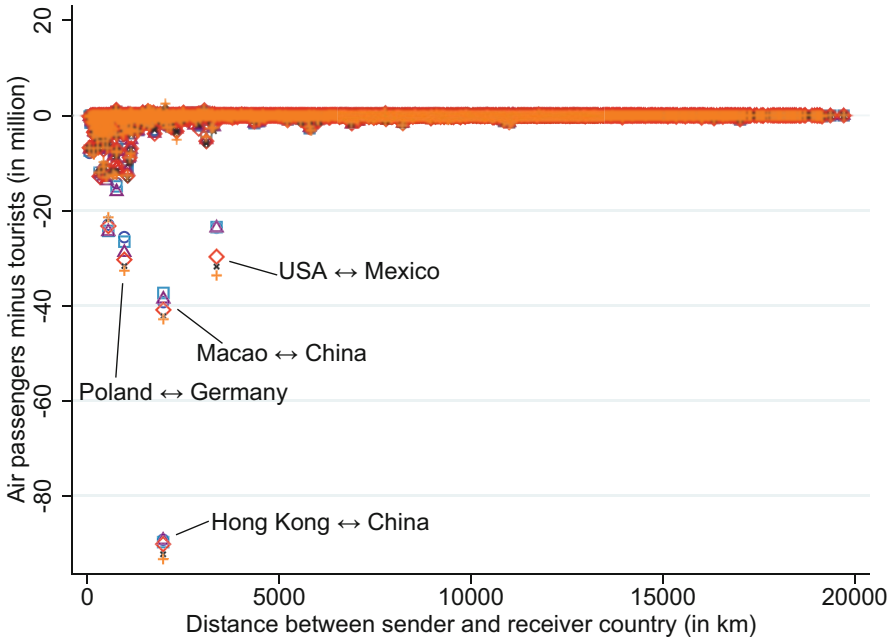


Fig. 9.2 The relation between geographic distance and divergences between the GMP-revised tourism dataset [1] and the KCMD-revised air passenger trend dataset [2]
Note: Different shapes denote different years. Distance is obtained from Mayer and Zignago (2006)

While rationale (b) is difficult to deal with but presumably marginal in statistical terms (see the remaining limitations described in Sect. 9.5), we treat rationale (a) by creating a correction factor that takes distance into account.

9.3.3 Creating the Distance-Adjusted Air Passenger Data [3]

The goal here is to adjust the KCMD-revised air passenger trend data [2] to correct for the fact that it underestimates mobility at short distances due to the use of alternative transportation means. To do so, we draw on the distance (in km) between country pairs. Our correction factor is specified as:

$$\left(\frac{k_{\max}}{k_{A\leftrightarrow B}}\right)^{1/c}$$

where $k_{\max}$ is the maximum possible distance between two countries, in this case 19,951.16 km (the distance between Paraguay and Taiwan), and $k_{A\ \leftrightarrow\ B}$ is the empirical distance between two countries A and B, based on CEPII's GeoDist dataset (Mayer & Zignago, 2006). The parameter c is chosen so that it maximizes the

correlation r between the GMP-revised tourism data [1] and the KCMD-revised air passenger trend data [2].[9] The rationale behind this is the assumption that [1] is not biased in terms of distance. Distance-adjusting [2] so that its correlation with [1] is maximized should thus lead to the best possible correction factor.

The result of this procedure is illustrated in Fig. 9.3a. After this adjustment, the correlation is $r(\max) = 0.7282$. Higher and lower c's lead to lower correlations. Figure 9.3b illustrates how the size of the resulting correction factor (based on the c that maximized the correlation) decreases as geographic distance increases between countries. The relationship resembles a fat-tailed power-law curve that is almost universally found to describe the spatial structure of human and animal mobility well (see Deutschmann, 2016 for an overview). Figure 9.3c shows the empirical distribution of resulting correction factors. For most cases, the correction is relatively small (correction factor < 1.5).

Figure 9.4 shows, on a log-log plot, how the GMP-revised tourism data [1] and the distance-adjusted air passenger data [3] relate to each other for all cases in which data from both sources is available. It reveals that, despite the distance-adjustments, the tourism data is still larger in about 70% of cases (i.e., more data points are located below the diagonal [solid line]). The adjustment can thus be considered conservative overall. The correlation is strong and clear, in line with hypothesis (b) in Sect. 9.3.2.

9.3.4 Creating the Global Transnational Mobility Dataset

In the final step, we merge the two revised data sources. As we hold that both the GMP-revised tourism data [1] and the distance-adjusted air passenger data [3] individually tend to under-estimate actual mobility flows (see Sect. 9.2), our final estimate is always the largest of the two when we have both kinds of information—that is, either [1] or [3]. When we lack [3], we take [1]; and vice versa. As final steps, we:

- Round decimals (non-integer estimates can occur due to the time-series decomposition applied by Gabrielli et al., 2019 and the correction factor introduced above).
- Add missing full country names and information on the world region a country is located in, based on the United Nations classification (drawing on Duncalfe, 2018).
- Exclude countries for which, after the merging procedure, no information was available.[10] Consequently, the dataset is reduced to the set of 196 countries used when creating the unified UNWTO dataset.

[9] We combine data from all available years and exclude cases with more than ten million trips to reduce the influence of these outliers on the calculations. On average, 31 cases are ignored per year (0.08% of the total).

[10] Countries and territories excluded are: Aruba, Anguilla, Cocos Islands, Cook Islands, Christmas Islands, Western Sahara, Falkland Islands, Faroe Islands, Guadeloupe, Grenada, Greenland, French

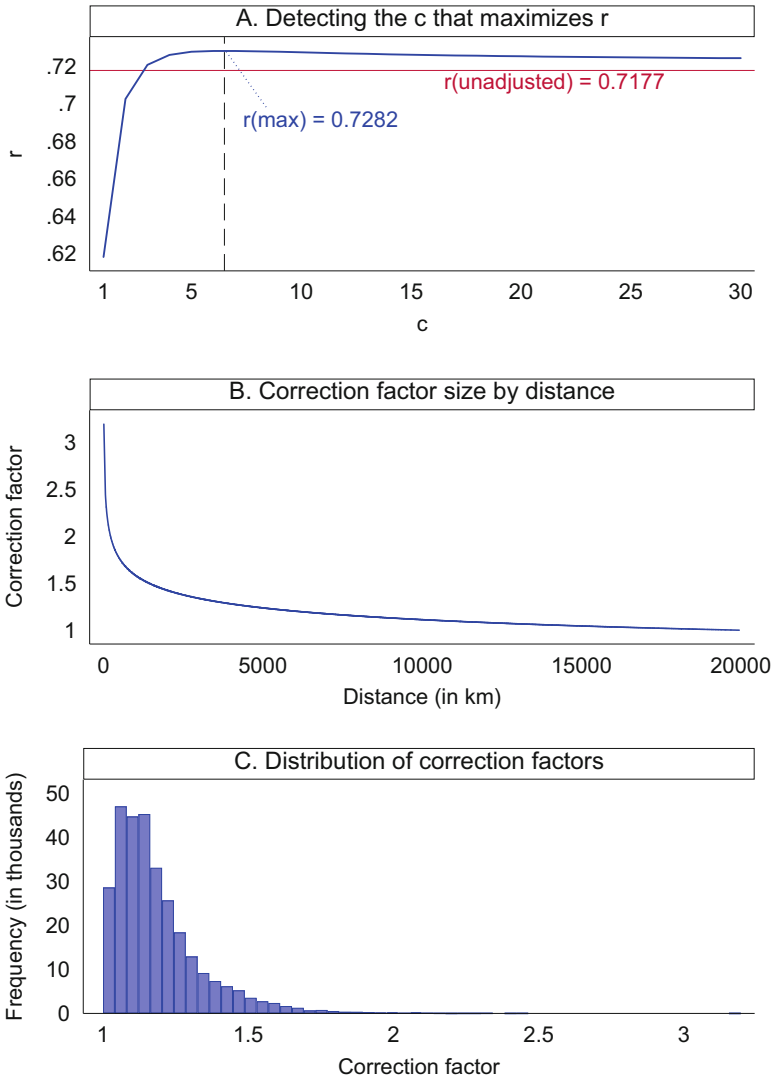


Fig. 9.3 Adjusting the distance-based correction factor for the KCMD-revised air passenger trend data to maximize the fit with the GMP-revised tourism data

The resulting Global Transnational Mobility Dataset can be explored on an interactive world map at the KCMD Dynamic Data Hub (https://bluehub.jrc.ec.europa.eu/migration/app/index.html; browse 'Datasets' – 'Mobility' – 'Global Transnational

Guiana, Montenegro, Northern Mariana Islands, Montserrat, Martinique, New Caledonia, Norfolk Islands, Pitcairn, Puerto Rico, French Polynesia, Reunion, Saint Helena, Saint Pierre and Michelon, Serbia, Tokelau, Taiwan, Wallis and Futuna Islands.

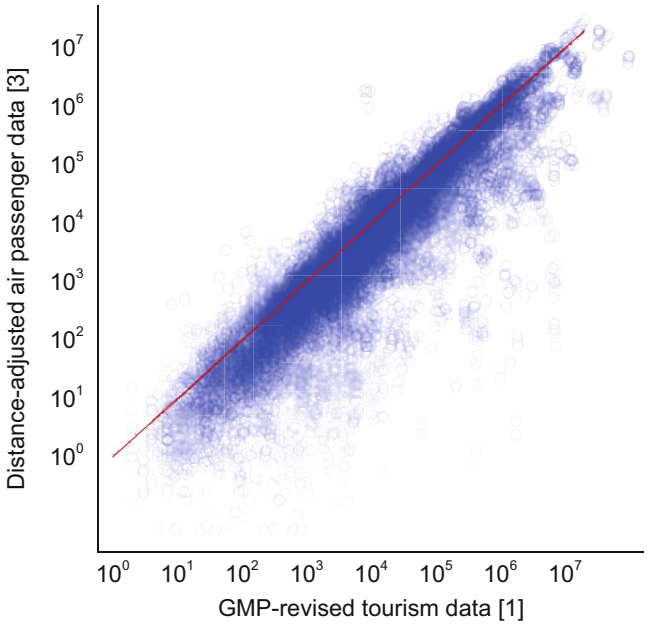


Fig. 9.4 The correlation between the distance-adjusted air passenger data [3] and the GMP-revised tourism data [1]

Mobility (KCMD-EUI)' – 'Estimated Trips'). More information can be found on the website of the Migration Policy Centre of the EUI (https://migrationpolicycentre.eu/projects/global-mobilities-project/), where the dataset can be downloaded. A list of the variables contained in the dataset can be found in the Appendix (Table 9.3).

9.4 Exploring the Dataset: Key Descriptive Findings

The Global Transnational Mobility Dataset covers 196 sender and receiving countries. Through the integration of two different sources, it is, to our knowledge, more comprehensive than all pre-existing information on worldwide cross-border mobility. Among other merits, its focus on transnational movements on a global scale helps to put migration in perspective, in both its geographical and demographic scope. The number of yearly migrant flows is very difficult to establish, and different alternative estimation methods have been proposed (Abel & Cohen, 2019; Abel & Sander, 2014; Azose & Raftery, 2019; Dennett, 2016). According to these methods, estimates range between 30 and 90 million migration episodes per year in the 2010–2015 period (Abel & Cohen, 2019, p. 8). Based on our new dataset, we estimate that, on average, about 2.55 billion yearly cross-border trips took place in the 2011–2015 period. Very crudely, thus, international migration episodes are

between 28 and 85 times (depending on migration estimates) less frequent than human movements across national borders in general. For specific regions, this ratio can be even higher. For example, in the European Union (for which actual yearly migration flow data is available), approximately 500–700 transnational trips occurred for every migratory move in 2016 (Deutschmann & Recchi, 2022).

While we leave to future research the full exploitation of the dataset's potential, also in conjunction with other datasets (not only on migration but also, for instance, on global trade, bilateral political relationships and many other potential predictors or predicted variables), the following pages offer a preliminary outline of several major takeaways.

9.4.1 *Worldwide Transnational Mobility Is Rapidly Increasing Over Time*

During the time frame under study, 2011 to 2016, transnational human mobility increased dramatically. In absolute terms, the number of estimated trips grew from about 2.3 billion in 2011 to about 2.9 billion in 2016. As Fig. 9.5a reveals, this growth is much larger than the growth in world population. This indicates that, collectively, humanity has indeed become more transnationally mobile. In this regard, transnational mobility is developing in line with cross-border communication, but in contrast to migration, which has not grown significantly faster than the world population (Czaika & De Haas, 2014; Deutschmann, 2021). This is also visible in Fig. 9.5b, which shows how, within the EU-28 (for which information on yearly migration flows is available), the number of transnational trips has grown much faster than both population and yearly migration flows (growth rates illustrated relative to the 2016 value). One important consequence of these diverging trends is that migration as a share of all transnational mobility is decreasing over time. In other words, temporary mobility has become more common relative to permanent migration.

The enormous increase in transnational mobility in a relatively short time frame raises questions on many grounds, like its environmental impact; its contribution to the spread of epidemics (Liu et al., 2020); its association with global systemic risks (Centeno et al., 2015); and, from a sociological perspective, social inequalities in access to these increased mobility opportunities. The latter issue is briefly touched upon in the following section.

9.4.2 *Transnational Mobility Tends to Cluster Within World Regions*

Figure 9.6a shows the mobility (in million trips) within world regions, using the United Nations M.49 Geoscheme as a base for assigning countries to regions. We

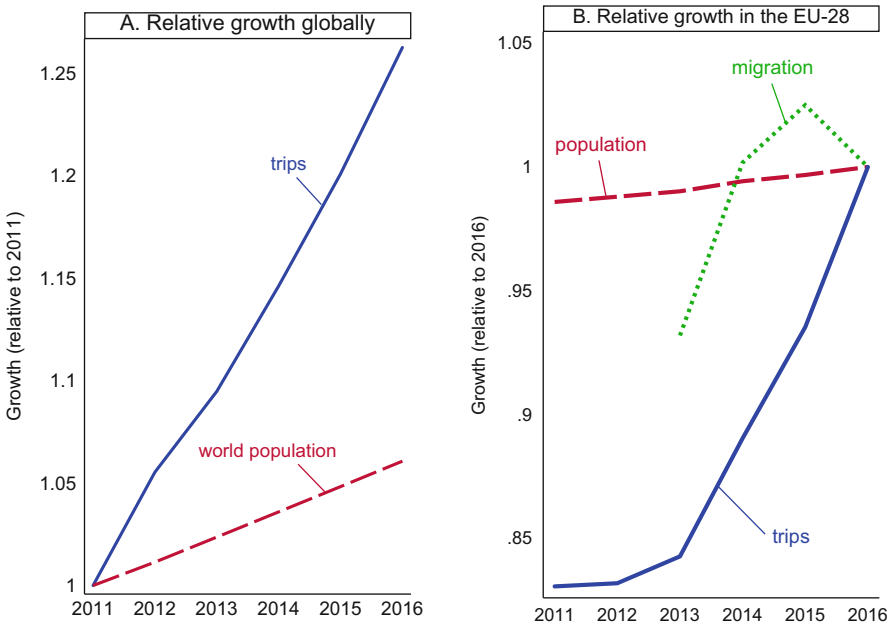


Fig. 9.5 Relative growth of mobility (and migration) globally and in the EU-28
Note: The graphs are based on the Global Transnational Mobility Dataset (trips), World Bank (2018) population data, and Eurostat migration data

find that Europe is the region with the highest number of intraregional trips, followed by Asia. The Americas are behind, and the smallest number of trips occur within Africa and Oceania.[11]

*Inter*regional mobility is far less common than intraregional mobility, with 80% of all mobility occurring within world regions in any given year (Deutschmann, 2020). However, there are differences between world regions in this regard (Fig. 9.6b): Intraregional mobility is more than five times more likely to occur than interregional mobility in the case of Europe; more than four times in the case of Asia; and almost three times in the case of the Americas. In the case of Africa, intraregional mobility is basically as likely as interregional mobility and in Oceania, intraregional mobility is half as likely as interregional mobility.

Note, however, that this comparison may be seen as 'unfair' since the pool of potential connections is obviously much larger in the case of interregional mobility

[11] Note that this simple measure may not be the best one to represent regional mobility. It is well possible that within Europe, for example, the high number of trips is driven by a subset of country pairs and that others participate very little in the intraregional network of transnational human mobility. Deutschmann (2021) proposes to use density-based measures as an alternative that allows to take into account between how many country pairs in a region meaningful amounts of mobility exist. Moreover, more sophisticated analyses would have to consider the varying population sizes of regions.

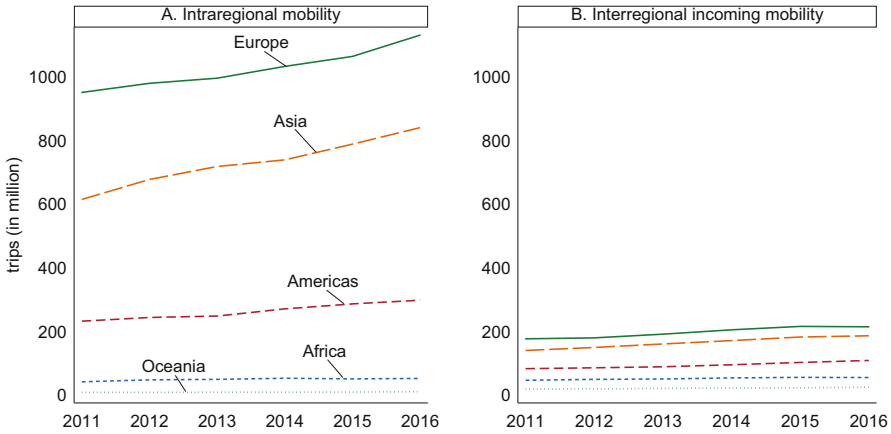


Fig. 9.6 Mobility within and between world regions

than in the case of intraregional mobility. A more sophisticated and 'just' comparison (which goes beyond the scope of this chapter) would be to compare intraregional mobility to mobility towards *specific* world regions. Past research has found that when this is done, mobility also tends to cluster within Africa and Oceania (Deutschmann, 2021).

In any case, Fig. 9.6a, b highlight the extreme stratification of opportunity to engage in transnational mobility at the global scale. Transnational mobility within Europe is about twenty times the amount of mobility within Africa, in spite of the much larger population of the latter continent. This global inequality in mobility chances has important sociological implications. For example, it has been shown that transnational human capital is an important resource that improves opportunities in life (Gerhards et al., 2017). Furthermore, transnational mobility shapes world views, attachment to other countries and cosmopolitan attitudes (Deutschmann et al., 2018; Helbling & Teney, 2015; Kuhn, 2015; Mau et al., 2008; Recchi, 2015). While these consequences of unequal access to transnational mobility chances have mainly been studied from a European viewpoint so far, a global perspective is largely missing. The Global Transnational Mobility Dataset may prove a good starting point for future analyses in this direction. The next section digs a little deeper into this global stratification by looking at the relationship between transnational human mobility and levels of prosperity.

9.4.3 Transnational Mobility Differs by Levels of Prosperity and Country Size

There is a relatively strong and significant relationship between a country's number of outgoing trips and the national level of prosperity, measured as GDP per capita in

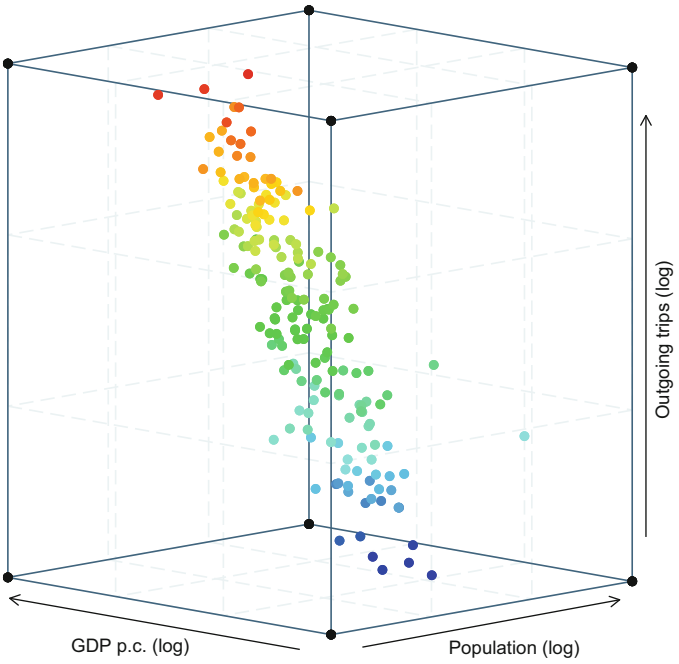


Fig. 9.7 The relation between mobility, population size, and GDP per capita

purchasing power parity based on World Bank data ($r = .63$). A similar pattern is found for the relationship between mobility and population size ($r = .58$). The three-dimensional graph in Fig. 9.7 illustrates the relationship between the three factors in combination. The distribution of dots, representing countries, follows a clear pattern, ranging from low GDP, small population and low mobility (bottom front corner) to high GDP, large population and high mobility (upper back corner). These insights are not entirely new but are showcased in a clear and robust way by this novel dataset. Future research may engage in more complex analyses, taking a larger set of factors into account and building more comprehensive multivariate models to study the antecedents and consequences of transnational human activity worldwide (see also Recchi et al., 2019b).

9.5 Discussion

A spate of migration and asylum-seeking crises has been hitting the world since the turn of the twenty-first century. The globe is on the move but, in spite of their salience in the media and public opinion, refugees and other migrants constitute only a tiny portion of the whole number of people crossing borders daily. According to various estimates, there were between 30 and 90 *million* migration episodes per year

in the early 2010s worldwide (Abel & Cohen, 2019). But according to our estimate, yearly border-crossings come close to 3 *billion* globally. By providing estimates of the amount of such transnational mobility beyond migration, the Global Transnational Mobility Dataset facilitates the study of the volume, directions and change of country-to-country human mobility on a worldwide scale.

This chapter described the procedures by which we have reached these estimates. While there is no single existing data source providing exact information on the number of people crossing national borders worldwide, we have argued that the two more complete and reliable sources (data on tourism and data on air passengers) show significant consistency and can be merged according to a few relatively simple combination rules.

We hope that our work will prove useful in two regards. First, we hope that the freely available GMP Global Transnational Mobility Dataset will be used to tackle questions related to mobility at the global scale. Potential applications are manifold and range from transnational mobility's unequal global structure and its social consequences to analyses that use the data to model the spread of infectious diseases. A first external study has already leveraged the dataset to model the spread of Covid-19 (Liu et al., 2020). Second, we hope that some aspects of our methodological approach can be transferred to other instances where researchers may consider merging two data sources, for instance, where regional migration data is available from several sources.

By focusing on yearly country-to-country flows of human *mobility* (whatever their duration), our dataset complements estimates of worldwide *migration* flows which refer to stays abroad longer than 12 months based on the conventional UN definition of migration. This dataset also improves upon previous usages of the UNWTO data (Deutschmann, 2016, 2021; Reyes, 2013) by capitalizing on an additional source and estimation methods. Finally, the Global Transnational Mobility Dataset parallels recent alternative attempts at measuring population mobility with digital sources (Fiorio et al., 2017; Hawelka et al., 2014; Messias et al., 2016; Rango & Vespe, 2017; Spyratos et al., 2018, 2019; State et al., 2013; Zagheni et al., 2017). Data triangulation across our data and digital estimates may prove useful to test the comparability of outcomes obtained through such different approaches.

Several important limitations remain. The first issue concerns the existence of grand-tour tourism and open-jaw flights. For instance, consider a traveler who goes on a round trip to Southeast Asia from Italy. She flies from Rome to Bangkok both on her way in and out and takes buses or rents a car to travel subsequently through Thailand, Vietnam, Laos, and Cambodia, before returning to Thailand to take her flight back home. According to the original UNWTO tourism data, there would be four trips: ITA → THA, ITA → VNM, ITA → LAO, and ITA → KHM. According to the GMP-revised tourism data [1], there would be eight trips: ITA → THA, THA → ITA, ITA → VNM, VNM → ITA, ITA → LAO, LAO → ITA, ITA → KHM, and KHM → ITA. According to the air passenger data (regardless of distance-adjustment), there would be two trips: ITA → THA, THA → ITA. In reality, however, there were six trips: ITA → THA, THA → KHM, KHM → VNM, VNM → LAO, LAO → THA, and THA → ITA. In this case, both sources and all

strategies lead to very different outcomes and none of them captures the transnational mobility that actually took place. This issue has no easy solution. Structurally, it should lead to a slight overestimation of long-distance mobility between world regions (which is most likely when such roundtrips are prone to occur). However, we argue that, compared to all global travels, these kind of journeys are rare and should not jeopardize the overall reliability of the dataset.

A second limitation consists of the following: by basing a substantial part of our mobility estimates on visitors who stayed overnight ('tourists' in the UNWTO terminology), we may be underestimating short-term border crossings. For instance by commuters who live in border regions and regularly go to the other side for work, leisure, or shopping. The following example is revealing in this regard: For the US, detailed data on land-border crossings are available (US Department of Transportation, 2018). Looking at mobility between the US and Canada, the distance-adjusted air passenger data (see Sect. 9.3.3) estimates about 20 million trips, while the GMP-revised tourism data (see Sect. 9.3.1) suggests around 33 million trips. The recorded land-border crossings, by contrast, are 103 million—98 million private car passengers alone. Many of these moves are not likely to be overnight stays. While it is hard to generalize from this example, it suggests that the mobility estimates in the Global Transnational Mobility Dataset (and the correction factor introduced in Sect. 9.3.3), although considerably larger than those provided by alternative global sources, are still quite conservative.

Finally, it is important to keep in mind that the Global Transnational Mobility Dataset contains mobility *estimates* rather than counts of actual, recorded trips. This is crucial. By applying a *statistical* approach to correct and adjust the data, we aimed to create a revised dataset that *on average* captures mobility between countries more accurately. In a minority of individual cases, this revision procedure might, however, lead to more inaccurate estimates. We would therefore like to remind that this dataset is well-suited to study structural features of transnational human mobility globally or for aggregates of countries. If the research interest is mobility between specific pairs of countries, the estimates in the Global Transnational Mobility Dataset should be taken with caution. Readers need to be aware of this limitation and should possibly compare the estimates to figures provided by alternative sources.

With these caveats in mind, we hope that this novel dataset will prove to be a valuable resource for all researchers interested in studying the human side of globalization. More particularly, this dataset can help embed migratory movements into the larger picture of transnational human mobility and better eschew the 'settlement bias' (Hugo, 2014) that recurrently weakens traditional migration studies. Attention to transnational mobility is especially needed to take into account less traditional and more reversible forms of migration (temporary, circular, shuttle, etc.). Also, it is needed more generally to remind us that international migrants are first of all people who cross borders – and therefore part and parcel of a mobile world.

Finally, there are several general lessons we want to share with readers who might be interested in following a similar merging strategy based on different datasets:

1. *Automatize!* The more the combination procedure is automatized, the easier it is to update datasets in the future. In our case, both the UNWTO and Sabre continuously update their datasets and it would be desirable to be able to quickly expand the time frame of the Global Transnational Mobility Dataset as these updates become available. Despite the availability of monthly air passenger volume projections for 2020 (Iacus et al., 2020), we were unfortunately not able to automatize the whole process (partly due to point [2] below). However, scholars interested in conducting a similar project with other sources are advised to automatize as much as possible to increase efficiency and facilitate future updates.
2. *Standardize!* One of the most time-consuming issues in combining mobility data from different sources is to bring the datasets into a mergeable format. One common obstacle occurs when only idiosyncratic country names rather than standardized country codes are available. For example, rather than using the standardized code COD, sources often only contain idiosyncratic names such as "Congo, Democratic Republic of", "D.R. Congo", "DR Congo", or "Congo, DR". We therefore appeal to all data-collecting organizations and individuals who publish such data to use standardized formats such as ISO 3166-1 alpha-3 country codes. Doing so will increase the potential for automatization and thereby increase efficiency. In our case, the UNWTO data files contain *numeric* ISO codes (e.g., "180" for D.R. Congo) while the air traffic data used the three-letter version (e.g., "COD"). While the conversion between the two is of course possible, it still constitutes one additional step that could be avoided by consistent usage of the more intuitive letter version. A further obstacle are changes in ISO country codes, such as the switch from ROM to ROU in the case of Romania due to a 2002 administrative decision. Such little changes often prevent flawless merging and lead to costly manual inspections. While we appreciate all existing standardization efforts, we believe there is still room for improvement.
3. *Annotate!* Good documentation is important, and we recommend annotating every single step in the procedure as clearly as possible and early on to increase inter-individual transparency.
4. *Keep it simple!* Several technically more sophisticated methods (e.g., multiple imputation) turned out not to lead to any useful information. We therefore developed the more straightforward correlation-maximizing approach presented above and the simple set of rules for combining the two sources. Often, in our view, it makes sense to stick to the classic KISS principle – Keep it simple, stupid!
5. *Globalize!* It is generally a good idea to start with the most comprehensive coverage and only drop data later on. It is always possible to make the dataset smaller but a lot more difficult to make it larger again.
6. *Be cautious!* In our view, merging data from different sources can bring advantages – and we see a lot of benefits of the GMP Global Transnational Mobility Dataset – but it may also carry risks, as already emphasized above. After all, different datasets are usually collected with different purposes in mind and are often based on different definitions and collection procedures. It is therefore

important to keep the resulting limitations in mind and reflect carefully on whether or not a certain combined dataset is well-suited for a specific research goal.

We hope that these recommendations can help other researchers who plan to embark on similar endeavors.

Appendix

Table 9.2 Categories in the UNWTO dataset

Code	Description	Preference
112	Arrivals of non-resident tourists at national borders, by country of residence	1st
111	Arrivals of non-resident tourists at national borders, by nationality	2nd
122	Arrivals of non-resident visitors at national borders, by country of residence	3rd
121	Arrivals of non-resident visitors at national borders, by nationality	4th
1912	Arrivals of non-resident tourists in all types of accommodation establishments, by country of residence	5th
1911	Arrivals of non-resident tourists in all types of accommodation establishments, by nationality	6th
712	Arrivals of non-resident tourists in hotels and similar establishments, by country of residence	7th
711	Arrivals of non-resident tourists in hotels and similar establishments, by nationality	8th

Table 9.3 Variables contained in the Global Transnational Mobility Dataset

Name	Description
source_name	Name of the country of origin
target_name	Name of the country of destination
source_iso3	ISO 3166-1 alpha-3 code of the country of origin
target_iso3	ISO 3166-1 alpha-3 code of the country of destination
year	Year, ranges from 2011 to 2016
estimated_trips	Estimated trips
dist	Geographic distance
source_region	Region of the country of origin
target_region	Region of the country of destination
source_subregion	Sub-region of the country of origin
target_subregion	Sub-region of the country of destination

Note: Geographic distance is obtained from CEPII's GeoDist dataset (Mayer & Zignago, 2006). Regions and subregions are based on the UN M.49 GeoScheme

References

Abel, G. J., & Sander, N. (2014). Quantifying global international migration flows. *Science, 343*(6178), 1520–1522.

Abel, G. J., & Cohen, J. E. (2019). Bilateral international migration flow estimates for 200 countries. *Scientific Data, 6*(1), 1–13.

Azose, J. J., & Raftery, A. E. (2019). Estimation of emigration, return migration, and transit migration between all pairs of countries. *Proceedings of the National Academy of Sciences, 116*(1), 116–122.

Centeno, M. A., Nag, M., Patterson, T. S., Shaver, A., & Windawi, A. J. (2015). The emergence of global systemic risk. *Annual Review of Sociology, 41*, 65–85.

Czaika, M., & De Haas, H. (2014). The globalization of migration: Has the world become more migratory? *International Migration Review, 48*(2), 283–323.

Dennett, A. (2016). *Estimating an annual time series of global migration flows–an alternative methodology for using migrant stock data* (pp. 125–142). Global Dynamics: Approaches from Complexity Science.

Deutschmann, E. (2016). The spatial structure of transnational human activity. *Social Science Research, 59*, 120–136.

Deutschmann, E. (2020). Visualizing the regionalized structure of mobility between countries worldwide. *Socius, 6*, 1–3.

Deutschmann, E. (2021). *Mapping the transnational world: How we move and communicate across borders, and why it matters*. Princeton University Press.

Deutschmann, E., Delhey, J., Verbalyte, M., & Aplowski, A. (2018). The power of contact: Europe as a network of transnational attachment. *European Journal of Political Research, 57*(4), 963–988.

Deutschmann, E., & Recchi, E. (2022). Europeanization via transnational mobility and migration. In M. Eigmüller, S. Büttner, & S. Worschech (Eds.), *Sociology of Europeanization*. De Gruyter Oldenbourg, 283–306.

Duncalfe, L. (2018). *ISO-3166 country and dependent territories lists with UN regional codes.* Available at: https://github.com/lukes/ISO-3166-Countries-with-Regional-Codes. Last accessed 08 Jan 2019.

Fiorio, L., Abel, G., Cai, J., Zagheni, E., Weber, I., & Vinué, G. (2017). Using twitter data to estimate the relationship between short-term mobility and long-term migration. *Proceedings of the 2017 ACM on Web Science Conference*, 103–110.

Gabrielli, L., Deutschmann, E., Natale, F., Recchi, E., & Vespe, M. (2019). Dissecting global air traffic data to discern different types and trends of transnational human mobility. *EPJ Data Science, 8*. https://doi.org/10.1140/epjds/s13688-019-0204-x

Gerhards, J., Hans, S., & Carlson, S. (2017). *Social class and transnational human capital. How middle and upper class parents prepare their children for globalization*. Routledge.

Hawelka, B., Sitko, I., Beinat, E., Sobolevsky, S., Kazakopoulos, P., & Ratti, C. (2014). Geo-located twitter as proxy for global mobility patterns. *Cartography and Geographic Information Science, 41*(3), 260–271.

Helbling, M., & Teney, C. (2015). The cosmopolitan elite in Germany. Transnationalism and postmaterialism. *Global Networks, 15*(4), 446–468.

Hugo, G. (2014). A multi sited approach to analysis of destination immigration data: An Asian example. *International Migration Review, 48*(4), 998–1027.

Iacus, S. M., Natale, F., Santamaria, C., Spyratos, S., & Vespe, M. (2020). Estimating and projecting air passenger traffic during the COVID-19 coronavirus outbreak and its socio-economic impact. *Safety Science, 104791*.

Kuhn, T. (2015). *Experiencing European integration: Transnational lives and European identity*. Oxford University Press.

Liu, Q., Liu, Z., Zhu, J., Zhu, Y., Li, D., Gao, Z., … Wang, Q. (2020). Assessing the global tendency of COVID-19 outbreak. *MedRxiv*. https://doi.org/10.1101/2020.03.18.20038224

Mau, S., Mewes, J., & Zimmermann, A. (2008). Cosmopolitan attitudes through transnational social practices? *Global Networks, 8*(1), 1–24.

Mayer, T., & Zignago, S. (2006). *GeoDist: The CEPII's distances and geo-graphical database* (MPRA paper no. 31243).

Messias, J., Benevenuto, F., Weber, I., & Zagheni, E. (2016). From migration corridors to clusters: The value of Google+ data for migration studies. In *Proceedings of the 2016 IEEE/ACM international conference on advances in social networks analysis and mining* (pp. 421–428).

Nye, J. S., & Keohane, R. O. (1971). Transnational relations and world politics: An introduction. *International Organization, 25*(3), 329–349.

Rango, M., & Vespe, M. (2017). *Big data and alternative data sources on migration: From case-studies to policy support. Summary report*. Joint Research Centre of the European Commission.

Recchi, E. (2015). *Mobile Europe: The theory and practice of free movement in the EU*. Palgrave Macmillan.

Recchi, E. 2017. *Towards a global mobilities database: Rationale and challenges.* Explanatory Note. MPC/EUI. Available at: http://www.migrationpolicycentre.eu/docs/GMP/Global_Mobilities_Project_Explanatory_Note.pdf. Last accessed 3 Mar 2019.

Recchi, E., Deutschmann, E., & Vespe, M. (2019a). *Estimating transnational human mobility on a global scale* (Robert Schuman Centre for Advanced Studies / Migration Policy Centre WP 2019/30). European University Institute.

Recchi, E., Deutschmann, E., & Chabriel, M. (2019b). *The global network of transnational mobility*. N-IUSSP, October. http://www.niussp.org/article/the-global-network-of-transnational-mobilityle-reseau-mondial-de-mobilite-transnationale/

Reyes, V. (2013). The structure of globalized travel: A relational country-pair analysis. *International Journal of Comparative Sociology, 54*(2), 144–170.

Sabre. (2020). *Market intelligence global demand data*. http://www.sabreairlinesolutions.com/home/software_solutions/airports/. Last accessed 28 Jun 2020.

Spyratos, S., Vespe, M., Natale, F., Weber, I., Zagheni, E., & Rango, M. (2018). *Migration data using social*. A European Perspective. JRC Technical Report. https://doi.org/10.2760/964282

Spyratos, S., Vespe, M., Natale, F., Weber, I., Zagheni, E., & Rango, M. (2019). Quantifying international human mobility patterns using Facebook Network data. *PLoS One, 14*(10). https://doi.org/10.1371/journal.pone.0224134

State, B., Weber, I., & Zagheni, E. (2013). Studying inter-national mobility through IP geolocation. In *Proceedings of the sixth ACM international conference on web search and data mining* (pp. 265–274).

United Nations World Tourism Organization (UNWTO). (2015). *Methodological notes to the tourism statistics database*. UNWTO.

U.S. Department of Transportation. (2018). *Border crossing entry data*. Available at: https://data.transportation.gov/Research-and-Statistics/Border-Crossing-Entry-Data/keg4-3bc2. Last accessed 9 Jan 2019.

Wimmer, A., & Glick Schiller, N. (2002). Methodological nationalism and beyond: Nation-state building, migration and the social sciences. *Global Networks, 2*(4), 301–334.

World Bank. (2018). *Population, total*. Available at: https://data.worldbank.org/indicator/SP.POP.TOTL. Last accessed 9 Jan 2019.

Zagheni, E., Weber, I., & Gummadi, K. (2017). Leveraging Facebook's advertising platform to monitor stocks of migrants. *Population and Development Review, 43*(4), 721–734.

Chapter 10
Google Trends as a Tool for Public Opinion Research: An Illustration of the Perceived Threats of Immigration

Reilly Lorenz, Jacob Beck, Sophie Horneber, Florian Keusch, and Christopher Antoun

10.1 Introduction

Traditionally, social science researchers have relied on surveys to produce population-level estimates of public opinion and behavior. However, surveys are not always feasible, and they generally require substantial time, effort, and money. Search queries made on Google's search engine, by contrast, can be obtained in aggregate form for free from the website Google Trends (GT). Consequently, researchers are starting to construct population measures based on these data. The research process typically involves selecting the keywords intended to measure a particular construct of interest, and then using GT to extract an estimate of the volume of Google search queries made, containing one or more of these keywords, in a particular time and place (Salganik, 2019). Unfortunately, this process yields measures that are of unknown accuracy.

In this chapter, we describe the fundamental ways that Internet search data differ from surveys, with a focus on the characteristics that could affect population estimates. Then, we describe a case study that empirically evaluates a measure of

R. Lorenz
Life Science Alliance, Heidelberg, Germany
e-mail: r.lorenz@life-science-alliance.org

J. Beck
Ludwig Maximilian University of Munich, Munich, Germany
e-mail: jacob.beck@stat.uni-muenchen.de

S. Horneber · F. Keusch (✉)
University of Mannheim, Mannheim, Germany
e-mail: shornebe@mail.uni-mannheim.de; f.keusch@uni-mannheim.de

C. Antoun
University of Maryland, College Park, MD, USA
e-mail: antoun@umd.edu

S. Pötzschke, S. Rinken (eds.), *Migration Research in a Digitized World*, IMISCOE Research Series, https://doi.org/10.1007/978-3-031-01319-5_10

the salience of the perceived threats of immigration in Germany based on Google Trends (GT) data. We conclude with a discussion of the issues that our case study raises with respect to the potential advantages and disadvantages of using GT data for social science research.

10.2 Google Trends as a Research Tool

Google Trends (https://trends.google.com/)—first launched on May 11, 2006—is run by Google Analytics. At the time of this research, Google has the highest market share of all search engines (about 90%), and Bing is the second most popular with only 2% of the market share (Statcounter, 2020). GT provides a search volume index (SVI) of a keyword, which is the relative popularity of a search term entered in Google's search engine, and measured as a share of a random sample of Google queries[1] in a specific time unit (e.g., day, week, or month) and location. The values of the SVI range from 0 to 100, with a value of 100 indicating a keyword's maximum share of all Google queries during a chosen time and location. For each other time unit, the SVI is calculated as a fraction of the maximum query share time unit. Thus, GT does not provide the absolute number of searches for a term, but rather an estimate of how the popularity of a keyword changes over time. SVI is available at the global and national level as well as the more fine-grained geographic level of a region or city (given that the number of queries for a term are sufficiently high enough to be in accordance with Google's privacy guidelines). It is possible to search for up to five terms simultaneously and compare their popularity within a chosen time and geographic area.

Recently, researchers have demonstrated that search engine queries can be used to study phenomena that are typically measured using surveys. For example, researchers have utilized GT data for studying consumer trends (Vosen & Schmidt, 2011), tracking of disease outbreaks such as influenza (Ginsberg et al., 2009), tracking of economic crises (Jun et al., 2018), and in migration research (e.g., Wladyka, 2013; Vicéns-Feliberty & Ricketts, 2016; Böhme et al., 2020). Chykina and Crabtree (2018) measured concerns about deportation among immigrants in the United States (US) based on the frequency of the search phrase "will I be deported." Stephens-Davidowitz (2014) measured racial animus in the US based on the volume of searches containing a racial epithet directed towards African Americans, and its association with voter preference in presidential elections. However, GT has only been found to be reliable by some of the methodological studies that have evaluated it using criterion measures. For example, some studies found that search queries for political candidates and parties were able to predict poll and election results (Askitas,

[1] See https://newsinitiative.withgoogle.com/training/lessons?tool=Google%20Trends&image=trends

2015; Hauge & Lied, 2017), whereas others found that search queries were not able to predict these outcomes any better than chance (Lui et al., 2011; Harford, 2014).

Given these mixed findings, it is critical to identify the features of Internet search data that distinguish them from survey data and other observational data, which could affect data quality. Search data have at least three potential advantages. First, individuals are presumably less influenced by social desirability pressures when making search queries about sensitive topics (e.g., drug use, racism, sexual practices, income, embarrassing health conditions, etc.) than when answering survey questions about these topics (especially when the surveys are interviewer-administered) due to concerns they may have about protecting their anonymity and privacy (Stephens-Davidowitz, 2014). Second, search queries are recorded at extremely high frequencies rather than at discrete time points, which makes it possible to study events (both expected and unexpected) over time without relying on retrospective survey questions for which forgetting might be a problem. Third, GT data are relatively low cost to obtain and easy to use. Accessing these data does not require advanced levels of programming skills or other data science expertise, and the data are virtually free to everyone with a computer and Internet access. The combination of real-time, low cost data is seen as a solution for the need for timely estimates in many areas, often referred to as *nowcasting* (Zagheni et al., 2017).

However, Internet search data have several potential disadvantages. First, they are anonymized and aggregated by geography, making it impossible to conduct individual-level analyses. Second, a search query must be interpreted by a researcher who makes inferences about the characteristics of a particular user, which makes it difficult to establish construct validity (e.g., searching for a particular political party/candidate is not a clear indication of the intention to vote for that party/candidate). Carneiro and Mylonakis (2009) have found that search queries on the same topic might even be entered differently, depending on a person's background, such as level of education, culture, and language. Third, Internet search data are collected from users of a particular search engine at a particular point in time, not from representative samples of the population. It has been well-documented that Internet users tend to be younger, higher educated, and wealthier than non-users (e.g., Anderson et al., 2019; Porter & Donthu, 2006). In addition, among Internet users, not everyone uses Google as a search engine (Mellon, 2013). For example, users with high privacy concerns may opt for using alternative search engines that put a strong focus on protecting users' privacy (e.g., DuckDuckGo[2]). Finally, a search engine may change—in how it's designed, who uses it, and how they use it—over time in ways that are out of researchers' control, which may confound real change in longitudinal data analysis (e.g., see Lazer et al., 2014).

We further explore these trade-offs in a case study using GT to measure perceived immigration-related threats in Germany, with a focus on the suitability of GT data for this purpose.

[2] However, at the time of the writing of this study, the market share of DuckDuckGo was less than 1% in Germany (Statcounter, 2020).

10.3 Case Study

For this study, we sought to measure the salience of negative opinions towards immigrants in Germany before and after the influx of refugees from the Middle East and North Africa in 2015. GT data were appealing for two reasons. First, our topic was sensitive, and we presumed that GT data would be less susceptible to social desirability bias than a survey-based measure that relied on respondents having to admit their anti-immigrant views. Second, we wanted to study trends, and GT data were available over the time period of interest.

Our main measures of interest pertain to the perceived threats posed by immigrants as an out-group, consistent with Group Threat Theory (Blumer, 1958) in the context of immigration (Zárate et al., 2004; van Klingeren et al., 2015). As shown in Table 10.1, we examined five types of perceived threats: *economic*, *cultural*, *excess*, *security*, and *sexual*. *Economic threat* represents the concerns of natives that immigration will result in a loss of their resources (e.g., lower wages, fewer jobs). *Cultural* concerns focus on whether immigrants will harm society in other ways (e.g., imposing their religious views, needing language accommodations in schools). *Excess threat*, for our purposes, refers to the perception that Germany was unfairly "burdened" by high numbers of migrants in comparison to other European countries. *Security threats* represent concerns about one's physical safety and safety from crime (Larsson, 2017; Fuchs, 2016). Last, *sexual threats* refer to security concerns about sexual violence (Pruitt et al., 2018), which is potentially salient in Germany because of some high-profile cases of sexual violence committed by male migrants (Johnson & Bräuer, 2016; Pruitt et al., 2018).

10.3.1 Keyword Selection

As Table 10.1 shows, we decided on four to five search terms for each threat category. We selected keywords using three steps. First, we acquired a corpus of immigration-related Facebook posts (further described in Lorenz, 2018). Second, we used automated text analysis to determine which terms occurred most often in the posts, which enabled us to discover relevant terms outside our frame of reference. Finally, from this large list of possible search terms, we manually selected terms that we deemed to be conceptually related to the different threat categories. In addition, we used two groups of five search terms each that were not expected to represent perceived threats of immigration, as a sensitivity check for our analysis. One group, labeled as *neutral terms*, contained migration-related terms that we deemed to be about the topic of migration but neutral in tone. The other group, labeled as *randomly selected terms*, contained terms generated by a random word generator.

Table 10.1 Keywords for five perceived threats of immigration, neutral, and random categories

Perceived threat	Google search terms (Original)	Google search terms (Translation)
Economic threat	Flüchtlinge Euro Flüchtlinge Begrüßungsgeld Asylanten Geld Wirtschaftsflüchtlinge Flüchtlinge Kosten	Refugees euro Refugees welcoming money Asylum seekers money Economic refugees Refugees costs
Cultural threat	Islamisierung Scharia Deutschland Islam in Deutschland Salafisten Deutschland Islamisierung Deutschland	Islamization Sharia Germany Islam in Germany Salafists Germany Islamization Germany
Excess threat	Asylflut zu viele Flüchtlinge Flüchtlingswelle in Deutschland Flüchtlingswelle Asylkrise	Asylum flood Too many refugees Refugee wave in Germany Refugee wave Asylum crisis
Security threat	kriminelle Flüchtlinge kriminelle Ausländer Kriminalität Flüchtlinge Deutschland Kriminalität Flüchtlinge Kriminalität Ausländer	Criminal refugees Criminal foreigners Criminality refugees Germany Criminality refugees Criminality foreigners
Sexual threat	Flüchtling vergewaltigt Flüchtlinge Vergewaltigung Vergewaltigung durch Flüchtlinge Flüchtlinge sexuelle Übergriffe	Refugee rapes Refugees rape Rape by refugees Refugee sexual assault
Neutral terms	Asylantrag Immigration Flüchtlinge Migration unbegleitete minderjährige Flüchtlinge	Asylum request Immigration Refugees Migration Unaccompanied minor refugees
Random terms	Alarm Geburtstag gefallen global Wildnis	Alarm Birthday Oblige Global Wilderness

10.3.2 Data and Methods

We extracted the GT data using R version 3.6.0 (R Core Team, 2018) and the R package gtrendsR (Massicotte, 2019).[3] The GT data were collected retrospectively from October 5, 2013, beginning with the first week after the 2013 German Federal

[3] Alternatively, one can also download Google Trends data directly from the GT website by specifying the relevant location and time frame. For further explanation, Google provides a "Trends Help" webpage at https://support.google.com/trends/answer/4365538?hl=en. When comparing multiple search terms over a longer period of time, as we do in our case study, using the R package saves time.

Election, until October 5, 2018 (note that Google Trends data on a weekly basis can only be collected for a maximum time period of 5 years). Geographically, we only included searches that came from German IP addresses. We extracted one dataset for each of the 34 search terms individually, which indicated how popular an individual search term was on a given week during the time period of interest. We calculated the weekly summary SVIs for each of the groups of keywords by averaging the individual SVIs across the individual keywords within a group.

To empirically evaluate our GT-based measures, we compared them to polling data for the Alternative für Deutschland (AfD), a German right-wing party that has run on a largely anti-immigrant platform. The AfD was founded in early 2013 with a Euro-critical orientation (Berning, 2017). In 2015, as many refugees from the Middle East and North Africa were coming to Germany, the AfD established a strong anti-immigrant stance and shifted towards a xenophobic right-wing populist orientation (Schmitt-Beck, 2017). Due to this clear positioning regarding the issue of immigration, and several radical anti-immigrant statements made by some AfD politicians, we assumed that the salience of perceived immigration-related threats in Germany should correlate with the polling outcome for the AfD. We also posited that the anti-immigrant searches may precede (e.g., by several weeks) any changes in the polling results for the AfD, consistent with the notion that individuals may gather information before formulating an opinion (e.g., Druckman et al., 2012; Lux, 2009; Chong & Druckman, 2010).

To measure the popularity of the AfD, we used data from weekly telephone surveys conducted by the Forsa Institute for Scientific Research (accessed on the website www.wahlrecht.de, which provides the results of German polls from different research institutes and also aggregates the poll outcomes). Forsa conducts weekly telephone surveys of the German electorate by asking respondents about their voting intentions in the next federal election (Forsa, 2019).[4] We opted to use data from Forsa because of the weekly frequency of the polls; other institutes publish polling results at most biweekly or less frequently. We chose to use polling data rather than election data because the AfD is a fairly new party without extensive data on its election results.[5] Although the polling data itself might have been subject to social desirability bias and other weaknesses, we assumed that the polls could still reliably measure support for the AfD (which on its face does not seem as sensitive as expressing anti-immigrant views). Indeed, polls predictions tend to provide fairly accurate reflections of election outcomes over time (Norpoth & Gschwend, 2003; Wlezien et al., 2013; Wright et al., 2014).

[4] Sample sizes range between 1001 and 2510. Around 95% of the surveys interviewed more than 2000 participants.

[5] A note on the data: Even if GT data were available for our chosen keywords in all regions, we, unfortunately, lacked the appropriate polling data to run an analysis on a regional basis. While the keyword data were available on a weekly basis for some of the German states (many German states did not surpass the privacy threshold, and therefore GT data were not provided), we did not have enough observations for poll or election data because the poll data were not collected often enough or had fewer than 30 observations. This issue highlights one of the drawbacks of using GT in combination with other data sources with respect to data availability and comparability.

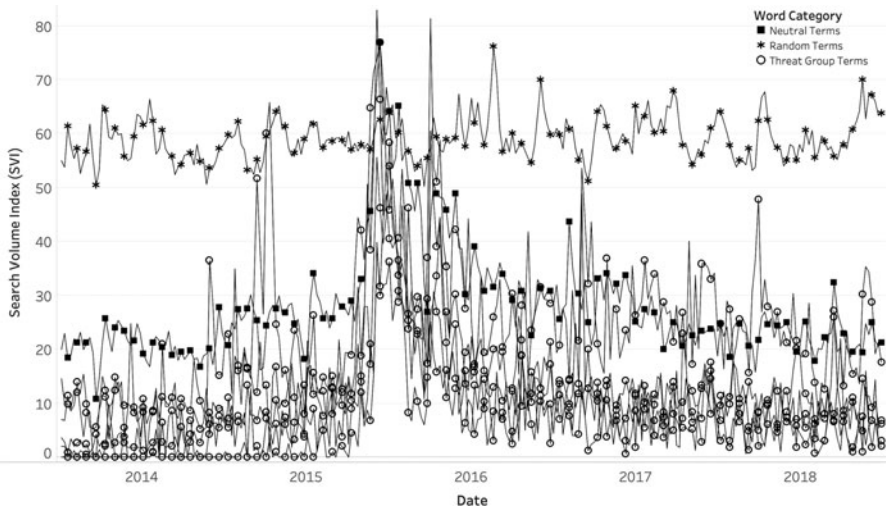


Fig. 10.1 Search Volume Index (SVI) from Google Trends for immigration-related threat group terms, neutral terms, and random terms in Germany from October 2013 to October 2018

We recoded the polling data to the calendar week in which the publication date occurred to be on the same timescale as the GT data. In a case where two polling estimates were published in the same week, we used the arithmetic mean. If no data were available for the week, the mean of the previous week and the following week was imputed. Thus, our analytical data set has no missing values.

We investigated the relationship between the GT data and polling results by using simple bivariate correlation coefficients. We also computed *real-time* correlations between the searches and poll results from the same period. In addition, we computed *temporally-lagged* correlations (Keane & Adrian, 1992; Podobnik & Stanley, 2008) between the keyword searches and poll results that occurred at a later time period than the searches.

10.3.3 Results

Figure 10.1 shows the SVIs for the five *threat* groups, as well as the *neutral* terms and *randomly generated* terms. While variation occurs across the *threat* groups with respect to search volume, the trends across the different *threat* groups is largely consistent: search volume increases in early 2015 and then again later that year, and then decreases to about the original volume in 2016 and later. The increases correspond to when the peak of the influx of refugees from the Middle East and North Africa to Germany occurred. While the SVI of *neutral* terms related to immigration seems to move in a similar pattern, it begins at a higher level before peaking in 2015, and also returns to this higher level compared to the search volume

Table 10.2 Pearson's r for groups of search terms and AfD polling results

Keyword Group	Correlation coefficient with AfD polling results
Economic threat	−.189**
Cultural threat	−.217***
Excess threat	−.129*
Security threat	.286***
Sexual threat	.057
Neutral terms	−.117
Random terms	.162**

*p < .05; **p < .01; ***p < .001

for the thread terms. This indicates that the volume of searches for the neutral terms was affected by the events in Germany in 2015, but not as much as the search volume for the threat terms. As expected, the SVI for the *randomly selected* terms stayed more or less consistent across our 5 year reference period. If the SVIs of the threat groups reveal actual changes in anti-immigrant opinions over time, we would expect them to correlate with the AfD polling results either in real-time or after a lag period.

As shown in Table 10.2, the real-time correlations between the aggregated Google searches and the AfD polling results varied substantially across the different threat groups. The search terms related to *security threat* were significantly and positively correlated with AfD polling results ($r = .29$; $p < .001$). The search terms related to *sexual threat* were also positively correlated with AfD polling results, although the effect did not reach statistical significance ($r = .06$; $p = .361$). Surprisingly, each of the other three *threat* groups were significantly and negatively correlated with AfD polling results. As expected, the *neutral* terms showed no significant correlation with AfD polling results ($r = -.12$; $p = .059$). However, the SVI of the group of *randomly selected* terms showed a significant positive correlation with AfD polling results ($r = .162$; $p = .009$).

The correlation results for individual search terms are provided in the Appendix (Table 10.3). Four individual search terms had a small ($r = .2$) to medium ($r = .5$) correlation with the criterion measure in the expected direction, including three terms in the security threat group—*kriminelle Flüchtlinge* (*criminal refugees*; $r = .31$; $p < .001$), *Kriminalität Flüchtlinge* (*criminality refugees*; $r = .32$; $p < .001$), and *Kriminalität Flüchtlinge Deutschland* (*criminality refugees Germany*; $r = .27$; $p < .001$)—as well as one term in the sexual threat group—*Flüchtlinge vergewaltigt* (*refugees rapes*; $r = .26$; $p < .001$). At the same time, four individual search terms had a small to medium negative correlation with the AfD polling results: *Flüchtlinge Begrüßungsgeld* (*refugees welcoming money*; $r = -.27$; $p < .001$), *Asylanten Geld* (*asylum seekers money*; $r = -.34$; $p < .001$), *Wirtschaftsflüchtlinge* (*economic refugees*; $r = -.28$; $p < .001$), and *Asylflut* (*asylum flood*; $r = -.25$; $p < .001$). Similarly, we found that two of the randomly selected terms had a medium correlation with the criterion—*Geburtstag* (*birthday*; $r = .51$; $p < .001$) and *gefallen* (*oblige*; $r = .36$; $p < .001$)—and one had a small negative correlation: *Wildnis* (*wilderness*; $r = -.22$; $p < .001$).

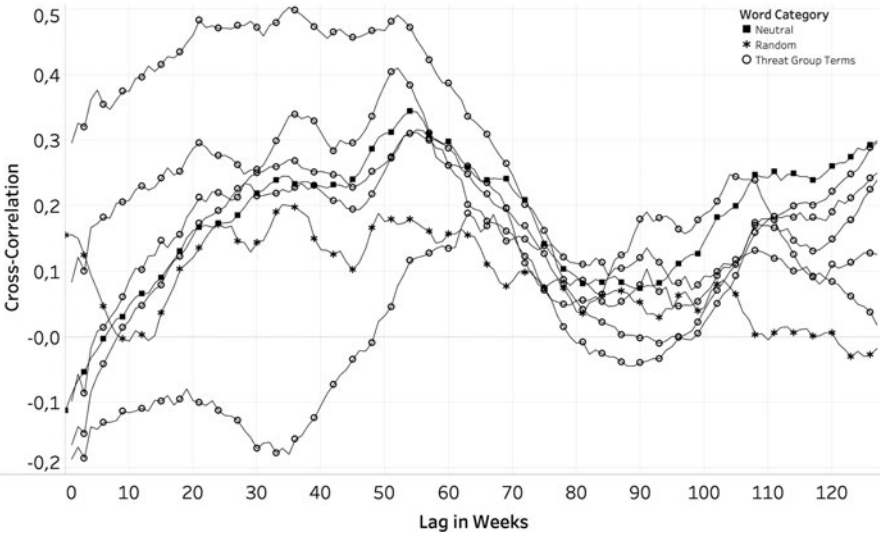


Fig. 10.2 Cross-correlations of Search Volume Index (SVI) from Google Trends for immigration-related threat group terms, neutral terms, and random terms with AfD polling results in Germany from October 2013 to October 2018

Next, we examined the temporally-lagged correlations with a time lag ranging from 1 to 127 weeks. Figure 10.2 shows these correlations for each of the search term groups. This figure reveals that all of the threat categories, even those that showed negative correlation values without a lag, reached medium positive correlation values once a lag was implemented. The peak correlation coefficients for the five threats ranged between .25 (*cultural threat* after 104 weeks) and .50 (*security threat* after 35 weeks). With the exception of the cultural threat group, the curves for the other threat groups showed similar patterns: the correlations increased with a lag up until 35–55 weeks of lag, and decreased rapidly until approximately 85 weeks of lag, and then stayed relatively stable.

The curve for the migration-related neutral terms showed a similar lag pattern and also peaked at a lag of 54 weeks. As expected, the group of randomly selected terms yielded the lowest maximum correlation with the AfD polls and stayed within a range of −.05 and .2, depending on the lag. This finding suggests that the connection between migration-related searches (regardless of their connotation) and AfD poll results was more than just random noise.

10.4 Discussion

Social science researchers are increasingly using alternative non-survey data sources to answer substantive research questions. The aim of this chapter was to explore the advantages and disadvantages of using one of these new data sources, Google

Trends (GT), based on a case study on the perceived threats of immigration. We found that GT data did not consistently correlate with the polling data for the right-wing German AfD in the expected direction in real-time, but rather was consistently predictive of future polling trends (35–104 weeks later) at a moderate level ($r = .25$ to .50), although the size of the correlations varied across time periods and groups of keywords. By contrast, although a group of randomly selected search terms had a small but significant positive correlation with current AfD polling results, it had the lowest correlations with future polling results. We take this finding as an indication that the correlation between the salience of specific threat-perceptions and AfD polling results is more than just random noise that could be expected from such an amount of data, and, moreover, it seems plausible that a sizable share of these searches was associated with virulent anti-immigrant attitudes.

Our case study highlighted several of the advantages of using GT data. First, we were able to gather the data quickly and at no cost. The data, covering a time period from 2013 to 2018, provided information on how the immigration debate changed in Germany over a long period. Conducting a longitudinal survey on this topic would have been expensive and not possible in retrospect. Second, we measured a topic that triggers social desirability concerns without relying on survey self-reports for which bias is a concern. We assumed that Google users had less concern when typing certain search terms as compared to openly admitting to anti-immigration sentiments in a survey. Third, we conducted a longitudinal analysis with measures recorded at discrete time points (weekly) over a relatively long period of time (5 years). As the AfD began to own the issue of immigration, Internet searches on immigration also increased. We were able to follow the timeline of when the public began searching for crime statistics with a connection to immigration. For example, a large spike in searches about immigration and crime and sexual threat perceptions occurred directly after the New Year's Eve events in Cologne at the start of 2016.[6]

Our case study also highlighted several of the disadvantages of using GT. First, with access to aggregate-level data only, we were unable to explore individual-level correlates (e.g., gender, education) of the perceived threats of immigration. If information on individual voters, such as partisanship, was available, we could have considered whether certain threat cues actually affected voters differently (Lahav & Courtemanche, 2012). Second, the search data were collected from Google users, who are most likely not a representative sample of the population of Germany, and probably not even representative of Internet users in Germany. Without access to sociodemographic and other auxiliary information about the searchers, we could not adjust for any potential bias due to the selectivity of the users who produced these data. Third, the search terms we selected were not equally valid measures of our construct of interest: some keywords were positively correlated with a variable that they ought to be related to (AfD support), whereas others

[6]The growing opposition to immigration in Germany often is attributed to the 2015/2016 New Year's Eve events in Cologne where a large number of sexual assaults were attributed to male immigrants (e.g., Ingulfsen, 2016).

were negatively correlated with the same variable. The choice of keywords cannot be validated directly, and so this is a significant issue for any research using GT data. Even though we found that the security threats correlated with the AfD polling data, another explanation for our results may be that the topic of crime and immigration is easier to operationalize than other topics using GT. For example, the idea that the Islamic culture will take over Germany was rather difficult to define using only keywords for Internet searches, and GT may not be able to capture these complicated nuances, especially since *culture* is a highly debated concept subject to personal opinion. Fourth, Google's algorithm automatically suggests search terms once a user begins typing. We checked whether the first words of our search terms returned suggestions that were particularly negative towards immigrants, and we found that, at the time of our data collection, this was not the case for our keywords. However, the algorithm may have returned different suggestions over time when users typed immigration-related words, and Google may change its algorithm in the future, which could potentially jeopardize the measurement of long-term trends.

Despite these issues, our case study demonstrates that GT data can be predictive of public opinion, which supports the notion that GT has value for social science researchers as a real-time monitoring tool or leading indicator of public opinion, and it may be especially well suited for measuring socially undesirable views. Future research should investigate which events or phenomena can be reliably measured using GT. Our methodology provides an approach for doing this through the validation of GT-based measures with benchmark survey data. Future research must also address the important questions regarding keyword selection. For example, in the absence of a validation measure, how should keywords be selected and how should they be aggregated into summary measures? These efforts will expand the ways in which social science researchers can leverage Internet search data to produce population-level estimates of public opinion and behavior.

Appendix

Table 10.3 Correlation overview of search terms with AfD poll outcomes

Keyword group	Google search terms (Translation)	Individual correlation coefficients	Combined correlation for perceived threat
Economic threat	Refugees euro	.003	−.189**
	Refugees welcoming money	−.272***	
	Asylum seekers money	−.344***	
	Economic refugees	−.281***	
	Refugees costs	.077	

(continued)

Table 10.3 (continued)

Keyword group	Google search terms (Translation)	Individual correlation coefficients	Combined correlation for perceived threat
Cultural threat	Islamization	−.195**	−.217***
	Sharia Germany	−.060	
	Islam in Germany	−.199**	
	Salafists Germany	−.143*	
	Islamization Germany	−.160**	
Excess threat	Asylum flood	−.247***	−.129
	Too many refugees	−.178**	
	Refugee wave in Germany	−.115	
	Refugee wave	−.086	
	Asylum crisis	.091	
Security threat	Criminal refugees	.305***	.286***
	Criminal foreigners	−.001	
	Criminality refugees Germany	.269***	
	Criminality refugees	.321***	
	Criminality foreigners	.043	
Sexual threat	Refugee rapes	.255***	.057
	Refugees rape	−.021	
	Rape by refugees	−.072	
	Refugee sexual assault	.061	
Neutral terms	Asylum request	−.004	-.117
	Immigration	−.072	
	Migration	−.151*	
	Refugees	−.162**	
	Unaccompanied minor refugees	−.078	
Random terms	Alarm	−.119	.162**
	Birthday	.511***	
	Oblige	.362***	
	Global	−.002	
	Wilderness	−.222***	

*p < .05; **p < .01; ***p < .001

References

Anderson, M., Perrin, A., Jiang, J., & Kumar, M. (2019). *10% of Americans don't use the internet. Who are they?* Pew Research Center. Retrieved from https://www.pewresearch.org/fact-tank/2019/04/22/some-americans-dont-use-the-internet-who-are-they/

Askitas, N. (2015). *Calling the Greek Referendum on the Nose with Google Trends* (IZA discussion paper). Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2708382

Berning, C. C. (2017). Alternative für Deutschland (AfD) – Germany's New Radical Right-wing Populist Party. *ifo DICE Report*, ifo Institut – Leibniz Institut für Wirtschaftsforschung an der Universität München, München, *15*(4), 16–19.

Blumer, H. (1958). Race prejudice as a sense of group position. *Pacific Sociological Review, 1*(1), 3–7.

Böhme, M. H., Gröger, A., & Stöhr, T. (2020). Searching for a better life: Predicting international migration with online search keywords. *Journal of Development Economics, 142*, 1–14.

Carneiro, H. A., & Mylonakis, E. (2009). Google Trends: A web-based tool for real-time surveillance of disease outbreaks. *Clinical Infectious Diseases, 49*, 1557–1564.

Chong, D., & Druckman, J. N. (2010). Dynamic public opinion. *American Political Science Review, 104*(4), 663–680.

Chykina, V., & Crabtree, C. (2018). Using Google Trends to measure issue salience for hard-to-survey populations. *Socius*. https://doi.org/10.1177/2378023118760414

Druckman, J. N., Fein, J., & Leeper, T. J. (2012). A source of bias in public opinion stability. *American Political Science Review, 106*(2), 430–454. https://doi.org/10.1017/S0003055412000123

Forsa. (2019). Methoden. (Version: May 22 2019) [Website Article]. Retrieved from https://www.forsa.de/methoden/

Fuchs, C. (2016). Racism, nationalism and right-wing extremism online: The Austrian Presidential Election 2016 on Facebook. *Momentum Quarterly, 5*(3), 172–196.

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature, 457*(7232), 1012–1014.

Harford, T. (2014). Big data: Are we making a big mistake? *Financial Times*.

Hauge, H. S., & Lied, T. B. (2017). *Explaining Election Outcomes Using Web Search Data: Evidence from the U.S. Presidential Elections 2008–2016*. Unpublished master's thesis). Norwegian School of Economics.

Ingulfsen, I. (2016, February 18). Why aren't European feminists arguing against the anti-immigrant right? *Open Democracy*. Retrieved from https://www.opendemocracy.net/en/5050/why-are-european-feminists-failing-to-strike-back-against-anti-immigrant-right/

Johnson, H., & Bräuer, T. (2016). Migrant crisis: Changing attitudes of a German city. *BBC News*. Retrieved from https://www.bbc.com/news/world-europe-36148418

Jun, S.-P., Hyoung, S. Y., & Choi, S. (2018). Ten years of research change using Google Trends: From the perspective of big data utilizations and applications. *Technological Forecasting & Social Change, 130*, 69–87.

Keane, R. D., & Adrian, R. J. (1992). Theory of cross-correlation analysis of PIV images. *Applied Scientific Research, 49*, 191–215.

Lahav, G., & Courtemanche, M. (2012). The ideological effects of framing threat on immigration and civil liberties. *Political Behavior, 34*(3), 477–505.

Larsson, A. O. (2017). Going viral? Comparing parties on social media during the 2014 Swedish election. *Convergence: The International Journal of Research into New Media Technologies, 23*(2), 117–131.

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The parable of Google flu: Traps in Big Data analysis. *Science, 343*(6176), 1203–1205.

Lorenz, R. E. (2018). *Right-wing extremism online: How the AfD frames immigration on Facebook*. Unpublished master's thesis. Universität Mannheim.

Lui, C., Metaxas, P. T., & Mustafaraj, E. (2011). *On the predictability of the U.S. elections through search volume activity*. Proceedings of the IADIS International Conference on e-Society, Avila, Spain.

Lux, T. (2009). Rational forecasts or social opinion dynamics? Identification of interaction effects in a business climate survey. *Journal of Economic Behavior & Organization, 72*(2), 638–655. https://doi.org/10.1016/j.jebo.2009.07.003

Massicotte, P. (2019). *gtrendsR: Perform and display Google Trends queries*. R package version 1.4.3.

Mellon, J. (2013). Internet search data and issue salience: The properties of Google Trends as a measure of issue salience. *Journal of Elections, Public Opinion and Parties, 24*(1), 45–72.

Norpoth, H., & Gschwend, T. (2003). Politbarometer und Wahlprognosen: Die Kanzlerfrage. In A. M. Wüst (Ed.), *Politbarometer*. VS Verlag für Sozialwissenschaften.

Podobnik, B., & Stanley, H. (2008). Detrended cross-correlation analysis: A new method for analyzing two nonstationary time series. *Physical Review Letters, 100*(8).

Porter, C. E., & Donthu, N. (2006). Using the technology acceptance model to explain how attitudes determine Internet usage: The role of perceived access barriers and demographics. *Journal of Business Research, 59*(9), 999–1007.

Pruitt, L., Berents, H., & Munro, G. (2018). Gender and the age in the construction of male youth in the European migration "crisis". *Journal of Women in Culture and Society, 43*(3), 687–709.

R Core Team. (2018). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Salganik, M. (2019). *Bit by bit: Social research in the digital age*. Princeton University Press.

Schmitt-Beck, R. (2017). The 'Alternative für Deutschland in the Electorate': Between single-issue and right-wing populist party. *German Politics, 26*(1), 124–148. https://doi.org/10.1080/09644008.2016.1184650

Statcounter. (2020). *Search engine market share Germany*. Retrieved from https://gs.statcounter.com/search-engine-market-share/all/germany

Stephens-Davidowitz, S. (2014). The cost of racial animus on a black candidate: Evidence using Google search data. *Journal of Public Economics, 118*, 26–40.

van Klingeren, M., Boomgaarden, H. G., Vliegenthart, R., & de Vreese, C. H. (2015). Real world is not enough: The media as an additional source of negative attitudes toward immigration, comparing Denmark and the Netherlands. *European Sociological Review, 31*(3), 268–283.

Vicéns-Feliberty, M. A., & Ricketts, C. F. (2016). An analysis of Puerto Rican interest to migrate to the United States using Google trends. *The Journal of Developing Areas, 50*(2), 411–430. https://doi.org/10.1353/jda.2016.0090

Vosen, S., & Schmidt, T. (2011). *Forecasting private consumption: Survey-based indicators vs. Google trends* (Ruhr Economic Papers 155). RWI – Leibniz-Institut für Wirtschaftsforschung, Ruhr-University Bochum, TU Dortmund University, University of Duisburg-Essen.

Wladyka, D. (2013). The queries to Google Search as predictors of migration flows from Latin America to Spain. *UTB/UTPA electronic theses and dissertations*, 10.

Wlezien, C., Jennings, W., Fisher, S., Ford, R., & Pickup, M. (2013). Polls and the vote in Britain. *Political Studies, 61*(1), 66–91. https://doi.org/10.1111/1467-9248.12008

Wright, M. J., Farrar, D. P., & Russell, D. F. (2014). Polling accuracy in a multiparty election. *International Journal of Public Opinion Research, 26*(1), 113–124. https://doi.org/10.1093/ijpor/edt009

Zagheni, E., Weber, I., & Gummadi, K. (2017). Leveraging Facebook's advertising platform to monitor stocks of migrants. *Population and Development Review, 43*, 721–734.

Zárate, M. A., Garcia, B., Garza, A. A., & Hitlan, R. T. (2004). Cultural threat and perceived realistic group conflict as dual predictors of prejudice. *Journal of Experimental Social Psychology, 40*, 99–105.

Chapter 11
Conclusion: Migration Research in Times of Ubiquitous Digitization

Sebastian Rinken and Steffen Pötzschke

Microprocessors and the Internet are outstanding cases in point for the increasingly frenetic pace of technological innovation. Starting in the 1950s and 1980s, respectively, computing speeds and interconnected data volumes have grown exponentially to become two related processes that have triggered profound changes in all kinds of productive, commercial, administrative, cultural, and social activities. Yet, when surveying the history of technology from a bird's-eye perspective, it is striking how many inventions—especially since the advent of industrialization about two centuries ago—have made a decisive mark. Mankind's collective path from subsistence communities to ubiquitous digitization is littered with milestones, and the number of important innovations is so large that it is difficult to arrive at a convincing shortlist. Was the lightbulb a more disruptive novelty than the automobile, or vice versa? How does the smartphone compare to the steam engine?

A focus on information and communication technologies (ICTs) helps us appreciate the truly epochal status of the digital revolution. Throughout the entire history of mankind, only one similarly momentous achievement stands out in this realm—the emergence of written language in ancient Mesopotamia (oral language is pre-technological, since it lacks non-biological hardware) (Majó, 2012, pp. 67–69). About 4000 years ago, the combination of novel coding (signs, alphabets) and storage technology (papyrus) started to smash the barriers of time and space associated with the physical range of humans' hearing and sight. Nowadays, the combination of novel coding (bits) and processing technology (microchips, Internet) is smashing the barriers of time and space erected by a range of mutually

S. Rinken (✉)
Spanish Research Council (CSIC), Institute for Advanced Social Studies (IESA), Córdoba, Spain
e-mail: srinken@iesa.csic.es

S. Pötzschke
GESIS – Leibniz Institute for the Social Sciences, Mannheim, Germany
e-mail: steffen.poetzschke@gesis.org

S. Pötzschke, S. Rinken (eds.), *Migration Research in a Digitized World*, IMISCOE Research Series, https://doi.org/10.1007/978-3-031-01319-5_11

incompatible (physical, electric, chemical, and electronic) means of storing and transmitting text, sound, and images. This unprecedented conversion to one universal format makes an ever-increasing volume of information available, at least potentially, to everybody, anywhere, anytime. By comparison, Gutenberg's celebrated invention was relatively minor, since while facilitating economies of scale—a feat that changed the world, to be sure—it left the extant information coding and transmission system largely unaltered.

By definition, since the business of scientists is distilling information into knowledge, the digital revolution profoundly affects the scientific endeavor. This insight is especially true for empirically-minded scientific disciplines that devote considerable effort to obtaining data in the first place. In survey research, as in medical trials, the process of information-gathering must follow strict rules for results to be valid, the most basic of which concern the selection and handling of study participants. However, in recent years, survey quality standards have become increasingly difficult to achieve due to growing coverage and non-response biases (Groves, 2011) and—especially when tackling sensitive items—persistent (if hard-to-measure) response bias (Krumpal, 2013). When surveying migrant populations, these challenges are exacerbated by added difficulties: international migrants, including refugees, tend to elude established sampling procedures, are often difficult to locate, may resist or resent the interviewee role, and require multi-lingual questionnaires and/or linguistic assistance—in short, they are notoriously hard-to-survey (Tourangeau et al., 2014).

It seems obvious that innovative ICTs are a game-changer in this context, since traditional time-space restrictions are particularly bothersome when targeting highly diverse and mobile populations. Yet, are migration scholars seizing the opportunities afforded them by new technologies?

The rationale of this book is based on the hypothesis that migration studies have even more to gain from the digital revolution than most other fields of social research. Each contribution in its own way encourages migration scholars to explore the added advantages granted them by innovative technologies and approaches. None of the authors, much less the editors, advocate an uncritical adoption of new technology: we all agree that its benefits have to be weighed carefully against its limitations, advantages put into perspective, and risks adequately managed. Yet, all the contributors agree that inertia is not an acceptable option, and so all the book chapters prod migration researchers to explore new data types and technological tools actively, rather than continuing to depend exclusively on accustomed data collection procedures. Because such experimentation inevitably entails a learning curve, we believe that migration scholars stand to benefit, both individually and collectively, even from mixed experiences. Before resuming a general discussion on how migration research is affected by the relentless process of digitization, we summarize the objectives, procedures, and results of each chapter.

11.1 Added Traction: New Tools for Sampling and Data Collection

The five contributions to the first part of the book address a variety of ways that new technology can improve the collection of "designed" data from purposefully sampled respondents. Three chapters address the twin problems of sampling and locating highly mobile, as well as oftentimes dispersed, populations, and two chapters focus on the challenges of target populations' markedly varied sets of linguistic competences.

In *Innovative Sample Designs for Studies of Refugees and Internally Displaced Persons*, Stephanie Eckman and Kristen Himelein focus on the procedures for drawing probability samples of forced migrants. More specifically, they discuss approaches to sampling that are apt for implementation in face-to-face surveys conducted in developing countries. They distinguish between three contexts of sample recruitment that depend on the characteristics of the study population: forced migrants living in camps, urban settings, and those who lack even a moderately stable place of residence ("on the move"). Regarding each of these scenarios, the authors explore a range of sampling options and highlight specific challenges and avenues to address them. The implementation of these options relies on new technologies to varying degrees and in various ways. One fascinating example is the use of satellite pictures or images collected by aerial drones (which can be deployed on-site by research teams) for generating real-time maps of migrant dwellings, which enable interviewers on the ground to employ aleatory (route-based) sampling plans. Other inspiring cases of new technological options include the use of geographic information system software and GPS-equipped interviewers in the recruitment of highly mobile migrants into a sample, and the incorporation of digital trace data in sampling strategies. While the tools of choice depend on each particular study's objectives, budget, and time-frame, the general take-home-message of this chapter is that regardless of the data collection mode, it is worthwhile for researchers to think creatively about how new technologies can improve the research process. Also, as this chapter shows, the technological enhancement of rather traditional data collection modes can be just as advantageous as the incorporation of new data types.

In *Targeting on Social Networking Sites as Sampling Strategy for Online Migrant Surveys: The Challenge of Biases and Search for Possible Solutions*, Anna Rocheva, Evgeni Varshaver, and Nataliya Ivanova shift the focus from sampling for face-to-face interviews to sampling for online surveys. Specifically, Rocheva and her colleagues analyze the use of advertisements in two of Russia's leading social networking sites (SNS)—Vkontakte and Odnoklassniki—to capture participants for various Internet surveys of migrant populations in Russia. By doing so, the authors contribute to the growing body of literature that is investigating the use of alternative recruitment procedures when reliable sampling frames are unavailable or not feasible. Previous work on the SNS-based sampling of hard-to-reach populations has dealt mostly with Western European countries, the USA, and Australia, thus the

extant literature refers mostly to Facebook, given its predominance in these markets. However, as the authors stress, a disproportionate focus on one particular SNS has serious limitations, since Facebook is not available in all countries, and is not necessarily the most-used social network in the countries in which it can be accessed. Furthermore, since each SNS employs its own algorithms, targeting options, and general procedures, the knowledge obtained about one such platform cannot be simply extrapolated to others. Therefore, although these authors' conclusions largely confirm those of previous research, their detailed examination of Vkontakte and Odnoklassniki constitutes a significant addition to the extant literature. On the upside, SNS-based sampling has been found to enable researchers to investigate highly dispersed populations within a short time frame. On the downside, however, such procedures generate non-probability samples, with the added limitation of uncertainty about the algorithm parameters that underpin the target selection of SNS-based advertisements. In Russia as elsewhere, these algorithm parameters are anxiously guarded as proprietary information by platform owners, and may change without researchers' knowledge. The ensuing combination of selection and self-selection biases of unknown proportions suggests that, as long as those conditions persist, the results of SNS-based sampling have to be considered with caution.

A similar note of caution is raised in the chapter *Web-Based Respondent-Driven Sampling in Research on Multiple Migrants: Challenges and Opportunities* by Ágata Górny and Justyna Salamońska, which explores how web-based respondent-driven sampling (web-based RDS) could be used to recruit multiple migrants into a web survey. Just like the authors of the previous chapter, they address non-probability sampling for online surveys, yet their context shifts from a heterogeneous and broadly defined target population residing in one specific country (as explored in Chap. 3) to a narrowly defined target group—Polish migrants who have resided in several foreign countries—scattered across a potentially large number of places, a situation that accentuates the "hard-to-identify" component of the manifold complications that typically make migrants a hard-to-survey population. Conceptually, the RDS approach resolves this difficulty by asking respondents to recruit as additional study participants all those among their friends and acquaintances who meet the target group definition. In combination with a diversified pool of first-round interviewees ("seeds"), this rule of recruiting *all* eligible contacts is expected to give RDS an edge, in terms of representativeness, over other non-probability sampling strategies. However, in this particular case, it turned out that the target definition (multiple migration experience) was not salient enough to generate extensive recruitment chains; therefore, first-round seeds accounted for the majority of participants. Accordingly, one of the conclusions of the chapter is the need for researchers to verify, prior to a study's launch, that the target population is defined in terms of a salient self-definition as a social group. Second, whereas traditional face-to-face RDS interviewers can explain to their respondents the importance of recruiting additional participants, this crucial step can be exceedingly challenging regarding web-based RDS. Third, again with a view to increasing the likelihood of referral to additional interviewees, the authors highlight the importance of keeping the questionnaire short and engaging. Fourth, the authors point out that

the management of incentives, a vital ingredient of the RDS methodology, also needs especially careful thought and preparation with respect to an online-only research setting in which staff and participants are literally scattered across the world. To recapitulate, this study will be extremely helpful to migration scholars interested in web-based RDS.

In *Computer-Assisted Migration Research: What We Can Learn about Source Questionnaire Design and Translation from the Software Localization Field*, Dorothée Behr provides a fascinating example of knowledge transfer. Diverging from the well-trodden path of scientists' insights being applied to other (political, economic, commercial, etc.) realms, on this occasion academics are at that transfer's receiving end. Behr draws on the know-how of multinational technology companies to outline the manifold steps, sophisticated workflows, and multi-professional expertise required for the seamless implementation of pluri-linguistic questionnaires in computerized surveys. Specifically, she details the procedures used by the software localization industry to ensure that the vast and continuously changing range of technology products are equipped with customized versions of instructions and user interfaces that consumers anywhere on the globe may require. Behr's examination of the complexity and resource requirements of this blueprint sends a sobering message to the oftentimes atomized and underfunded community of migration scholars: state-of-the-art multilingual questionnaires for digital surveys require painstaking forward planning and a seamless cooperation of many distinct professionals, and thus, they require top-drawer organizational capabilities and a substantial budget. Small-scale surveys entailing only two or three languages may still be manageable with more artisanal means, but Behr's chapter illustrates how, in migration studies and other cross-national survey operations, the research landscape is evolving towards increasingly large and complex management structures. Although only unusually well-resourced projects can hope to emulate the procedures that Behr outlines, her study should appeal to a much broader audience, since it exemplifies how cross-cultural adaptation, rather than the niche concern of specialized scholars, is quite literally mainstream business.

The second of our chapters on linguistic matters—*Surveying Illiterate Individuals: Are Audio Files in Computer-Assisted Self-Interviews a Useful Supportive Tool?* by Florian Heinritz, Gisela Will, and Raffaela Gentile—provides methodological reflections on a research design that had been optimized on substantive grounds. Thus, rather than employing distinct methodological options alternatively in an experimental setting, several such options were combined in the fieldwork. Despite the ensuing limitations of this approach, the chapter contains interesting observations on the tools that can be used with an especially hard-to-survey population—international migrants who lack reading skills in their native tongue. Heinritz and colleagues addressed this challenge by preparing audio recordings for all the questionnaire items in a range of languages; in addition, they translated the questionnaire into these languages and deployed fieldwork staff competent in these languages. Initially, the audio recordings were meant to be used by respondents who preferred to administer part of the (computer-assisted) questionnaire themselves, with a view to preserving the anonymity of chosen

responses and thus preventing response bias due to social desirability concerns. However, since the interviewers were present at all times throughout all the interviews, the distinction between computer-assisted self-interview (CASI) and computer-assisted personal interview (CAPI) was blurred in practice, and the intended safeguard against social desirability bias became largely elusive. In addition to illustrating the need for researchers to carefully envision and pretest the whole fieldwork process to detect unanticipated glitches, this study cautions that the incorporation of technologically advanced features does not automatically guarantee enhanced data quality.

11.2 A New Dimension: Leveraging "Found" Data for Migration Research

The four contributions to the second part of the collection address options and tools for accessing and using *found data*, i.e., data that were either collected actively by third parties or generated passively—without a prior research design or specified scientific purpose—by digital sensors or devices.

Sebastian Rinken's and José Luis Ortega's *Leveraging the Web for Migration Studies: Data Sources and Data Extraction* provides an introduction to the second part of the collection. They explore the implications of the "data revolution" for migration research, i.e., the availability of ever-increasing amounts of mostly unstructured data through the Internet. Rinken and Ortega argue that such new data sources are particularly useful for migration studies, given the limitations of traditional research techniques and data sources. In addition to highlighting the wealth of third-party surveys and administrative datasets accessible on the Internet via a range of generalist data portals, specialized sites, data repositories, and search engines, the main contribution of this chapter is its discussion of some of the techniques that enable researchers to extract non-structured data from the Internet. Rather than a hands-on crash course, this introduction to *web-scraping* as a data collection method aims to alert migration researchers of the need to broaden their skill set, both as individuals and as cross-disciplinary teams. In Rinken's and Ortega's view, unstructured data could offer extraordinary opportunities for gaining insights into migration flows, integration patterns, and migration-related attitudes, to name three areas of outstanding relevance. As they strive to advertise this potential, the authors may at times strike an overenthusiastic note in that important issues, such as data protection, privacy considerations, and data quality, are hinted at rather passingly and certainly warrant more sustained consideration. Thus, in the context of the overall structure of the collection, this chapter serves as an appetizer inviting migration scholars to actively explore new data types and their ensuing research options. Although Rinken and Ortega do not suggest that traditional types of data will disappear, they anticipate that their added value, when compared to "found" data, will become less and less obvious as digitization affects an ever-growing share

of more and more people's daily lives. Thus, their advice to migration scholars is not to sit on the fence, but rather enter the fray!

The remaining three chapters in the book's second part exemplify, each in its own way, how migration scholars can leverage new data sources for their research objectives. In *How Canada's Data Ecosystem Offers Insights on the Options for Studying Migration in an Unprecedented Era of Information*, Howard Ramos and Michael Haan focus on the use of administrative records as resources for scientific inquiry, paying special attention to the interconnections between different datasets. Drawing mainly on their intimate knowledge of the Canadian data environment, Ramos and Haan highlight innovative approaches that facilitate the utilization of different administrative data sources and their linkage for research purposes. While attesting to the value that such data hold for migration studies, the authors identify a number of challenges that need to be addressed to fully harness their potential. A first issue concerns the diversity of definitions and measurement options employed by distinct data providers. With respect to this concern, the authors insist on the need for scientists and practitioners to develop standardized definitions and instruments. This solution seems quite ambitious even at a national level, not to mention a cross-national perspective, considering the breadth of statistical operations involved and the diversity of the specific goals pursued by distinct data providers and data collection operations. A second major challenge concerns data access and curation. Even in Canada, where administrative records are increasingly available to researchers, access usually requires physical presence at a specific institution. This rather anachronistic prerequisite poses added difficulties especially for scholars based in other countries. As for curation, making administrative records available to the scientific community requires a considerable additional workload, and thus investment in qualified personnel by the data providers. Finally, Ramos and Haan also draw attention to the fact that the secondary use of administrative sources raises important data security and data protection issues. Notwithstanding these challenges, they urge researchers and data producers to cooperate nationally and internationally with a view to creating the data infrastructures and data handling protocols necessary for making administrative data more readily available for scientific analysis.

In *Assessing Transnational Human Mobility on a Global Scale*, Emanuel Deutschmann, Ettore Recchi, and Michele Vespe switch the focus to an even more ample notion of *found data*, the truly huge volumes of data that are gathered for purely operational reasons, without any inherent relation to researchers' conceptual definitions and needs. Deutschmann and his colleagues combine datasets on world-wide tourism and air passenger traffic to generate a plausible estimate of a target variable that is not provided by any of those two sources, namely, *cross-border mobility*. Since they manage to uncover something that was "visibly absent" in those sources, their dataset merging procedures seem to be touched by magic. Put more prosaically, the authors depart from a painstakingly crisp description of the content covered by their two baseline sources. Next, they analyze the relation of these found data with their information needs. By detecting the overlapping kinds of information and defining rules of transformation for specific data categories, they are able to outline a pathway toward a merged dataset. While the exact procedures and steps are

obviously specific to each study, such a formalized approach (that translates neatly to mathematical formulae) is likely to be appropriate whenever scholars face analogous challenges. Thus, the contribution by Deutschmann and his colleagues is an excellent example of how extant data collections can be repurposed for research needs. By the same token, this chapter showcases the virtues of datasets that cover entire populations—in this case, all of the world's cross-border overnight stays and air travel.

Finally, in *Google Trends as a Tool for Public Opinion Research: An Illustration of the Perceived Threats of Immigration*, Reilly Lorenz, Jacob Beck, Sophie Horneber, Florian Keusch, and Christopher Antoun provide a stimulating example of how scientific inquiry can benefit from data that are readily available on the Internet. Lorenz and her colleagues demonstrate that the usefulness of search engine data (specifically, Google's Trends feature) depends not only on the careful selection of search terms, time periods, and territorial references, but also on recognizing inherent limitations. Google Trends provides information on relative frequencies regarding chosen search terms and reference periods, but not on user profiles or absolute frequencies, although estimates of these absolute frequencies are available to the paying customers of Google Ads. Rather than directly revealing specific behavioral or attitude patterns, such data speak to the relative salience of search terms. Side-stepping these limitations, the authors' strategy of external validation detects a lagged aggregate correlation of negatively worded search queries on perceived immigration threats with voting intentions for a virulent anti-immigrant party. Their study shows not only what kinds of threat perceptions are associated most closely with the electoral fortunes of right-wing populism, but also that immigration-related concerns translate into anti-immigrant voting preferences with several months of delay (at least in the particular case examined here—voting intentions for the "Alternative für Deutschland" party from 2013 through 2019). Since it is difficult to see how these findings could have been obtained with traditional survey instruments, they highlight the virtues of Internet-based research tools (granularity, customizability, timeliness, etc.). Also, while little knowledge exists so far on the potential inhibitions of choosing query terms, search engine use seems less prone to social desirability bias than surveys, even those self-administered online. However, this chapter also suggests that, with a view to detecting individual-level covariates or determinants, surveys continue to be an important part of researchers' toolkits.

11.3 The New Frontier: Distilling Knowledge from Accrued Data

To point out the obvious, the papers collected in this volume do not constitute a representative sample of migration scholars' use of innovative technology. Even if that claim had been plausible at the book's inception, its veracity would have

diminished inevitably by the time of its publication. Since the opportunities afforded to researchers by emerging technologies and the uses thus enabled by these options are constantly evolving, any conclusions related to the subject-matter of this book are necessarily tentative. Such caution regarding the adoption of new technologies applies to any research domain, yet it seems especially appropriate with regard to migration studies. For example, in comparison to the vibrant exchange between survey researchers and computational social scientists at the biannual Big Data Meets Survey Sciences (BigSurv) conferences (see Hill et al., 2020 for a collection of papers from the first such event), the migration research community appears to be relatively slow at taking up new technological options, especially with regard to mining the web for actual insights. Although we can only speculate about the reasons, it seems plausible to assume that data access issues, on the one hand, and insufficient data management skills, on the other, contribute decisively to this situation. Many migration researchers would probably argue that the complex nature of their field of study requires data of a more qualitative kind than those traced, in one way or another, on the Web. However, in our view, this line of argument underestimates the huge potential of next-generation multi-method approaches.

Admittedly, a comparison to the avant-garde of interdisciplinary cross-fertilization between data scientists and survey researchers is somewhat unfair to migration scholars. However, setting the bar high seems appropriate to raising the game, so this book has aimed at spurring some added diligence (for a complementary effort see Salah et al., forthcoming). Initiatives such as the Big Data for Migration Alliance (https://data4migration.org) and the HumMingBird project (https://hummingbird-h2020.eu) suggest that the migration-research landscape could change particularly fast in the coming years due to a combination of a relatively low "market share" of ICTs and other innovative technologies (as gauged by the papers presented at IMISCOE's annual conferences, for example), on one hand, and the added mileage that can be obtained potentially by their adoption, on the other. This book does not aim to provide a step-by-step guide to competence-building, but rather offers an incentive for migration researchers to constantly assess what kind of empirical evidence is most appropriate for achieving their goals, and how to obtain such data.

The collection's nine chapters (not including this closing chapter and the introduction) exemplify, or showcase, two main ways that innovative technology may contribute to enhancing the methodological arsenal of migration studies. The five contributions gathered in the book's first part explore how ICT and other emerging technologies help to improve the viability and quality of conceptually traditional studies, i.e., researcher-defined data collections pursuing extrapolation with respect to oftentimes very specific target populations. Despite their intrinsic difficulties, inescapable limitations, and considerable cost, such sample-based studies continue to be held in high esteem by academics and migration-managing institutions alike. This is true especially in times of intensifying migration flows. However, even in those countries with excellent systems of public statistics, information on newcomers' characteristics, needs, and skills cannot be delivered adequately by extant administrative sources or general-population surveys, nor can qualitative studies

alone provide the input needed for planning and implementing the services and procedures required in such circumstances. More generally, academics, practitioners, and evidence-focused policymakers cherish the possibility of converting survey items into predictors of key outcome variables. With due respect for other sources and approaches, surveys of migrant populations may therefore be seen to have been the linchpin or "frontier" of migration studies in recent decades. Although some of the techno-methodological options described in this book's first part (e.g., the use of social network sites for respondent recruitment) are of interest for qualitative data collections as well, their main field of application is the improvement of migrant surveys.

Shifting gears, the four contributions to the collection's second part address researchers' use of data that, instead of relying on samples, cover entire populations (at least in principle), and instead of deriving from specified research designs, were produced for administrative or purely operative reasons. In the book's second part, such data come into focus not with respect to sample design and implementation, but in terms of the actual clues they provide about people's behaviors and mindsets.

Even though social scientists have extracted information from administrative records and censuses for decades, the Internet now enables access to a previously unimaginable wealth of such sources. This situation opens up exciting new research options, especially when various datasets can be linked, and provided that privacy can be protected and informed consent achieved across an enlarged user base—two big "ifs" that require careful attention (Ramos and Haan, in this volume). Even in light of these constraints, administrative records represent the more accessible part of the new data universe, since they have been generated in purposeful ways and are conceptually rather similar to researcher-defined data collections (Connelly et al., 2016). This affinity is illustrated by the fact that a large share of the contributions to Hill et al. (2020) refer to the combination of survey data with administrative datasets.

In contrast, from a researcher's point of view, most types of Internet data are messy and oftentimes arcane. Different from censuses, which collect the same (relatively sparce) information from a given territory's residents at long intervals, and different from administrative records, which refer to well-defined (yet typically rather isolated) events or procedures, Internet data transcend traditional time-space restrictions and accustomed notions of what counts as data in the first place. In the new data universe, space and time cease to be constitutive features of a study's data collection plan, since they become customizable parameters for data extraction. By the same token, the traditional business of operationalizing variables of interest gives way to the perhaps even more challenging task of *repurposing* extant information, an endeavor that presupposes a capacity to identify relevant bits of information among overwhelming quantities of non-sense. Also, once a pattern is recognized, it requires interpretation, yet the dataset may (and most likely will) lack information concerning *explanatory* variables.

In short, *big data* is a categorically different kind of input into the research process – with far-reaching consequences. The specialized literature offers a range of descriptive adjectives, including *organic* (Groves, 2011), *found* (Connelly et al., 2016) and *non-sampled* (Hill, 2020). We would like to suggest the term *accrued*

data as an addition to this semantic cluster to highlight their status as mere derivatives of behaviors or even, increasingly, interconnected digital gadgets. Whatever the label, it seems vital to flag the categorical difference separating purposefully designed research data from data that lack intrinsic meaning, relevance, and even intelligibility.

The frontier has moved.

In our view, the leap from more or less circumscribed, conceptually-driven data collections to unbounded, operationally-driven data generation entails differences at least as momentous as those between the two broad categories of research designs (qualitative *versus* quantitative) that have co-existed, competed, and complemented one another in the social sciences throughout the past century or so. Of course, both survey-based and internet-based data require (and support) computer-assisted quantitative analyses. However, the accustomed statistical techniques are inappropriate for handling the enormous, and exponentially growing, volume of data that are by-products of an ever-expanding range of everyday activities (Spiegelhalter, 2014). As the role of automated computation increases, so does the incidence of spurious correlations. In the emerging world of data-driven social sciences, a crucial challenge is to derive *meaningful* knowledge from a deluge of information that undermines acquainted notions of pertinence and transparency (Kitchin, 2014).

The abundance of real-time data at zero or very low marginal cost risks obscuring, rather than highlighting, significant patterns. The sheer volume of information may blind researchers to its skewed nature. For various reasons, it would be misleading to understand these new data as mere reflections of the world as is (Vinck et al., 2019). Digital devices, software applications, and algorithms are not only conceived for specific purposes and subject to coverage, malfunctioning, and/or misreporting biases of oftentimes unknown magnitude, but also are designed and written by human beings and therefore, directly or indirectly, are influenced by their creators' socializations and cultural backgrounds. Furthermore, the predominance of private-sector data generates access hurdles and secrecy regarding fundamental definitions and procedures, as is the case with proprietary algorithms. In short, the opportunities afforded by big data's volume, velocity, and variety (cf. Laney, 2001) also entail pressing concerns regarding validity and veracity (McCoach et al., 2020).

The skills honed with a view to collecting and analyzing researcher-defined datasets are anything but obsolete in this context; instead the opposite is the case. As Hill et al. (2020) have stressed, initial excitement about big data's competitive advantages in terms of scope, timeliness, and cost has given way to the realization that many of its most notorious challenges can be reframed in terms such as coverage bias, imputation error, or total error (Biemer & Amaya, 2020; Sen et al., 2021). With respect to the interdisciplinary cooperation envisioned by Hill et al. (2020), data science requires survey researchers' input especially with regard to conceptualization, context, and data quality. Also, domain expertise of a more qualitative nature will continue to be indispensable for interpreting any patterns observed in organic data (Salah et al., 2019). These considerations suggest that survey research (and other "low-volume" data types and techniques) will continue to play a relevant role in the methodological portfolio of the social sciences in general and migration

research in particular. However, we anticipate that their contributions will be defined increasingly by next-generation "mixed methods" approaches. In a research landscape increasingly shaped by creative combinations of sampled and non-sampled data (Groves, 2011) and highly interdisciplinary teams (King, 2014), administrative data will be used to ease the burden for respondents, just as surveys will be employed to add depth and context to the behavioral and attitudinal patterns revealed by digital traces. Historical retrospect supports this prediction: throughout the second half of the twentieth century, rather than being supplanted by the increasing sophistication of surveys, qualitative studies continued to play their part. Just as in-depth interviews or focus-groups can offer more fine-grained and contextualized insights into motivations than surveys, so can surveys provide better insights into motivations when compared to organic data.

Research ethics is a second area in which accrued data represent a momentous departure from established procedures. With regard to surveys, administrative datasets, and censuses, ethical considerations have long been guided by the principles of privacy protection and informed consent. Since migrants are oftentimes vulnerable to abuse and discrimination, rigorous procedures are required to prevent any possibility of privacy breaches. However, in the brave new world of data mining, these principles are both partially unviable and patently insufficient, at least as traditionally conceived. As Deutschmann et al. and Lorenz et al. illustrate in their contributions to this collection, *a posteriori* repurposing of data may not be covered by explicit authorizations from their original sources (in these two specific cases, international travelers and search engine users), and such consent may not be necessarily pertinent with respect to huge aggregate datasets that do not contain any *personal* information. In other contexts, especially when user-created content is at the center of analysis, reflection may be needed on the exact nature of any given consent. More specifically, a question arises as to whether approaches that are technically covered by services' terms of use nevertheless need additional and specific consent by individuals when they become participants in, or objects of, research (Leurs & Prabhakar, 2018; Vinck et al., 2019). However, the need for ethical conduct transcends the personal sphere, for example, automated big-data analysis may reveal the real-time location of vulnerable groups. The new data universe requires an anticipatory incorporation of ethical considerations into computing procedures ("ethics by algorithm", cf. Dignum, 2018). The emerging consensus in the debate on big-data ethics demands complementing the range of principles that govern bioethics (beneficence, non-maleficence, autonomy, and justice) with an added imperative of explicability: the procedures of artificial intelligence must be made intelligible to lay citizens, and accountability for any misguided outcomes must be assured (Floridi et al., 2018). In more general terms, the use of accrued data, especially, yet not exclusively, in research on migrants and other minority groups, necessitates that scholars reflect on the potentially unintended consequences of their research. Upholding the "do no harm" principle must always take precedence over scientific curiosity and the allure of testing innovative procedures of knowledge production. Thus, on the one hand, scholars need to contemplate the consequences of publishing specific findings, especially on vulnerable

subpopulations such as irregular migrants. On the other hand, researchers need to consider that political actors may employ newly developed methods to carry out their own agendas (Franklinos et al., 2020; Vinck et al., 2019). Another challenge is to make the data accrued by private corporations accessible to scientific repurposing. The forthcoming regulatory battle at this regard presupposes the public's trust in appropriate ethical safeguards.

To resume, we envision a future of multi-disciplinary collaboration where migration specialists with various disciplinary backgrounds, survey methodologists, and data scientists upgrade and transfer skills with one another. If pushed to single out one take-home message from this book's contributions, we advise migration researchers to actively participate in the process of multi-disciplinary skill building in terms of project design, networking, and human resources development (training and staffing). Without wishing to scare people into action, the history of technology abounds in examples of doom spelled by the failure to adapt to disruptive technology. Without any hype, it seems safe to rate ubiquitous digitization among the most disruptive waves of technological innovation ever.

That disruption continues relentlessly as Internet-wary cohorts diminish and ever-increasing shares of daily behaviors become subject to digitized scrutiny. In time, sample-based data collections might come to be seen as anachronistic. Remember those chemistry-based photographs?

References

Biemer, P. P., & Amaya, A. (2020). Total error frameworks for found data. In C. A. Hill, P. P. Biemer, T. D. Buskirk, L. Japec, A. Kirchner, S. Kolenikov, & L. E. Lyberg (Eds.), *Big data meets survey science* (pp. 131–161). Wiley. https://doi.org/10.1002/9781118976357.ch4

Connelly, R., Playford, C. J., Gayle, V., & Dibben, C. (2016). The role of administrative data in the big data revolution in social science research. *Social Science Research, 59*, 1–12. https://doi.org/10.1016/j.ssresearch.2016.04.015

Dignum, V. (2018). Ethics in artificial intelligence: Introduction to the special issue. *Ethics and Information Technology, 20*(1), 1–3. https://doi.org/10.1007/s10676-018-9450-z

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines, 28*(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5

Franklinos, L., Parrish, R., Burns, R., Caflisch, A., Mallick, B., Rahman, T., Routsis, V., Sebastián López, A., Tatem, A., & Trigwell, R. (2020). *Key opportunities and challenges for the use of big data in migration research and policy* [Preprint]. https://doi.org/10.14324/111.444/000042.v1

Groves, R. M. (2011). Three eras of survey research. *Public Opinion Quarterly, 75*(5), 861–871. https://doi.org/10.1093/poq/nfr057

Hill, C. A. (2020). Moving social science into the fourth paradigm. In C. A. Hill, P. P. Biemer, T. D. Buskirk, L. Japec, A. Kirchner, S. Kolenikov, & L. E. Lyberg (Eds.), *Big data meets survey science* (pp. 713–731). Wiley. https://doi.org/10.1002/9781118976357.ch24

Hill, C. A., Biemer, P. P., Buskirk, T. D., Japec, L., Kirchner, A., Kolenikov, S., & Lyberg, L. E. (Eds.). (2020). *Big data meets survey science. A collection of innovative methods* (1st ed.). Wiley. https://doi.org/10.1002/9781118976357

King, G. (2014). Restructuring the social sciences: Reflections from Harvard's Institute for Quantitative Social Science. *PS: Political Science and Politics, 47*(1), 165–172. https://doi.org/10.1017/S1049096513001534

Kitchin, R. (2014). Big data, new epistemologies and paradigm shifts. *Big Data & Society, 1*(1), 2053951714528481. https://doi.org/10.1177/2053951714528481

Krumpal, I. (2013). Determinants of social desirability bias in sensitive surveys: A literature review. *Quality & Quantity, 47*(4), 2025–2047. https://doi.org/10.1007/s11135-011-9640-9

Laney, D. (2001). 3-D data management: Controlling data volume, velocity and variety. *META Group Research Note, 6*(70), 1.

Leurs, K., & Prabhakar, M. (2018). Doing digital migration studies: Methodological considerations for an emerging research focus. In R. Zapata-Barrero & E. Yalaz (Eds.), *Qualitative research in European migration studies* (pp. 247–266). Springer. https://doi.org/10.1007/978-3-319-76861-8_14

Majó, J. (2012). Evolución de las tecnologías de la comunicación. In M. de Moragas (Ed.), *La comunicación: De los orígenes a internet* (pp. 65–89). Gedisa.

McCoach, D. B., Dineen, J. N., Chafouleas, S. M., & Briesch, A. (2020). Reproducibility in the era of big data: Lessons for developing robust data management and data analysis procedures. In C. A. Hill, P. P. Biemer, T. D. Buskirk, L. Japec, A. Kirchner, S. Kolenikov, & L. E. Lyberg (Eds.), *Big data meets survey science* (1st ed., pp. 625–655). Wiley. https://doi.org/10.1002/9781118976357.ch21

Salah, A. A., Korkmaz, E. E., & Bircan, T. (forthcoming). *Data science for migration and mobility*. Oxford University Press.

Salah, A. A., Pentland, A., Lepri, B., & Letouzé, E. (2019). *Guide to mobile data analytics in refugee scenarios. The "data for refugees challenge" study*. Springer. https://doi.org/10.1007/978-3-030-12554-7

Sen, I., Flöck, F., Weller, K., Weiß, B., & Wagner, C. (2021). A Total error framework for digital traces of human behavior on online platforms. *Public Opinion Quarterly, 85*(S1), 399–422. https://doi.org/10.1093/poq/nfab018

Spiegelhalter, D. J. (2014). The future lies in uncertainty. *Science, 345*(6194), 264–265. https://doi.org/10.1126/science.1251122

Tourangeau, R., Edwards, B., Johnson, T. P., Wolter, K. M., & Bates, N. (2014). *Hard-to-survey populations*. Cambridge University Press.

Vinck, P., Pham, P. N., & Salah, A. A. (2019). "Do no harm" in the age of big data: Data, ethics, and the refugees. In A. A. Salah, A. Pentland, B. Lepri, & E. Letouzé (Eds.), *Guide to mobile data analytics in refugee scenarios. The "data for refugees challenge" study* (pp. 87–99). Springer. https://doi.org/10.1007/978-3-030-12554-7_5