

Hans-Peter Blossfeld
Jutta von Maurice
Michael Bayer
Jan Skopek *Editors*

Methodological Issues of Longitudinal Surveys

The Example of the National
Educational Panel Study

Methodological Issues of Longitudinal Surveys

Hans-Peter Blossfeld • Jutta von Maurice
Michael Bayer • Jan Skopek (Eds.)

Methodological Issues of Longitudinal Surveys

The Example of the National Educational
Panel Study

Editors

Hans-Peter Blossfeld
European University Institute
Florence, Italy

Michael Bayer
LIfBi
Bamberg, Germany

Jutta von Maurice
LIfBi
Bamberg, Germany

Jan Skopek
European University Institute
Florence, Italy

ISBN 978-3-658-11992-8

ISBN 978-3-658-11994-2 (eBook)

DOI 10.1007/978-3-658-11994-2

Library of Congress Control Number: 2016932501

Springer VS

© Springer Fachmedien Wiesbaden 2016

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made.

Printed on acid-free paper

This Springer VS imprint is published by Springer Nature

The registered company is Springer Fachmedien Wiesbaden GmbH

Content

Hans-Peter Blossfeld, Jutta von Maurice, Michael Bayer and Jan Skopek
Foreword XI

I. Introduction

Jutta von Maurice, Hans-Peter Blossfeld and Hans-Günther Roßbach
The National Educational Panel Study:
Milestones of the Years 2006 to 2015 3

Frank J. Infurna, Denis Gerstorf, Nilam Ram and Jutta Heckhausen
Analytic Strategies for the Study of Adaptation to Major Life Events:
Making the Most of Large-Scale Longitudinal Surveys 19

II. Sampling, Recruiting, and Fieldwork Management

Hans Walter Steinhauer and Sabine Zinn and Christian Aßmann
Weighting Panel Cohorts in Institutional Contexts 39

Sabine Zinn
Variance Estimation with Balanced Repeated Replication:
An Application to the Fifth and Ninth Grader Cohort Samples
of the National Educational Panel Study 63

<i>André Müller-Kuller, Sonja Meixner and Michaela Sixt</i> Challenges in Gaining Access: The School Cohorts of the National Educational Panel Study	85
<i>Ina-Sophie Ristau and Stephanie Beyer</i> Cooperation and Communication Within Scientific Organizations: The Role of Survey Coordination	99
<i>Michaela Sixt, Martin Goy and Georg Besuch</i> The Concept of Individual Retracking in NEPS—Approach, Practice, and First Empirical Evidence From Starting Cohorts 3 and 4	111
<i>Götz Lechner, Julia Göpel and Anna Passmann</i> Challenges and Intentions of Target-Specific Public Relations Work	133
III. Longitudinal Measurement of Educational Processes: Surveys and Constructs	
<i>Anja Sommer, Claudia Hachul and Hans-Günther Roßbach</i> Video-Based Assessment and Rating of Parent-Child Interaction Within the National Educational Panel Study	151
<i>Doreen Müller, Tobias Linberg, Michael Bayer, Thorsten Schneider and Florian Wohlkinger</i> Measuring Personality Traits of Young Children— Results From a NEPS Pilot Study	169
<i>Florian Wohlkinger, Michael Bayer and Hartmut Ditton</i> Measuring Self-Concept in the NEPS	181
<i>Cornelia Kristen, Melanie Olczyk and Gisela Will</i> Identifying Immigrants and Their Descendants in the National Educational Panel Study	195
<i>Johann Carstensen, Anja Gottburgsen and Monika Jungbauer-Gans</i> Measuring Health in a Longitudinal Education Study	213
<i>Christiane Gross and Katharina Seebaß</i> The Standard Stress Scale (SSS): Measuring Stress in the Life Course	233

<i>Lena Nusser, Jana Heydrich, Claus H. Carstensen, Cordula Artelt and Sabine Weinert</i>	
Validity of Survey Data of Students with Special Educational Needs— Results From the National Educational Panel Study	251
<i>Hildegard Schaeper and Thomas Weiß</i>	
The Conceptualization, Development, and Validation of an Instrument for Measuring the Formal Learning Environment in Higher Education . . .	267
<i>Kerstin Hoenig, Reinhard Pollak, Benjamin Schulz and Volker Stocké</i>	
Social Capital, Participation in Adult Education, and Labor Market Success: Constructing a New Instrument	291
<i>Gunther Dahm, Oliver Lauterbach and Sophie Hahn</i>	
Measuring Students' Social and Academic Integration—Assessment of the Operationalization in the National Educational Panel Study	313
<i>Katrin Drasch, Corinna Kleinert, Britta Matthes and Michael Ruland</i>	
Why Do We Collect Data on Educational Histories Over the Life Course the Way We Do? Core Questionnaire Design Decisions in Starting Cohort 6—Adults	331
<i>Annette Trahms, Britta Matthes and Michael Ruland</i>	
Collecting Life-Course Data in a Panel Design: Why and How We Use Proactive Dependent Interviewing	349
<i>Michael Ruland, Katrin Drasch, Ralf Künster, Britta Matthes and Angelika Steinwede</i>	
Data-Revision Module—A Beneficial Tool to Support Autobiographical Memory in Life-Course Studies	367
<i>Florian Janik, Oliver Wölfel and Merlind Trepesch</i>	
Measurement of Further Training Activities in Life-Course Studies	385

IV. Longitudinal Measurement of Skills: Competence Testing

<i>Karin Berendes and Sabine Weinert</i>	
Selecting Appropriate Phonological Awareness Indicators for the Kindergarten Cohort of the National Educational Panel Study: A Theoretical and Empirical Approach	401

<i>Stephan Jarsinski, Sarah Frahm, Inge Blatt, Wilfried Bos and Michael Kandera</i> Assessing Spelling Competence Development in the National Educational Panel Study	427
<i>Aileen Edele, Kristin Schotte and Petra Stanat</i> Assessment of Immigrant Students' Listening Comprehension in Their First Languages (L1) Russian and Turkish in Grade 9: Test Construction and Validation	441
<i>Kathrin Lockl, Marion Händel, Kerstin Haberkorn and Sabine Weinert</i> Metacognitive Knowledge in Young Children: Development of a New Test Procedure for First Graders	465
<i>Anna Südkamp, Steffi Pohl, Jana Heydrich and Sabine Weinert</i> Including Students With Special Educational Needs in the Competence Assessment of the NEPS— Results on the Comparability of Test Scores in Reading	485
<i>Christian Aßmann, Christoph Gaasch, Steffi Pohl and Claus Carstensen</i> Estimation of Plausible Values Considering Partially Missing Background Information: A Data Augmented MCMC Approach	503
<i>Kerstin Haberkorn, Steffi Pohl, Claus Carstensen and Elena Wiegand</i> Scoring of Complex Multiple Choice Items in NEPS Competence Tests	523
 V. Assessing Data Quality	
<i>Thomas Bäumer and Hans-Günther Roßbach</i> Measurement of Preschool Quality Within the National Educational Panel Study—Results of a Methodological Study	543
<i>Cornelia Gresch, Rolf Strietholt, Michael Kandera and Heike Solga</i> Reading-Aloud Versus Self-Administered Student Questionnaires: An Experiment on Data Quality	561
<i>Franziska Fellenberg, Heiko Sibbers, Birgit Jesske and Doris Hess</i> Quality Assurance in the Context of Data Collection	579

VI. Data management, Coding, Dissemination, and User Support

<i>Jan Skopek, Knut Wenzig, Daniel Bela, Tobias Koberg, Manuel Munz and Daniel Fuß</i>	
Data Dissemination, Documentation, and User Support	597
<i>Jan Skopek, Tobias Koberg and Hans-Peter Blossfeld</i>	
RemoteNEPS—An Innovative Research Environment	611
<i>Knut Wenzig, Christian Matyas, Daniel Bela, Ingo Barkow and Marc Rittberger</i>	
Management of Metadata: An Integrated Approach to Structured Documentation	627
<i>Daniel Bela</i>	
Applied Large-Scale Data Editing	649
<i>Jan Skopek and Manuel Munz</i>	
Life-Course Data and the Longitudinal Classification of Education	669
<i>Tobias Koberg</i>	
Disclosing the National Educational Panel Study	691
<i>Manuel Munz, Knut Wenzig and Daniel Bela</i>	
String Coding in a Generic Framework	709
<i>Ralf Künster</i>	
Visualizing Life Courses With the TrueTales View	727

Foreword

Hans-Peter Blossfeld, Jutta von Maurice, Michael Bayer and Jan Skopek

Introduction

This book is the second volume focusing on important methodological issues of longitudinal studies using the example of the National Educational Panel Study (NEPS) in Germany. Today, the NEPS is one of the biggest longitudinal data-collection endeavors in social sciences in Europe and even beyond. The first volume described the main research aims, the basic design, the organization, and the setup of the NEPS (Blossfeld, Roßbach, & von Maurice, 2011). In this second volume a rich compendium documenting important methodological challenges, solutions, and achievements that emerged in developing a major longitudinal study are extensively described and discussed.

The aim of the NEPS is to collect rich large-scale longitudinal data on life courses, in particular the educational careers and competence developments of individuals and their consequences in terms of health and political behavior, career pathways, job success, employment behaviors, and income trajectories from early childhood to late adulthood. The basic survey design of the NEPS—a multicohort sequence design—involves six large independent panel samples (the so-called starting cohorts that are then followed-up in regular data sweeps over long time spans. In 2009, the NEPS started to collect data on (1) 6-month-old babies (Early Childhood cohort), (2) children in Kindergarten 2 years before regular school enrolment, (3) fifth graders at the age of about 10, (4) ninth graders (the 15-year-olds that are also analyzed in the PISA study by the OECD), (5) first-year students in higher education, that is, at traditional universities and universities of applied sciences, and (6) adults at the age of 23 to 64. In addition, the NEPS conducted additional secondary school studies in two selected German Federal States. The NEPS has developed and implemented a comprehensive range of longitudinal survey instruments and competence tests, sampling strategies, fieldwork procedures as well as an infrastructure for data edition, data dis-

semination, and user support. More than 200 scientists from different disciplines such as sociology, psychology, education sciences, economics, demography, statistics, and experts in sociological research methods are working on the NEPS. In January 2014, the NEPS project was institutionalized as a Leibniz Institute for Educational Trajectories (Leibniz-Institut für Bildungsverläufe, LIfBi). This support of the government ensures a long-term data infrastructure for national and international educational research in Germany. The total number of target persons included in the NEPS longitudinal study is about 60,000. In addition, educators, teachers, school principals, and parents associated with these 60,000 target persons are interviewed in order to include their familial, regional, and school contexts. Since 2012, a remarkable number of Scientific Use File data sets have been released to the international scientific community. Today, the number of scholars around the world who are using NEPS data for longitudinal empirical research has increased to more than 1,000 users. Consequently, the NEPS has become the most important data source for sociological, educational, economical, and psychological longitudinal research in Germany and beyond.

By now, several years after the start of the NEPS, an abundance of methodological challenges have been mastered and valuable knowledge about new solutions and tools have been developed for the NEPS. The aim of this volume entitled “*Methodological Issues of Longitudinal Surveys—The Example of the National Educational Panel Study*” is to address important user-relevant issues of the NEPS. The central idea of this book is to report and discuss the specific methodological problems of longitudinal studies and the practical solutions that have been found in the various NEPS disciplines while building up an attractive, efficient, and powerful large-scale multicohort panel database. In particular, the book demonstrates new standards in the collection and distribution of large-scale longitudinal data. In a nutshell, the 40 short and to-the-point chapters in this book capture a broad variety of relevant methodological issues ranging from sampling and weighting, recruiting and fieldwork management, designing longitudinal surveys, constructs, and competence tests, improving data quality, editing and documenting data on a large-scale basis, disseminating data to researchers, as well as establishing an effective public relations and communications service for a large panel study. Addressing an impressive array of methodological challenges and solutions, 93 authors—all of them longitudinal experts from different fields and backgrounds—have contributed to this unique volume.

The Approach of the Book

A key goal of the book is the discussion of important methodological challenges in today’s longitudinal designs and suggestions for their practical solutions as they have been achieved by the NEPS. Hence, contrary to other books on the market, this book is not intended to be just another theoretical primer in survey research. Rather, this book presents a well-selected collection of applied methodological topics and prac-

tical issues that had to be solved in building up a large-scale survey project but are hardly ever discussed in any available textbooks on survey research today. For instance, the book will provide not only chapters on sampling, weighting, and measurement of concepts in the context of longitudinal designs, but also on topics such as how to practically access and follow up target populations in a school sampling context, how to coordinate and manage multiple surveys, how to build up target-specific public relations services, or how to establish the highest standards of quality management in the context of longitudinal data collection. Moreover, the book provides a variety of valuable contributions for users of longitudinal data in the field of data management, dissemination, and user support—all of which are undoubtedly crucial for modern longitudinal survey projects and for the NEPS users, but which are still virtually untouched in the current literature.

Beyond sampling and data-collection issues, a core focus of this book is the longitudinal measurement of educational processes and skills over the life course. Several chapters cover a series of innovative methodological approaches that have been implemented in the NEPS, such as dependent interviewing for seamlessly collecting life-course data, video-based assessments of early childhood behavior, or measuring migration background, personality traits, health, stress, or further training activities. A major mission of the NEPS is the longitudinal assessment of competencies and skills of age-graded populations on a representative basis, which is largely uncharted territory in psychometrics.

This volume mainly targets an audience of survey researchers, practitioners in survey methodology, and the broader scientific community using the NEPS and other longitudinal data for their analyses. In general, it will be interesting for applied life-course researchers, psychologists, demographers, sociologists, economists, and educational researchers who are interested in large-scale assessments and educational careers. Consistent with the strategy of tackling real-life methodological problems in large-scale surveys, the volume explicitly does not follow the approach of a conventional textbook. Rather, it serves as a reference book for applied longitudinal methodology. While connected chapters are grouped together under relevant themes, all chapters can be read independently depending on a particular reader's interest. Notwithstanding, we believe that the book may also be of great value for introducing undergraduate and postgraduate students to the longitudinal methodology of the social sciences.

Synopsis

The book is organized into six parts. A first part provides a brief introduction to the National Educational Panel Study while also reporting on important milestones that have been achieved during the establishment phase of NEPS between 2009 and 2014. Moreover, analytical strategies to advance our knowledge of how life events change

the life course and shape developmental trajectories across different educational stages are discussed. A second part of the book contains six chapters tackling crucial issues of multicohort and institutional sampling, recruiting of survey participants in a multiactor design, and management of complex multiagency fieldwork processes. A third part is dedicated to the longitudinal measurement of educational processes, one of the major challenges of the NEPS. A collection of 14 chapters touches upon innovative topics such as video-based assessment of infants, measurement of personality traits, self-concept, health, stress, social capital, multigenerational migration background, social and academic integration, as well as the collection of initial and further educational biographies using modern techniques and tools for collecting seamless life history data. Next to collecting data on educational trajectories, a second goal of the NEPS is to assess competencies and skills throughout the entire life span. Part four provides seven chapters focusing on several methodological issues in assessment and statistical scaling of competence data. Particularly, these chapters document significant new experiences in assessing competencies among more difficult target groups such as students with special educational needs or students with migration backgrounds. Part five is devoted to the assessment of data quality in the NEPS. Evidence on data quality from experimental studies is presented and the importance of quality assurance units in large-scale studies is demonstrated. The NEPS has successfully built up a robust infrastructure, not only for collecting data but also for disseminating and delivering longitudinal data to the wider scientific community. Hence, part six of the book deals with innovative methods, techniques, and tools of data management, data coding, and data dissemination in the context of a large-scale longitudinal survey project. Eight chapters deal with highly relevant questions such as how researchers need to manage and document large-scale survey data, how to disseminate data of different disclosure levels while maximizing research utility, or how to build up a powerful program for user support and training.

Hans-Peter Blossfeld, European University Institute, Florence

Jutta von Maurice, Leibniz Institute for Educational Trajectories, Bamberg

Michael Bayer, Leibniz Institute for Educational Trajectories, Bamberg

Jan Skopek, European University Institute, Florence

I. Introduction

The National Educational Panel Study: Milestones of the Years 2006 to 2015

Jutta von Maurice, Hans-Peter Blossfeld and Hans-Günther Roßbach

Abstract

Funded by the Federal Ministry of Education and Research (Bundesministerium für Bildung und Forschung—BMBF), the National Educational Panel Study (Nationales Bildungspanel—NEPS) started in August 2008 with the aim of collecting urgently required longitudinal data about educational processes and competence development from early childhood to late adulthood. As of January 2014, the NEPS is now situated at the newly founded Leibniz Institute for Educational Trajectories (Leibniz-Institut für Bildungsverläufe—LIfBi). The NEPS provides these data to the scientific community as quickly as possible after each data-collection sweep. During the years 2006 to 2015, several important milestones have been achieved by the NEPS team: First, an interdisciplinary network of excellence has been built up including the best educational researchers and research institutions in Germany—initially with the University of Bamberg as the home of the NEPS center. Second, a clear structure for the NEPS has been developed focusing on five substantively oriented pillars and eight life-course stages. Third, a multi-cohort sequence design was defined in order to be able to quickly collect and disseminate data on different educational stages and to enable an easy comparison of different cohorts. Fourth, six cohorts with more than 60,000 target persons (plus some 40,000 context persons) were sampled in educational institutions or based on register data. Fifth, innovative longitudinal instruments were designed by an interdisciplinary team of researchers bringing together relevant theories, concepts, and variables from various disciplines. Sixth, procedures in order to collect representative data based on different samples have been defined, following up individuals through their educational pathways. Seventh, an effective infrastructure for the dissemination of data to the scientific community in Germany and abroad, a program of introductory user courses, and a user support center have been set up. Data from all six NEPS cohorts have been released to date. More than

1,000 researchers from various disciplines are already using NEPS data. Eighth, the institutionalization under the umbrella of the Leibniz Association has created a long-term perspective for NEPS to establish itself as an infrastructure facility for educational research.

1 Interdisciplinary Network of Excellence

After the publication of the first PISA results, a lack of longitudinal data on educational processes and competence development has become painfully evident in Germany. As a response to this data situation, the Federal Ministry of Education and Research (Bundesministerium für Bildung und Forschung—BMBF) took the initiative to discuss with leading German researchers setting up a new panel study that would focus on educational processes over the life span. A first, very preliminary draft of an educational panel study had already been presented by Fickermann (BMBF) in 2004. These discussions were also accompanied by a paradigmatic change in the way in which research was going to be funded by the BMBF. The BMBF did not only try to initiate research projects and research programs, but it committed itself to following strict scientific standards in the selection, evaluation, and funding decisions—which meant, in general, an earlier and more intensive involvement of the scientific community in the funding procedure of the Ministry (Buchhaas-Birkholz, 2009).

After 2004, the BMBF and several leading researchers started to further develop the idea of collecting longitudinal data about educational processes and competence development from early childhood to late adulthood. Based on this aim it became obvious that many different disciplines and experts with profound expertise on the various educational stages (such as early childhood, school age, age of vocational and university study choices, participation in university, vocational training, and lifelong learning) would have to be involved. However, several attempts to initiate a consortium of educational experts failed.

In the summer of 2006, Hans-Peter Blossfeld was asked by the BMBF to form an interdisciplinary network with the aim of collecting representative longitudinal data on educational processes over the life course. He immediately accepted the offer and started to build up a consortium. Within this network he included not only well-known colleagues from different faculties of the University in Bamberg, but also the most prominent experts from different disciplines, as well as the most important educational research institutions from all over Germany. In 2007, first basic ideas of what a National Educational Panel Study (NEPS) could look like were written down by the interdisciplinary consortium of excellence (for the advantages of research in interdisciplinary networks see also Blossfeld & von Maurice, 2012). This first draft of the NEPS was financially supported by the BMBF and then submitted to the German Research Foundation (Deutsche Forschungsgemeinschaft—DFG). In early summer of 2007, the DFG organized a workshop with an international group of highly

renowned scientists to discuss the ideas of the NEPS Consortium. The evaluators were excited and unanimously supported the decision that an elaborated proposal for a NEPS should be worked out by the Consortium. In the summer of 2008, this full-fledged proposal was completed and submitted to the DFG. The group of international evaluators enthusiastically approved and supported the ideas of the NEPS. The NEPS was then immediately rolled out in August 2008 and officially opened in a ceremony with the Federal Minister of Education and Research (Bundesministerin für Bildung und Forschung), Annette Schavan, and the Bavarian State Minister of Sciences, Research, and the Arts (Bayerischer Staatsminister für Wissenschaft, Forschung und Kunst), Wolfgang Heubisch, in February 2009. Right from the beginning, the NEPS had been part of the Framework Programme for the Promotion of Educational Research (Rahmenprogramm zur Förderung der empirischen Bildungsforschung; Bundesministerium für Bildung und Forschung, 2008). For an overview of the NEPS please see Blossfeld, Roßbach, and von Maurice (2011).

The NEPS Consortium has been quite stable since 2009. Only small extensions were introduced—mostly in connection with main researchers within the network being appointed to other institutions. Alongside LIfBi and the University of Bamberg, there are presently 18 different institutions—with a large number of professors actively engaged in NEPS—collaborating within the NEPS network as contracted partners:

- Berlin Social Science Center (WZB)
- Centre for European Economic Research (ZEW) in Mannheim
- Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU)
- German Centre for Research on Higher Education and Science Studies (DZHW) in Hannover
- German Institute for International Educational Research in Frankfurt (DIPF)
- Humboldt-Universität in Berlin
- Ifo Institute—Leibniz Institute for Economic Research at the University of Munich
- Institute for Employment Research in Nuremberg (IAB)
- Institute for School Development Research at TU Dortmund University
- Justus Liebig University in Giessen
- Leibniz Institute for Science and Mathematics Education (IPN) at Kiel University
- Leibniz Universität Hannover
- Leipzig University
- Ludwig-Maximilians Universität in Munich (LMU)
- Universität Hamburg
- University of Mannheim
- University of Siegen, and
- University of Tübingen.

Besides these contracted partners excellent researchers from several additional institutions were integrated especially in the NEPS instrument development: european fo-

rum for migration studies at the University of Bamberg, German Institute for Adult Education—Leibniz Centre for Lifelong Learning in Bonn, German Youth Institute in Munich, Max Planck Institute for Human Development in Berlin, Ruhr University in Bochum, State Institute for Family Research at the University of Bamberg, State Institute of Early Childhood Research in Munich, Technical University of Munich, and University of Kassel.

The NEPS is highly active in building up and collaborating with other panel studies and other data infrastructure facilities in Germany and abroad. A close relationship exists with some important panel and other large-scale studies in Germany: For example, the research group Educational Processes, Competence Development, and Selection Decisions in Preschool and School Age (BiKS); the German Family Panel (pairfam); the Programme for the International Assessment of Adult Competencies (PIAAC); the Programme for International Student Assessment (PISA); the Survey of Health, Aging, and Retirement in Europe (SHARE); and—last but not least—the German Socio-Economic Panel (SOEP). Moreover, NEPS has become a member of the Leibniz Education Research Network (Leibniz-Forschungsverbund Bildungspotenziale—LERN). Besides these German collaborations there is an especially strong cooperation with similar other European or even non-European longitudinal studies: For example, the Growing up in France Study (elfe), the Millennium Cohort Study (MCS) in Great Britain, different longitudinal studies of the Australian Bureau of Statistics, the Educational Research Institute in Warsaw in Poland, the Human Sciences Research Council in South Africa, the Institute for Research and Development of Education at the Charles University in Prague in the Czech Republic, as well as the infrastructure facility Micro data Online Access (MONA) at Statistics Sweden. The aim of this cooperation is not only to continuously foster the quality of the NEPS instruments and NEPS data, but also to adjust survey instruments between different panel studies in order to allow for joint data analyses as early as possible. Moreover, methodological, survey-methodological, and technical aspects are discussed and best practice solutions are conjointly developed. The active interaction with researchers from outside the NEPS Consortium can be illustrated by more than 270 publications, more than 600 presentations, and more than 100 research visits conducted in the years 2009–2015.

In August 2012, Hans-Peter Blossfeld—who moved to a chair at the European University Institute (EUI) in Florence, Italy—handed over the position of Principal Investigator to Hans-Günther Roßbach, who has been actively involved in the NEPS since the preparatory phase.

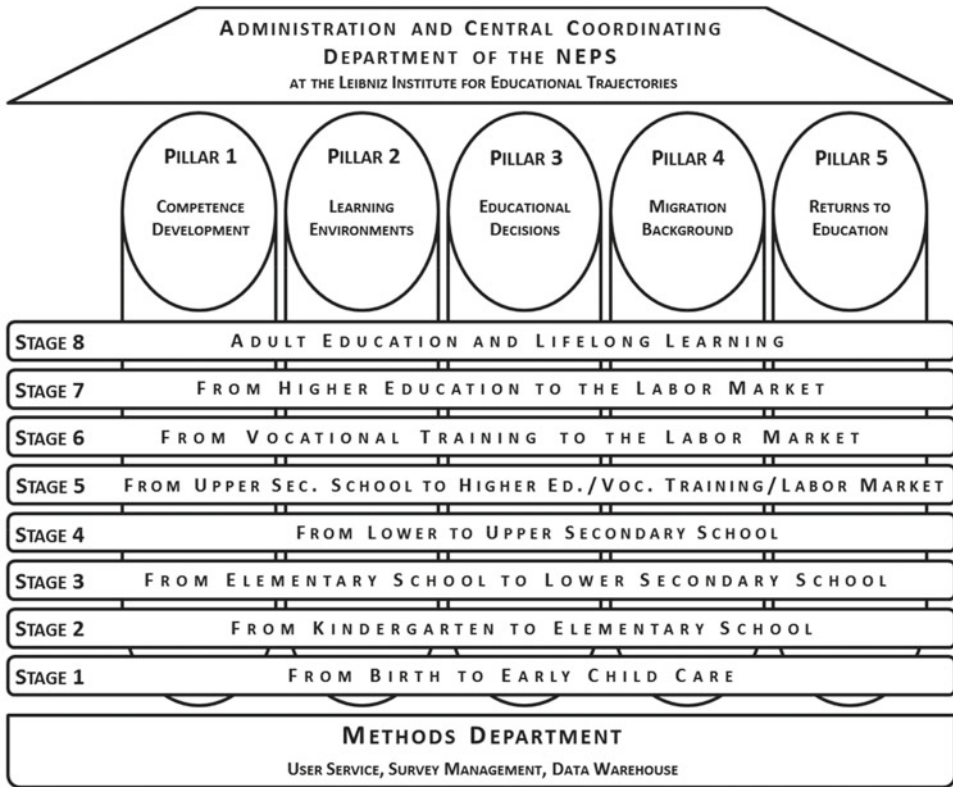
2 Setting Up a Five-Pillar and Eight-Stages Structure

The NEPS focuses on five carefully selected substantive dimensions over the entire life span. These five dimensions guarantee the homogeneity of the theoretical concepts and—as far as appropriate—the instruments used at very different educational phases from early childhood until late adulthood:

- Pillar 1 (Weinert et al., 2011; Artelt, Weinert, & Carstensen, 2013) focuses on competence development and the effects of individual competencies for educational decisions and trajectories. The competence measurements cover domain-general cognitive abilities, domain-specific cognitive competencies (with a focus on German language, mathematics, and science), metacompetencies (such as metacognition and information and communication technologies literacy), and stage-specific competencies (e. g., related to curriculum or job-related abilities and skills).
- Pillar 2 (Bäumer, Preis, Roßbach, Stecher, & Klieme, 2011) deals with the various formal, informal, and nonformal learning environments within the NEPS. The team of Pillar 2 includes the quantity and quality of the various learning environments in their analyses and also focuses on transitions between as well as cumulating effects of different learning environments.
- Pillar 3 (Stocké, Blossfeld, Hoenig, & Sixt, 2011) is concerned with educational decision-making over the entire life span and with measuring the effects of social inequality. The team of Pillar 3 is also responsible for the design and collection of sociodemographic data within the NEPS.
- Pillar 4 (Kristen et al., 2011) focuses on the situation of people with a migration background in the different educational stages of the NEPS. There is a specific emphasis on the effects of the mother language and the available networks of migrants.
- Pillar 5 (Gross, Jobst, Jungbauer-Gans, & Schwarze, 2011) addresses the economic returns to education (such as income and career trajectories) as well as non-economic benefits (such as satisfaction, health, participation) in the different life stages.

The central theoretical concepts of all five pillars are implemented from early childhood to late adulthood—keeping the measurements as comparable as possible. Additionally, personality aspects and motivational concepts are also integrated in the NEPS instruments in order to supplement the five-pillar structure (Wohlkinger, Ditton, von Maurice, Haugwitz, & Blossfeld, 2011). The educational phases over the life span are structured according to eight stages, giving the NEPS a strong internal structure (see Figure 1).

Figure 1 Pillars and stages of the National Educational Panel Study

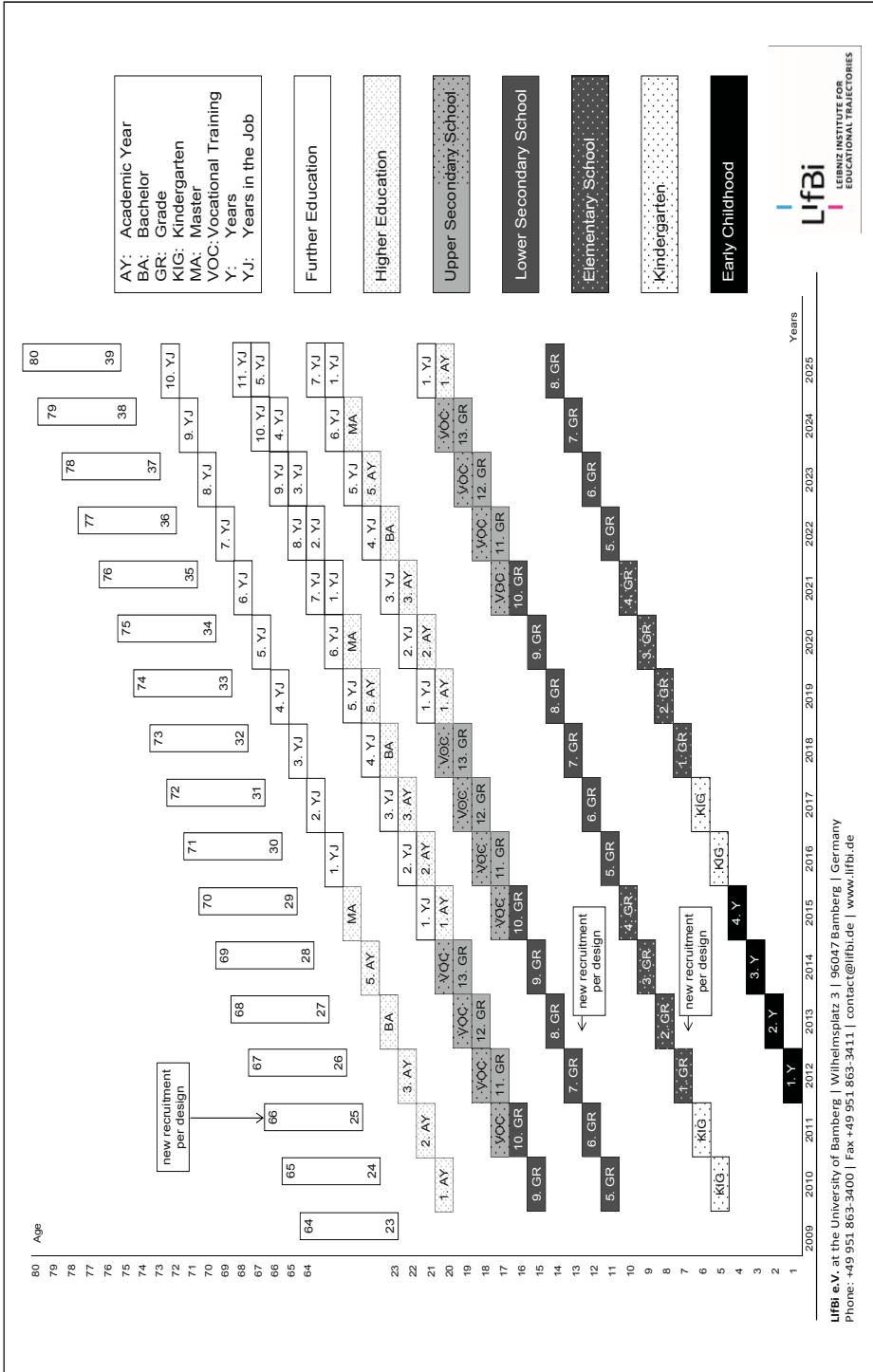


3 Multicohort Sequence Design

Studying educational processes and competence development over the life span and following the basic principles of the life-span perspective (Elder, Johnson, & Crosnoe, 2003), a longitudinal design is called for (see Blossfeld & von Maurice, 2011). The NEPS Consortium has implemented six starting cohorts along carefully selected transition points in the individual educational life course. All six cohorts started between 2009 and 2012, allowing for the development of integrated instruments and the joint specification of relevant data-collection procedures (see Figure 2).

Figure 2 demonstrates the basic design of the NEPS—pointing out the idea of following individuals over their individual life spans. Two aspects are not visible in Figure 2: First, NEPS tracks individuals irrespective of standard educational careers. Hence, also “departing and individual” trajectories are represented in the NEPS data (e.g., early school enrollment, repetition of classes, and dropout from higher education). Second, before bringing the main NEPS studies into the field, several steps of

Figure 2 Multicohort sequence design of the National Educational Panel Study



instrument development (especially in the area of competence measurement all domain-specific tests have to be newly developed by the NEPS Consortium), constant instrument improvement, studies of mode effects, linking studies of competence tests across successive age groups, as well as pilot studies for field procedures, and the appropriateness of materials (e.g., information material, testing material, and incentives) must be accomplished. Besides the main studies of the NEPS, about 100 pilot or preliminary studies are part of the data-collection plan for 2009–2015.

4 Samples

The six cohorts of the NEPS contain more than 60,000 target persons and, additionally, nearly 40,000 context persons (such as parents, caregivers, educators, teachers, and principals). All samples are carefully drawn as individual or institutional samples based on well-documented selection routines. They are all representative for the situation in Germany (see also Aßmann et al., 2011):

- Starting Cohort 1—Early Childhood (SC1) is drawn as an individual sample of children born between February and June 2012 (via population registers). A first wave started in August 2012, collecting data when the children were about six to eight months old. We have direct measurements of the children’s competencies as well as measures of parent-child interaction. More information—for example, about the children’s and families’ background, extrafamilial care arrangements, health, and joint activities of mothers and children—are collected in a computer-assisted personal interview (CAPI) with the mothers. The sample size at Wave 1 is nearly 3,500 (for Stage 1 see Schlesiger, Lorenz, Weinert, Schneider, & Roßbach, 2011).
- Starting Cohort 2—Kindergarten (SC2) is drawn via an indirect sampling procedure. In 2010 a random sample of elementary schools was drawn and information about the Kindergartens that were supplying those schools with children was collected. Based on this information, a random sample of Kindergartens was drawn. This procedure was necessary, as no complete list of Kindergartens within Germany had been available and, therefore, no sampling frame on the Kindergarten base could be defined. Within the Kindergartens, those children were selected who—based on their birth date—were scheduled for school enrollment in 2012. As details of school enrollment differ between the 16 Federal States (Bundesländer), the selected range of birth dates had to be adapted to the respective regulations. It was not possible to sample complete units of Kindergarten classes, as groups in German Kindergartens are age-mixed. Data collection within the first wave started in January 2011. Key to SC2 is a direct measurement of children’s competencies (with the child’s answers being documented by a well-trained interviewer) and a computer-assisted telephone interview (CATI) with the parents. Ad-

ditionally, questionnaires (paper-based assessment, PBA) for educators and heads of staff were used in order to collect some information about the institutional learning environment. SC2 started with about 3,000 children in more than 250 Kindergartens. When entering elementary school, the sample was extended by an additional sampling procedure; the cohort was increased by 5,315 additional first graders in 2012/2013 (for Stages 2 and 3 see Berendes et al., 2011).

- Starting Cohort 3—Grade 5 (SC3) is sampled as a fifth-graders cohort on the basis of a sample frame of all schools across Germany. In two classes of Grade 5 within each sampled school (if available, of course), we have a clustered sample. NEPS uses competence tests and student PBA questionnaires, PBA questionnaires for teachers and heads, as well as CATI interviews with parents. By using this instrumentation, a broad area of topics can be addressed to the appropriate respondent, taking a multi-informant perspective into account. Wave 1 started in December 2010 and ended with a sample of more than 6,000 children from nearly 300 institutions (for Stage 4 see Frahm et al., 2011); in 2012, 2,205 additional students in Grade 7 were sampled.
- Starting Cohort 4—Grade 9 (SC4) is built up as a ninth-graders cohort in quite identical fashion as SC3 concerning sampling and instrumentation. It also started its first measurement wave in December 2010 and realized a sample of about 16,500 children within about 650 institutions. Compared to SC3, SC4 became highly complex after the two measurement waves in Grade 9, as students started to leave the school context and entered vocational training or the transition system (for Stages 4, 5, and 6 see Frahm et al., 2011; Wagner et al., 2011; Ludwig-Mayerhofer et al., 2011).
- Starting Cohort 5—First-Year Students (SC5) focuses on college freshmen and realizes a clustered sample of students from selected study areas from German universities and universities of applied sciences. NEPS uses a strong multimethod approach, combining CATI, competence measures in group-testing settings (mainly PBA but also computer-based), or online testing, as well as online questionnaires. At the first wave more than 31,000 students were recruited using postal and personal recruitment strategies. Wave 1 started in November 2010 with a CATI in which detailed information of roughly 18,000 college freshmen could be collected (for Stage 7 see Aschinger et al., 2011).
- Starting Cohort 6—Adults (SC6) is concerned with lifelong learning and adult education. The sampling procedure was complex (as a sample from 2007 could successfully be integrated) using a register-based sampling procedure of people born between 1944 and 1986. In odd measurement waves a mixed CATI-CAPI interview was conducted, whereas in even measurement waves a competence test in the respondents' homes was administered. In measurement Wave 1—starting in November 2009—a sample of about 11,500 respondents was built up; in 2011/2012 we supplemented the sample by 5,208 newly sampled adults (for Stage 8 see Allmendinger et al., 2011).

The NEPS Consortium is working hard to keep sampling, data collection, data documentation, and data dissemination procedures as comparable and appropriate as possible for all six cohorts. Special challenges (e.g., groups of respondents with a high dropout risk) are dealt with carefully.

5 Collaborative Instrument Development

In order to make the collaboration of the members of the NEPS Consortium more convenient, it was agreed to use rather standardized instrument development and instrument documentation procedures and to follow jointly agreed timetables. Clear communication procedures and responsibilities in combination with the high expertise and commitment of all teams within the Consortium are prerequisites for an effective and in-time collaboration.

The internal work and the outside communication are highly structured by the five-pillar- and eight-stages structure of the NEPS. In each main study of the NEPS, the teams of all five pillars (Pillar 1 is responsible for the competence tests and Pillars 2–5 for the questionnaires) develop tests and questions for their respective focus topic. These items are handed over to the responsible stage team. The stages add stage-specific concepts and combine the bulk of items into a draft version of the instrument, thus giving a special focus to the “script” of the complete instrument. This version is checked for length and discussed within the Consortium in several steps of instrument improvement. Finally, all supporting materials (description of procedures for approval of data protection aspects, letters and information material for participants, and training manual for interviewers) are developed for the main NEPS studies. These steps are supported by the Central Coordination Unit at the Leibniz Institute for Educational Trajectories in Bamberg, which is in close contact with the contracted data-collection institutes. It also monitors all field procedures. This team is also responsible for public relations and incentives, corporate design, formal aspects in all materials, as well as data protection regulations.

6 Data Collection

All data collection within the different preliminary and pilot studies as well as the main studies of the NEPS are conducted by two highly experienced data-collection institutes: The Data Processing and Research Center (DPC) in Hamburg (as part of the International Association for the Evaluation of Educational Achievement, IEA) is responsible for all data collection within Kindergartens and within schools; the infas Institute for Applied Social Sciences conducts all surveys in individual settings (newborns, school leavers, adults), as well as the parent interviews and data collection in the freshmen cohort. To realize a panel study such as the NEPS, efficient working

procedures have had to be developed not only between the NEPS Consortium and the responsible data-collection institutes, but also forms of collaboration between the two data-collection institutes themselves have had to be precisely defined. Especially aspects relating to data protection must be handled carefully by all the involved institutions.

All NEPS data-collection procedures are clearly documented. Special emphasis is given to a profound interviewer training, ranging from a short refresher training (of experienced interviewers already engaged in previous NEPS studies) to an intense several-days-training (for highly complex data-collection procedures especially in the newborns cohort).

In NEPS, people are followed within their respective starting cohort independent of their individual educational pathway through life. This requires effective tracking mechanisms including checklists of participants' status within the institutional contexts, address inquiry procedures by postal service, and the use of all contact information given by the individual (including phone, mobile phone, e-mail, and postal addresses). Especially the tracking of school leavers in SC4 has proven particularly challenging, as individual life courses are highly plural and transitions are multiple in this target group.

The NEPS Consortium has decided to allow for temporary dropouts of respondents. Panel progression has shown that a substantial proportion of people who did not participate in one of the NEPS waves would later reenter the NEPS in subsequent waves. A final exclusion from the sample will thus be made in most cohorts when no information about the target person can be gathered for at least two years.

To achieve a high quality in data collection, the survey institutes have introduced a number of very effective measures, such as interviewer reports, direct supervision within the CATI field, and respondents' feedback questionnaires in the CAPI setting. Moreover, the Quality Management team of the NEPS Consortium, as well as those NEPS working groups that are directly involved in the respective substudy, is also regularly engaged in shadowing—that is, observing in situ—of interviewing or testing sessions with members of their own staff.

7 Data Dissemination

The NEPS is set up as an infrastructure facility for the scientific community. The primary aim is to collect and to disseminate the best possible data about educational processes and competence development. Data are disseminated no later than 18 months after the end of the field phase. To achieve this goal, incoming data are checked carefully for completeness and inconsistencies, undergo some routines of anonymization, and are edited and documented.

All data documentation is available via the NEPS website for data users (<https://www.neps-data.de/>). The information available is broken down by starting cohorts.

Data documentation contains, among other things, instruments, data manuals, code-books, as well as detailed information on sampling, weighting, data editing, and anonymization. In addition to a wealth of sophisticated written documentation, data users are supported by an online information system with tools for searching the NEPS instruments (NEPSplorer), an extensive user training program, as well as a telephone and an e-mail-hotline.

Data of all six cohorts have already been disseminated to the scientific community. There are three modes of data access: (1) download from the NEPS website, (2) remote access technology (RemoteNEPS), and (3) on-site access. Data available in these three modes differ in their level of anonymization. The data disseminated so far are being used by more than 800 researchers dealing with very different research topics. Whereas the majority of data users are still located in Germany, already a quickly increasing proportion of international data users has emerged. This is possible because all instruments and all documentation materials are also available in English.

8 Institutionalization as a Leibniz Institute

Several steps had to be taken to create a long-term perspective for the NEPS by integrating the panel study into the newly founded Leibniz Institute. In July 2011, the Bavarian State Minister of Sciences, Research, and the Arts (Bayerischer Staatsminister für Wissenschaft, Forschung und Kunst), Wolfgang Heubisch, submitted a request to the President of the Leibniz Association (Leibniz-Gemeinschaft), Karl Ulrich Mayer, to permanently institutionalize the NEPS under the umbrella of the Leibniz Association. After several further steps, the Joint Science Conference (Gemeinsame Wissenschaftskonferenz) then decided in April 2012 to promote the affiliation of NEPS as a Leibniz Institute. As part of this evaluation procedure a group from the German Council of Science and Humanities (Wissenschaftsrat) visited the NEPS Consortium in December 2012 in order to assess the work conducted so far. Based on a very positive evaluation report—labeling the NEPS a “unique and outstanding infrastructure facility” (Wissenschaftsrat, 2013, p. 63)—the Leibniz Association included the LIfBi as their new member as of January 2014. Following this, all necessary formal steps such as the formulation of rules and regulations and the entry in the local register of associations as well as building up a self-sufficient administration department were then successfully achieved. Also, the central committees—the Board of Trustees and the Scientific Advisory Board—were assigned and could meet for their first sessions. Finally, a cooperation agreement with the University of Bamberg was put in place and effective groups for structuring the further development of the NEPS were built up as part of the network structure guided by a mutually approved Network Charter.

9 Outlook

The NEPS has thus mastered its starting phase. Appropriate methods of collaboration have been developed and all six starting cohorts have successfully finished their first data-collection waves. Challenges of tracking panel participants and challenges in building up user-friendly longitudinal data products for researchers with different levels of methodological expertise have been faced and responded to. Due to the very positive panel progress, data collection in all cohorts will be continued over the coming years. First discussions have started to address the aspect of cohort succession. The NEPS team is strongly committed to our joint aims and objectives. They are willing and capable of solving the many different challenges associated with a dynamic multicohort sequence design.

References

- Allmendinger, J., Kleinert, C., Antoni, M., Christoph, B., Drasch, K., Janik, F., ... Ruland, M. (2011). Adult education and lifelong learning. In H.-P. Blossfeld, H.-G. Roßbach, & J. von Maurice (Eds.), *Zeitschrift für Erziehungswissenschaft, 14. Education as a lifelong process: The German National Educational Panel Study (NEPS)* (pp. 283–299). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Artelt, C., Weinert, S., & Carstensen, C. H. (Eds.). (2013). Assessing competencies across the lifespan within the German National Educational Panel Study (NEPS) [Special issue]. *Journal for Educational Research Online, 5*(2), 5–14. Retrieved from: <http://www.j-e-r-o.com/index.php/jero/issue/view/24>
- Aschinger, F., Epstein, H., Müller, S., Schaeper, H., Vöttiner, A., & Weiß, T. (2011). Higher education and the transition to work. In H.-P. Blossfeld, H.-G. Roßbach, & J. von Maurice (Eds.), *Zeitschrift für Erziehungswissenschaft, 14. Education as a lifelong process: The German National Educational Panel Study (NEPS)* (pp. 267–282). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Aßmann, C., Steinhauer, H. W., Kiesl, H., Koch, S., Schönberger, B., Müller-Kuller, A., ... Blossfeld, H.-P. (2011). Sampling designs of the National Educational Panel Study: Challenges and solutions. In H.-P. Blossfeld, H.-G. Roßbach, & J. von Maurice (Eds.), *Zeitschrift für Erziehungswissenschaft, 14. Education as a lifelong process: The German National Educational Panel Study (NEPS)* (pp. 51–65). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Bäumer, T., Preis, N., Roßbach, H.-G., Stecher, L., & Klieme, E. (2011). Education processes in life-course-specific learning environments. In H.-P. Blossfeld, H.-G. Roßbach, & J. von Maurice (Eds.), *Zeitschrift für Erziehungswissenschaft, 14. Education as a lifelong process: The German National Educational Panel Study (NEPS)* (pp. 87–101). Wiesbaden: VS Verlag für Sozialwissenschaften.

- Berendes, K., Fey, D., Linberg, T., Wenz, S. E., Roßbach, H.-G., Schneider, T., & Weinert, S. (2011). Kindergarten and elementary school. In H.-P. Blossfeld, H.-G. Roßbach, & J. von Maurice (Eds.), *Zeitschrift für Erziehungswissenschaft*, 14. *Education as a lifelong process: The German National Educational Panel Study (NEPS)* (pp. 203–216). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Blossfeld, H.-P., Roßbach, H.-G., & von Maurice, J. (Eds.). (2011). Education as a lifelong process: The German National Educational Panel Study (NEPS) [Special Issue]. *Zeitschrift für Erziehungswissenschaft*, 14. Wiesbaden: VS Verlag für Sozialwissenschaften.
- Blossfeld, H.-P., & von Maurice, J. (2011). Education as a lifelong process. In H.-P. Blossfeld, H.-G. Roßbach, & J. von Maurice (Eds.), *Zeitschrift für Erziehungswissenschaft*, 14. *Education as a lifelong process: The German National Educational Panel Study (NEPS)* (pp. 19–34). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Blossfeld, H.-P., & von Maurice, J. (2012). Chancen für die Forschung durch interdisziplinäre Netzwerkbildung: Das Beispiel des Nationalen Bildungspanels. *Gegenworte: Hefte für den Disput über Wissen*, 28, 39–43.
- Buchhaas-Birkholz, D. (2009). Die “empirische Wende” in der Bildungspolitik und in der Bildungsforschung. Zum Paradigmenwechsel des BMBF im Bereich der Forschungsförderung. *Erziehungswissenschaft*, 20(39), 27–33.
- Bundesministerium für Bildung und Forschung (2008). *Rahmenprogramm zur Förderung der empirischen Bildungsforschung* [Framework Programme for the Promotion of Educational Research] (Schriftenreihe Bildungsforschung 22). Bonn, Berlin: Bundesministerium für Bildung und Forschung.
- Elder, G. H. Jr., Johnson, M. K., & Crosnoe, R. (2003). The emergence and development of life course theory. In J. T. Mortimer, & M. J. Shanahan (Eds.), *Handbook of the life course* (pp. 3–19). New York: Springer US.
- Fickermann, D. (2004, November). *Überlegungen zu einem Bildungspanel aus Bundessicht*. Paper presented at the meeting of the German Youth Institute [Deutsches Jugendinstitut, DJI], München.
- Frahm, S., Goy, M., Kowalski, K., Sixt, M., Striethold, R., Blatt, I., ... Kanders, M. (2011). Transition and development from lower secondary to upper secondary school. In H.-P. Blossfeld, H.-G. Roßbach, & J. von Maurice (Eds.), *Zeitschrift für Erziehungswissenschaft*, 14. *Education as a lifelong process: The German National Educational Panel Study (NEPS)* (pp. 217–232). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Gross, C., Jobst, A., Jungbauer-Gans, M., & Schwarze, J. (2011). Educational returns over the life course. In H.-P. Blossfeld, H.-G. Roßbach, & J. von Maurice (Eds.), *Zeitschrift für Erziehungswissenschaft*, 14. *Education as a lifelong process: The German National Educational Panel Study (NEPS)* (pp. 139–153). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Kristen, C., Edele, A., Kalter, F., Kogan, I., Schulz, B., Stanat, P., & Will, G. (2011). The education of migrants and their children across the life course. In H.-P. Blossfeld, H.-G. Roßbach, & J. von Maurice (Eds.), *Zeitschrift für Erziehungswissenschaft*, 14.

- Education as a lifelong process: The German National Educational Panel Study (NEPS)* (pp. 121–137). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Ludwig-Mayerhofer, W., Solga, H., Leuze, K., Dombrowski, R., Künster, R., Ebralidze, E., ... Kühn, S. (2011). Vocational education and training and transitions into the labor market. In H.-P. Blossfeld, H.-G. Roßbach, & J. von Maurice (Eds.), *Zeitschrift für Erziehungswissenschaft, 14. Education as a lifelong process: The German National Educational Panel Study (NEPS)* (pp. 251–266). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Schlesiger, C., Lorenz, J., Weinert, S., Schneider, T., & Roßbach, H.-G. (2011). From birth to early child care. In H.-P. Blossfeld, H.-G. Roßbach, & J. von Maurice (Eds.), *Zeitschrift für Erziehungswissenschaft, 14. Education as a lifelong process: The German National Educational Panel Study (NEPS)* (pp. 187–202). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Stocké, V., Blossfeld, H.-P., Hoenig, K., & Sixt, M. (2011). Social inequality and educational decisions in the life course. In H.-P. Blossfeld, H.-G. Roßbach, & J. von Maurice (Eds.), *Zeitschrift für Erziehungswissenschaft, 14. Education as a lifelong process: The German National Educational Panel Study (NEPS)* (pp. 103–119). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Wagner, W., Kramer, J., Trautwein, U., Lüdtke, O., Nagy, G., Jonkmann, K., ... Schilling, J. (2011). Upper secondary education in academic school tracks and the transition from school to postsecondary education and the job market. In H.-P. Blossfeld, H.-G. Roßbach, & J. von Maurice (Eds.), *Zeitschrift für Erziehungswissenschaft, 14. Education as a lifelong process: The German National Educational Panel Study (NEPS)* (pp. 233–249). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Wohlkinger, F., Ditton, H., von Maurice, J., Haugwitz, M., & Blossfeld, H.-P. (2011). Motivational concepts and personality aspects across the life course. In H.-P. Blossfeld, H.-G. Roßbach, & J. von Maurice (Eds.), *Zeitschrift für Erziehungswissenschaft, 14. Education as a lifelong process: The German National Educational Panel Study (NEPS)* (pp. 155–168). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Weinert, S., Artelt, C., Prenzel, M., Senkbeil, M., Ehmke, T., & Carstensen, C.H. (2011). Development of competencies across the life span. In H.-P. Blossfeld, H.-G. Roßbach, & J. von Maurice (Eds.), *Zeitschrift für Erziehungswissenschaft, 14. Education as a lifelong process: The German National Educational Panel Study (NEPS)* (pp. 67–86). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Wissenschaftsrat (2013). *Stellungnahme zum Nationalen Bildungspanel (NEPS)*. Retrieved from: <http://www.wissenschaftsrat.de/download/archiv/2999-13.pdf>

About the authors

H.-P. Blossfeld
European University Institute (EUI), Florence, Italy.

J. von Maurice
Leibniz Institute for Educational Trajectories (LifBi), Bamberg, Germany.
e-mail: jutta.von-maurice@lifbi.de

H.-G. Roßbach
Leibniz Institute for Educational Trajectories (LifBi), Bamberg.
Chair of Early Childhood Education, University of Bamberg, Bamberg.

Analytic Strategies for the Study of Adaptation to Major Life Events: Making the Most of Large-Scale Longitudinal Surveys

Frank J. Infurna, Denis Gerstorf, Nilam Ram and Jutta Heckhausen

Abstract

Longitudinal surveys are essential for studying developmental change across the lifespan and have been instrumental in contributing to a better understanding of how people change from childhood through adolescence, adulthood, and into old age. This chapter reviews some of the strengths of longitudinal surveys for studying the adaptation and self-regulation of individuals who experience major life events across their adult lives. First, large national longitudinal surveys are highly instructive and necessary in order to prospectively collect data on sufficiently large sub-samples of people who are confronted with certain life events as the survey unfolds. Second, having access to prospective data from such sub-samples enables us to thoroughly track developmental changes in the nature, correlates, and outcomes of adaptation and self-regulation with the experience of major life events. Third, we discuss how multi-phase growth curve models can be used to distinguish between pre-event changes, reaction, and adaptation in order to examine individual differences in each of these phases and to explore individual and contextual variables that may serve as risk- or protective factors. Finally, we consider how embedded micro-longitudinal study designs and propensity score matching techniques may increase the advantages of panel surveys for studying adaptation and self-regulation across adulthood. In sum, combining the strengths of longitudinal surveys with contemporary methods of analysis can put researchers in a position to advance their knowledge of how life events shape developmental change trajectories across the entire lifespan.

1 Introduction

Longitudinal surveys are essential for examining how individuals change or remain stable over a given period of time (Baltes & Nesselroade, 1979). Longitudinal surveys are especially important for examining the extent to which major life events (e. g., disability, spousal loss, and unemployment) may or may not influence developmental trajectories of change across domains of functioning (Diener, Lucas, & Scollon, 2006; Hultsch & Plemons, 1979). Major life events have been shown to come with considerable changes in daily routines. For example, the incidence of pathology, spousal loss, and unemployment typically results in substantial declines in well-being (Fauth et al., 2012; Infurna et al., 2013; Lucas, 2007). However, the effects of these events are often not uniform. People differ in how they anticipate, deal with, and adjust to the events (Bonanno, 2004; Carver, 1998; Infurna & Luthar, in press). Tracking individuals as they go through such experiences enables researchers to make use of longitudinal surveys to examine such patterns of change and the multitude of different risk- and protective factors that contribute to heterogeneity.

Our focus in the present chapter is to highlight the utility of longitudinal surveys for examining developmental change and adaptation in relation to the experience of major life events. In conjunction with our aim, we focus on how the National Educational Panel Survey (NEPS) can be used to help answer research questions about the effects of major life events on psychological adjustment. The NEPS comprises a multiple cohort (i. e., newborn, preschool, various school-age cohorts, college students, and a wide age range of adults) large-scale (approx. total sample of 100,000) longitudinal study of Germans who are assessed on an annual basis using an extensive battery consisting of competence-related, economic, sociological, psychological, and health information. The NEPS thus provides the opportunity to study developmental change and adaptation to life events in each of these cohorts before, during the time of, and after these life events happen. For example, researchers can begin to examine employment outcomes for individuals transitioning from college to the workforce. Furthermore, researchers are in a position to examine the long-term sequelae of major life events that may occur in childhood and adolescence and how these sequelae impact later developmental outcomes in adulthood. We have organized the chapter into four sections. First, we discuss why longitudinal surveys are needed to gain access to large samples to study subgroups of the population who experience major life events such as disability, spousal loss, and unemployment. Second, we discuss how longitudinal surveys allow for the thorough tracking of developmental changes before, at the time of, and after major life events, as well as of correlates and the consequences of such events. Third, we highlight how we can capitalize on the flexibility of multi-phase models of change to better understand the different processes underlying the anticipation, reaction, and adaptation to an event. Our fourth and final section foreshadows how incorporating micro-longitudinal study designs within longitudinal surveys can enable the further understanding of the mechanisms involved in

the adaptation to major life events and how advances in contemporary methodology, such as propensity score matching procedures, can be used as a methodological tool to advance our understanding of change in relation to major life events.

2 Longitudinal Surveys and Sample Size

Major life events can be broadly defined as internal or external occurrences that signify a qualitative shift or role transformation in one's life (Frederick & Loewenstein, 1999; Diener et al., 1999; Hultsch & Plemons, 1979). For example, a more controllable role transformation would be getting married, experiencing childbirth, or starting a career, whereas a less controllable role transformation would be suffering from a threatening health event or becoming unemployed. Experiencing a major life event can result in a wide range of responses or changes across a variety of domains of functioning. Well-being is one of the most studied domains for examining change in relation to major life events, and its pattern of change typically consists of multiple phases: reaction and adaptation. The reaction phase refers to changes in the time surrounding the life event (which could be months or years). For example, individuals typically experience a substantial drop in well-being with spousal loss (Lucas et al., 2003), whereas positive life events, such as marriage or childbirth, are associated with an increase or boost in well-being (Diener et al., 2006; Lucas, 2007). The phase following the reaction to a major life event is called adaptation. In the context of major life events, adaptation broadly refers to whether or not an individual returns to his or her previous level of functioning after he or she has experienced the event (Frederick & Loewenstein, 1999). For example, unemployment typically results in sustained lower levels of well-being as compared with the years prior to unemployment (Lucas, 2007). Furthermore, the initial decrease in well-being during the time surrounding spousal loss (reaction) is typically followed by the return of well-being levels to previous levels after several years (Lucas et al., 2003). We note that our description of well-being change in relation to major life events mainly focuses on the model-implied (average) pattern of change of reaction and adaptation. However, there are large between-person differences in reaction and adaptation such that individuals may follow different pathways of change in relation to major life events. For example, Bonnano (2004) explains that individuals may follow four different trajectories (i. e., resilient, chronic, delayed, or recovered), with most individuals being resilient and not experiencing any (lasting) changes in functioning associated with the major life event (for discussion, see Infurna & Luthar, in press).

We next assert that longitudinal surveys are an essential tool for studying developmental change and adaptation to major life events across domains of functioning. We use spousal loss as an example throughout this chapter to illustrate this point. Our concentration on spousal loss is due to its status as one of the most stressful and detrimental events that could occur in someone's life (Holmes & Rahe, 1967). This focus

also provides the opportunity to discuss in more detail how longitudinal surveys can be used to study developmental change and adaptation to major life events.

The ultimate goal when studying major life events is to examine how they impact functioning in the time leading up to, surrounding, and following event occurrence. There are several advantages of longitudinal surveys for studying developmental change and adaptation in relation to major life events. First, interdisciplinary longitudinal surveys assess relatively large samples of participants repeatedly, which allows researchers to identify segments of the population that have experienced a major life event. For example, the incidence rate of widowhood for men and women across the entire lifespan in the United States is 3.5 and 7.8 per 1,000 individuals, respectively (Elliott & Simmons, 2011; Lee, 2002; Spraggins, 2003). Second, longitudinal surveys repeatedly assess participants at a regular interval, which enables the examination of how participants develop and change prior to, surrounding, and following a major life event. This examination is critical because unlike experimental conditions in which there are typically two groups, namely control and experimental, researchers cannot require participants in a study to experience an event such as spousal loss. Therefore, longitudinal surveys provide the opportunity and flexibility to study “natural experiments” by identifying these events that naturally occur in the life course and isolating the various components of change that may occur. Third, examining developmental change processes in relation to major life events permits targeting the “stressful” times in which individuals’ reactive and regulatory systems are in action, that is the times during which individual differences in how these systems function will stand out (Gerstorf & Ram, 2012; Hultsch & Plemons, 1979). As such, natural events provide unique opportunities to study the mechanisms underlying successful development (Rutter, 2007). For example, losing a spouse is a devastating event that can lead to dramatic changes in one’s well-being and health. Research on this transition can shed light on factors that contribute to adjustment, recovery, and even growth. For example, supportive social relationships, one’s ability to fulfill personal and social responsibilities, and the capacity for positive emotions and generative experiences are typically associated with resilience when confronted with major life events (Bonanno et al., 2002, 2004; Frederickson et al., 2003). Therefore, it is of utmost importance to be in a position to study not only average change, but also what some of the risk- and protective factors that moderate these changes are.

3 Prospective Tracking of Developmental Change

Longitudinal surveys allow for the identification of individuals who have experienced specific major life events. Once individuals who have experienced the major life event of interest have been identified, we can then examine how particular domains of functioning change in relation to event occurrence. The yearly assessments as implemented in surveys like the NEPS enable the capturing of anticipatory and re-

sponsive changes to life events as they unfold and allow data availability on the date of the event to have information on the amount of time that has elapsed (for discussion, see Uglanova & Staudinger, 2012). For example, empirical evidence suggests that well-being is relatively stable across adulthood and old age (Charles et al., 2001). Research in the past decade has shown that well-being change in adulthood and old age may be driven by processes beyond that of chronological age, such as major life events (Diener et al., 2006; Lucas, 2007).

More specifically, aligning individuals in relation to a major life event allows researchers to examine the nature of change and the consequences of such events. When examining change in relation to major life events, we are interested in examining change in the time leading up to, surrounding, and following the experience of spousal loss, as well as long-term outcomes thereof (e.g., mortality, incidence of disease). The repeated assessments can help researchers distinguish the defined components of change. For spousal loss, we are interested in defining and distinguishing between *anticipation*, *reaction*, and *adaptation*. The time leading up to spousal loss can be represented by an *anticipatory period* characterized by stability or declines in well-being. Changes (e.g., declines in well-being) during the anticipation phase can be considered an active process that may help individuals cope with the impending loss of their loved one or, in contrast, be indicative of a loss of resources and an inability for emotional regulation (Heckhausen, Wrosch, & Schulz, 2010; Kastenbaum & Costa, 1977). The *reaction period* refers to one's changes in well-being at the time surrounding spousal loss. Are individuals able to maintain their levels of functioning despite the devastating experience of spousal loss, or does this loss result in a precipitous drop (Uglanova & Staudinger, 2012)? The time following spousal loss is referred to as the *adaptation period*. This phase examines whether individuals are able to return back to levels of functioning that are similar to those several years prior to spousal loss (Lucas, 2007). Lastly, we can target *long-term outcomes* of the major life event, such as mortality following spousal loss. Several studies have shown that spousal loss is predictive of physical health declines and mortality (Elwert & Christakis, 2008; Mendes de Leon, Kasl, & Jacobs, 1993; Schulz & Beach, 1999; Stroebe, Schut, & Stroebe, 2007). The continuous tracking of participants in longitudinal surveys enables researchers to examine the long-term consequences of poor adaptation to a major life event. For example, sorrow after the loss of a loved one may not be associated with mortality (reaction), but failure to return to a normal emotional life after a certain period of time (adaptation) may be detrimental and increase one's mortality hazard.

Not all individuals exhibit the same pattern of well-being change with spousal loss, and in fact, there are large between-person differences in how individuals react and adapt to life-altering events (Carver, 1998; Wortman & Silver, 1989). For example, Bonanno (2004) suggests that most individuals are resilient and able to adapt by recovering relatively quickly or even maintaining their pre-loss well-being, whereas other individuals experience steep loss-related declines in well-being and are only

able to adapt slowly (for discussion, see Infurna & Luthar, in press). Reasons for heterogeneity in trajectories of change following major life events include situational and individual factors (Bonanno, 2004; Carver, 1998; Hultsch & Plemons, 1979), which may have differing roles depending on the phase. For example, older age and greater health problems of the spouse may result in stronger well-being declines in the years preceding spousal loss (anticipation) because spousal loss may be considered an expected event with anticipatory declines being instrumental for adaptation in the following years (Jopp & Smith, 2006; Schulz et al., 2003). During the time surrounding spousal loss (reaction), social network integration and supportive relationships may serve to protect against the negative impact of the stress of losing a spouse because people have a larger pool of individuals to go to, which may help with coping and protect against well-being declines (Bonanno, 2004; Cohen & Wills, 1985). Following spousal loss, educational attainment may lead to better adaptation through the knowledge and use of adaptive and compensatory strategies (Adler et al., 1994).

There is much to be gained from using longitudinal surveys to examine developmental change and adaptation in relation to major life events. First, researchers are able to compare and contrast the magnitude of effects major life events have on particular domains of functioning. Up to this point, most of the research has focused on well-being change in relation to major life events. However, whether the pattern of change is similar across psychological factors, such as goal (dis)engagement strategies as well as cognition and health, remains an open question. For example, does spousal loss only result in substantial declines in well-being and not in cognitive functioning? Compared across major life events, could events centered around pathology (e.g., disability) have a greater impact beyond the well-being domain and influence cognition and health in contrast to events centered on work or family that may only impact the well-being domain? Future research will be able to disentangle such propositions by examining whether the eventual onset of the life event drives the change and whether the levels and rates of change in the years preceding have implications for the eventual onset of such events. Second, researchers can pinpoint the time in relation to the major life event that is most stressful for the individual and which areas of functioning are at their limits. This has intervention implications for helping to maintain one's levels of functioning in times of great disruption (Rae et al., 2010). For example, interventions that focus on positive activities, such as cultivating one's strengths, visualizing an ideal future self, and performing kind acts, are shown to boost one's well-being (for discussion, see Lyubomirsky & Layous, 2013). Lastly, it is important to examine not only how levels of functioning differ following a major life event as compared with prior, but also whether the rate of change is affected. We have found that depressive symptoms show shallower increases in the years following cancer diagnoses as compared with the years leading up to cancer diagnosis (Infurna et al., 2013). The developmental rate of change leading up to a major life event, such as a cancer diagnosis, may be indicative of an eventual underlying pathology that will lead to an increased risk for pathology incidence. Not only can one's absolute levels

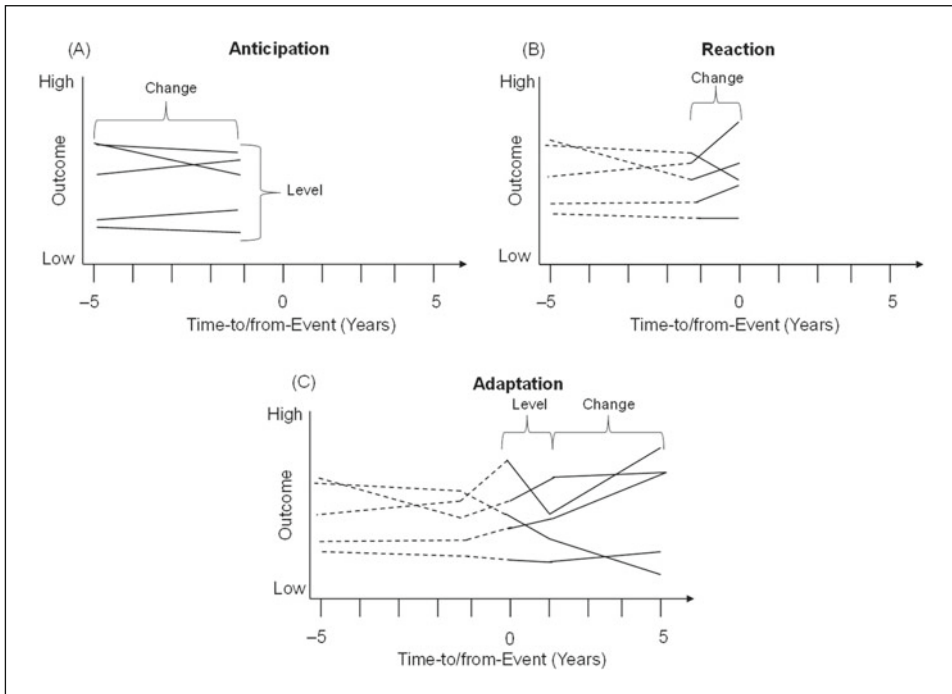
of functioning be altered by a particular life event, but the event could subsequently put an individual on a positive or negative trajectory of change. Change in the developmental rate following a life event, such as spousal loss or cancer diagnosis, could be indicative of individuals' seeking out support or using ambulatory care that results in less steep declines in domains of functioning. Future research bears the burden of examining such notions further and discovering what the implications of an altered rate of change for outcomes following the particular life event are.

4 Multi-phase Growth Models

In the previous sections, we detailed how longitudinal surveys are instrumental for studying developmental change in relation to major life events. These research studies often need large-enough sample sizes and the specified measures of interest to examine the nature and correlates of change in relation to major life events. This next section focuses on how analytical models, particularly multi-phase growth models (McArdle & Nesselroade, 2003; Ram & Grimm, 2007; Singer & Willett, 2003), can be used to answer our research questions and hypotheses. As we have discussed, when individuals are confronted with major life events, one would expect that different phases can be distinguished and that different sorts of risk- and protective factors play a role in these phases. In the case of examining developmental change in relation to spousal loss, phases to be distinguished include anticipation (i. e., time prior to spousal loss), reaction (i. e., time surrounding spousal loss), and adaptation (i. e., time following spousal loss). For example, low perceived control may protect against well-being declines with spousal loss because it indicates an acknowledgement that spousal loss is due to factors beyond one's own control; conversely, high perceived control following spousal loss may lead to better adaptation through individuals' engagement in the proper coping and goal-engagement strategies (Skinner, 1995). Using advanced methods such as multi-phase growth models, we can indeed model whether (different constellations of) perceptions of control moderate changes in well-being in relation to major life events using a large-sample and frequent-assessment dataset such as the NEPS and thereby arrive at a better understanding of the phenomena we are interested in.

Multi-phase growth curve analysis comprises a flexible set of models that allow researchers to isolate particular components along a time series when examining change in relation to major life events (McArdle & Nesselroade, 2003; Ram & Grimm, 2007; Singer & Willett, 2003). Figure 1 graphically illustrates how such a model can be used to distinguish the components involved in how the outcome of interest changes in relation to the major life events of anticipation, reaction, and adaptation. Part A of Figure 1 displays how levels and rates of change in the outcome may vary several years prior to event incidence (anticipation). Anticipation can be broadly assessed as change in the years leading up to the major life event. In the case of widowhood, an-

Figure 1 Illustrating the components or phases of developmental change in relation to major life events. These three components include anticipation (A), reaction (B), and adaptation (C). Anticipation refers to individuals' levels and rates of change in functioning prior to the major life event (A). The reaction phase refers to how individuals may display differential rates of change with the incidence of the major life event (B). Lastly, differential levels and rates of change may be exhibited in the years following the major life event, which is referred to as adaptation (C). Each line in Figure 1 displays a hypothetical trajectory of change for individuals who experience spousal loss.



Anticipatory changes in well-being may begin up to several years prior to spousal loss, possibly due to the worsening health of the dying spouse. Part B of Figure 1 graphically illustrates the reaction phase and how change may take different forms with life event incidence as well as that this may differ across individuals. The reaction phase is typically quantified as the time surrounding the life event (i. e., during the year when the event occurred). For methodological reasons, this is typically defined as the difference in well-being or another outcome between the wave immediately prior to the event and the wave when the event was first observed or reported. This explains why reaction appears to refer to something that happens before the event (i. e., between -1 and 0). However, identifying the date or month of the event permits moving towards having more nuanced approaches for studying reactions to major life events through

examining change via monthly intervals (see Uglanova & Staudinger, 2012). Lastly, Part C of Figure 1 illustrates how individuals show differential level and change in the years following event incidence. Adaptation for some individuals may be immediate (one year) or take several years for others. Adaptation may take different forms: (a) whether individuals' levels of functioning in the years following the life event will return back to prior levels and (b) how individuals' rates of change following the event may or may not be similar to those in the years leading up to the life event. For example, anticipatory declines in well-being leading up to spousal loss may result in individuals' being able to better adapt and show stronger increases in well-being in the years following widowhood.

The components of the multi-phase growth model shown in Figure 1 can be used to answer research questions regarding developmental change and adaptation in relation to major life events. As a first step, we can model the average trajectory of change in relation to the event of interest. Furthermore, by estimating variance in each of the growth components, we can determine whether there are between-person differences. Second, researchers may be interested in examining whether between-person difference factors, such as socio-demographic, cognition, and physical health factors, moderate such associations. The lines in Figure 1 represent trajectories for hypothetical participants and, in particular, that there can be a great deal of heterogeneity in how individuals anticipate, react to, and adapt to life-altering events (Carver, 1998; Infurna & Luthar, in press; Wortman & Silver, 1989). This is indeed the case with spousal loss such that not all individuals exhibit the same pattern of well-being change in relation to spousal loss. The task would be to examine whether various risk- and protective factors, such as social support or coping strategies, buffer against declines in the time surrounding the major life event and better adaptation in the time that follows. This would be done, for example, by inserting social support into the model as a moderator of well-being change during the anticipation and adaptation phases.

5 Implications for the National Educational Panel Study (NEPS)

The NEPS offers various opportunities for tracking developmental change and adaptation in relation to major life events. First and foremost, the design of the NEPS allows for addressing research questions centered on major life events from the initiation of the study. Beginning with the second wave, researchers can use the NEPS to examine change following major life events, such as spousal loss or the incidence of disease, through annual observations across domains of functioning. The NEPS can be used, for example, to examine whether goal engagement or disengagement strategies are best for optimizing well-being following spousal loss. Furthermore, do goal (dis)engagement strategies display similar associations on developmental outcomes in the context of major life events at different phases of the lifespan? Second, the

NEPS surveys participants from the entire lifespan, that is infancy through old age, which opens up the opportunity to study the impact of major life events that are more likely to occur in specific areas of the lifespan and compare their effects depending on one's own point in the lifespan. For example, researchers using the NEPS will be in a position to compare and contrast the effects of spousal loss for a period of the lifespan when it would be atypical (i. e., young adulthood and midlife) to typical (i. e., old age). Spousal loss in young adulthood and midlife could be associated with more substantial drops in well-being due to its being considered an "off-time" event as compared with old age, at which point it is considered an "on-time" event (Neurgarten & Hagestad, 1976). Thinking more broadly beyond just spousal loss, the NEPS can help examine whether the timing of major life events plays a role in shaping developmental change across the lifespan. Thus, the lifespan nature of the NEPS puts researchers in the unique position of studying major life events from across the entire lifespan and investigating their implications for developmental change and adaptation, such as the transition from school to the work force, unemployment, retirement, marital transitions, and the onset of disease.

Another advantage of the lifespan sample of the NEPS is the ability to assess whether (or not) life events have cumulative effects across the entire lifespan, effectively allowing researchers to move more towards a prospective approach. For example, empirical evidence suggests that psychological stress in childhood is associated with an increased susceptibility to chronic disease in old age (Miller, Chen, & Parker, 2011). The longitudinal design of the NEPS allows for more specifically examining how early life events, such as psychological stress in childhood, transpire over time to affect development in adulthood through possible psychosocial and biological mechanisms that may underlie these associations. For example, child maltreatment may be linked to adult mental and physical health problems via emotion processing and risky health behavior (Infurna, Rivers, Reich, & Zautra, 2015; Miller et al., 2011; Repetti, Taylor, & Seeman, 2002).

Third, previous research has largely centered on well-being change in relation to spousal loss and more generally to major life events. The extensive assessment battery of the NEPS allows researchers to take a multivariate approach by examining how other components may or may not be affected by the major life event and also allows them to target mediators and moderators of change in prominent areas, such as well-being. For example, how are motivational processes of primary and secondary control strategies, such as goal engagement, affected by spousal loss (e. g., Heckhausen et al., 2010)? It could be expected that spousal loss would result in an initial decline in goal engagement strategies and an increase in goal re-engagement strategies as individuals turn their focus to more attainable goals. This, especially, could be the case when the surviving spouse may have been involved in caregiving-related activities. Examining change in psychosocial constructs with major life events can lead to mediation analyses aimed at their role in accounting for well-being change. For example, declines in well-being with spousal loss could be due to or accounted for by the loss

of emotional support from one's network or a change in goal engagement strategies. Lastly, the extensive psychosocial battery can be used to examine various risk- and protective factors that moderate change or adjustment with major life events. For example, do perceptions of control and social support provide an additive or multiplicative effect for increasing one's likelihood for adaptation following widowhood?

6 Future Directions

This final section discusses future directions that can be used to more thoroughly examine the extent to which domains of functioning change in relation to major life events. In particular, we discuss propensity score matching procedures as a statistical method of analysis to further our understanding of how major life events influence developmental trajectories of change. We also concentrate on how the incorporation of micro-longitudinal designs (e.g., measurement-burst designs) within the context of macro-longitudinal studies of change can complement and allow for taking a more process-oriented approach to studying the underlying mechanisms and pathways.

Propensity score matching is a class of methods in which the objective is to create a case-matched "control" group to compare with the "treatment" group (Rubin, 1974). This technique is a way to move towards making potentially causal inferences with observational data and has largely been used in prevention and intervention research. Moreover, it has recently been incorporated in psychological research (Foster, 2010; Rutter, 2007; Stuart, 2010). The relevance for major life events would be the creation of a "control" group to compare with participants who have experienced a particular life event in order to examine whether there are differences in the levels and rates of change in the outcome of interest. The objective would be to move towards determining whether a particular life event may "cause" developmental changes in particular areas of functioning.

This procedure would consist of two steps. First, researchers would need to identify covariates, or factors by which to identify participants to include in the control group. Covariates would need to be selected based on how likely they would be to be associated with the treatment condition or major life event. For example, socio-demographic and behavioral factors are typically associated with disease incidence; therefore, these factors would be essential to include as covariates to ensure that the two groups would be similar on these factors prior to conducting further analyses. The selected covariates would then be used to estimate a propensity score using logistic regression to indicate the likelihood of an individual's being assigned to the treatment condition (i.e., major life event; Stuart, 2010). In the second step, participants who had experienced the major life event would then be matched to participants who had not experienced the major life event based on the propensity score, which would represent the predicted likelihood of being assigned to the treatment or major life event group. Once a "control" group had been determined, the next step would be to

conduct analyses to examine whether there were differences in the levels and rates of change in the outcome of interest between the two groups. For example, had individuals who had experienced spousal loss already exhibited lower levels of and steep drops in well-being in the years leading up to spousal loss? Focusing on the transition from adolescence to young adulthood, Jackson and colleagues (2012) utilized propensity score matching to create two groups of participants in Germany who did or did not experience military training. In comparing these two groups, they found that military training resulted in lower levels of agreeableness. Further, highly informative applications of propensity score matching techniques are readily available in the literature (e.g., Gerstorf et al., 2015).

Macro-longitudinal studies allow for examining developmental change over years or decades. For example, multiple longitudinal surveys have shown that well-being remains relatively stable across the adult lifespan, even into older ages (Charles et al., 2001; Mroczek & Spiro, 2005). These designs allow researchers to gain insight into the long-term course of development and, as we have discussed in this chapter, developmental change and adaptation in relation to major life events. However, longitudinal surveys are limited in their ability to discern the underlying mechanisms driving change. To obtain the necessary data, longitudinal studies may look to embed micro-longitudinal or measurement-burst designs within the macro-longitudinal design (for discussion, see Nesselrode, 1991; Ram & Gerstorf, 2009). At the micro-time scale, researchers obtain multiple reports or assessments over a relatively short span of time (e.g., hours, days) via a diary, ecological momentary assessment, or ambulatory procedures (Bolger et al., 2003; Hoppmann & Riediger, 2009; Sliwinski, 2008). This enables the examination of individuals in the daily context and the procurement of reports of stressors, emotions, behaviors, and physiological indicators that can be linked to longitudinal change. Furthermore, measurement-burst designs can help distinguish among intra-individual change and variability that may occur at different time scales (for discussion, see Sliwinski, 2008). When combined with data from longitudinal studies assessing change over years or decades, this can shed light on mechanisms of developmental change (Gerstorf, Hoppmann, & Ram, 2014). For example, Ram and colleagues (2011) found that cognitive aging over approximately 13 years of time was associated with greater cognitive plasticity, less cardiovascular lability, and less emotional diversity over a two-week period in older adults. Embedding this sort of design in longitudinal studies more regularly can provide the opportunity to examine daily functioning both prior to and following major life events. In the specific case of spousal loss, research has been able to study risk- and protective factors associated with well-being change following event occurrence. For example, Ong and colleagues (2005) observed that reporting more daily control was linked to less daily anxiety and buffered against the impact of stressors on well-being in a sample of recently bereaved persons.

7 Conclusion

Longitudinal surveys are essential for studying and examining developmental change across the lifespan. In this chapter, we have discussed the advantages of longitudinal surveys for examining developmental change and adaptation in relation to major life events. Our discussion additionally focused on how major life events can be studied in the NEPS. The NEPS offers many fruitful avenues to examine how major life events may or may not shape developmental change across the lifespan. First, large-scale longitudinal surveys are essential tools for capturing sufficiently large sub-samples of individuals who are confronted with certain life events as the study unfolds. Second, prospective data from longitudinal surveys allows researchers to prospectively assess developmental change and adaptation in relation to major life events. Third, multi-phase growth models can be used to distinguish between the components of level and rate of change with the experience of major life events. Fourth, future research could examine further components of developmental change and adaptation via the utilization of micro-longitudinal designs and propensity score matching methods.

References

- Adler, N. E., Boyce, T., Chesney, M. A., Cohen, S., Folkman, S., Kahn, R. L., & Syme, S. L. (1994). Socioeconomic status and health: The challenge of the gradient. *American Psychologist*, *49*(1), 15–24.
- Baltes, P. B., & Nesselroade, J. R. (1979). History and rationale of longitudinal research. In J. R. Nesselroade, & P. B. Baltes (Eds.), *Longitudinal research in the study of behavior and development* (pp. 1–39). New York, NY: Academic Press.
- Bisconti, T. L., Bergeman, C. S., & Boker, S. M. (2006). Social support as a predictor of variability: An examination of the adjustment trajectories of recent widows. *Psychology and Aging*, *21*(3), 590–599.
- Bolger, N., Davis, A., & Rafaeli, E. (2003). Diary methods: Capturing life as it is lived. *Annual Review of Psychology*, *54*(1), 579–616.
- Bonanno, G. A. (2004). Loss, trauma, and human resilience: Have we underestimated the human capacity to thrive after extremely adverse events? *American Psychologist*, *59*(1), 20–28.
- Bonanno, G. A., Wortman, C. B., Lehman, D. R., Trweed, R. G., Haring, M., Sonnega, J., ... Nesse, R. M. (2002). Resilience to loss and chronic grief: A prospective study from preloss to 18-months postloss. *Journal of Personality and Social Psychology*, *83*(5), 1150–1164.
- Bonanno, G. A., Papa, A., LaLande, K., Westphal, M., & Coifman, K. (2004). The importance of being flexible: The ability to both enhance and suppress emotional expression predicts long-term adjustment. *Psychological Science*, *15*(7), 482–287.

- Carver, C. S. (1998). Resilience and thriving: Issues, models, and linkages. *Journal of Social Issues, 54*(2), 245–266.
- Charles, S. T., Reynolds, C. A., & Gatz, M. (2001). Age-related differences and change in positive and negative affect over 23 years. *Journal of Personality and Social Psychology, 80*(1), 136–151.
- Cohen, S., & Wills, T. A. (1985). Stress, social support, and the buffering hypothesis. *Psychological Bulletin, 98*(2), 310–357.
- Diener, E., Lucas, R. E., & Scollon, C. N. (2006). Beyond the hedonic treadmill: Revising the adaptation theory of well-being. *American Psychologist, 61*(4), 305–314.
- Elliott, D. B., & Simmons, T. (2011). *Marital events of Americans: 2009* (American Community Survey Reports No. 13). Washington, DC: U. S. Census Bureau.
- Elwert, F., & Christakis, N. A. (2008). The effect of widowhood on mortality by the causes of death of both spouses. *American Journal of Public Health, 98*(11), 2092–2098.
- Fauth, E. B., Gerstorf, D., Ram, N., & Malmberg, B. (2012). Changes in depressive symptoms in the context of disablement processes: The role of demographic characteristics and social support. *The Journals of Gerontology: Series B: Psychological Sciences and Social Sciences 67*(2), 167–177.
- Foster, E. M. (2010). Causal inference and developmental psychology. *Developmental Psychology, 46*(6), 1454–1480.
- Frederick, S., & Loewenstein, G. (1999). Hedonic adaptation. In D. Kahneman, E. Diener, & N. Schwarz (Eds.), *Well-being: The foundations of hedonic psychology* (pp. 302–329). New York, NY: Sage.
- Gerstorf, D., Hoppmann, C. A., & Ram, N. (2014). The promise and challenges of integrating multiple time-scales in adult developmental inquiry. *Research in Human Development, 11*, 75–90. doi: 10.1080/15427609.2014.906725
- Gerstorf, D., Hülür, G., Drewelies, J., Eibich, P., Düzel, S., Demuth, I., ... Lindenberger, U. (2015). Secular changes in late-life cognition and well-being: Towards a long bright future with a short brisk ending? *Psychology and Aging, 30*, 301–310. doi: 10.1037/pag0000016
- Gerstorf, D., & Ram, N. (2012). Late-life: A venue for studying the mechanisms by which contextual factors influence individual development. In S. K. Whitbourne, & M. J. Sliwinski (Eds.), *Handbook of Adulthood and Aging* (pp. 49–71). New York, NY: Wiley.
- Heckhausen, J., Wrosch, C., & Schulz, R. (2010). A motivational theory of life-span development. *Psychological Review, 117*(1), 32–60.
- Holmes, T. H., & Rahe, R. H. (1967). The social readjustment rating scale. *Journal of Psychosomatic Research, 11*(2), 213–218.
- Hoppmann, C. A., & Riediger, M. (2009). Ambulatory assessment in lifespan psychology: An overview of current status and new trends. *European Psychologist, 14*(2), 98–108.
- Hultsch, D. F., & Plemons, J. K. (1979). Life events and life-span development. In P. B. Baltes, & O. G. Brim, Jr. (Eds.), *Life-span development and behavior* (Vol. 2, pp. 1–36). New York, NY: Academic Press.

- Infurna, F. J., Gerstorff, D., & Ram, N. (2013). The nature and correlates of change in depressive symptoms with cancer diagnosis: Reaction and adaptation. *Psychology and Aging, 28*(2), 386–401.
- Infurna, F. J., & Luthar, S. S. (in press). Resilience to major life stressors is not as common as thought. *Perspectives on Psychological Science*.
- Infurna, F. J., Rivers, C. T., Reich, J., & Zautra, A. J. (2015). Childhood trauma and personal mastery: Their influence on emotional reactivity to everyday events in a community sample of middle-aged adults. *PLoS ONE, 10*(4), 1–21.
- Jackson, J. J., Thoemmes, F., Jonkmann, K., Lüdtko, O., & Trautwein, U. (2012). Military training and personality trait development: Does the military make the man, or does the man make the military? *Psychological Science, 23*(3), 270–277.
- Jopp, D., & Smith, J. (2006). Resources and life-management strategies as determinants of successful aging: On the protective effect of selection, optimization, and compensation. *Psychology and Aging, 21*(2), 253–265.
- Lee, G. R. (2002). Widowhood. *Encyclopedia of Aging*. Retrieved from <http://www.encyclopedia.com/doc/1G2-3402200429>
- Lucas, R. E. (2007). Adaptation and the set-point model of subjective well-being: Does happiness change after major life events? *Current Directions in Psychological Science, 16*(2), 75–79.
- Lucas, R. E., Clark, A. E., Georgellis, Y., & Diener, E. (2003). Reexamining adaptation and the set point model of happiness: Reactions to changes in marital status. *Journal of Personality and Social Psychology, 84*(3), 527–539.
- Lyubomirsky, S., & Layous, K. (2013). How do simple positive activities increase well-being? *Current Directions in Psychological Sciences, 22*(1), 57–62.
- McArdle, J. J., & Nesselroade, J. R. (2003). Growth curve analysis in contemporary psychological research. In J. Shinka, & W. Velicer (Eds.), *Comprehensive handbook of psychology: Research methods in psychology* (Vol. 2, pp. 447–480). New York, NY: Wiley.
- Mendes de Leon, C. F., Kasl, S. V., & Jacobs, S. (1993). Widowhood and mortality risk in a community sample of the elderly: A prospective study. *Journal of Clinical Epidemiology, 46*(6), 519–527.
- Miller, G. E., Chen, E., & Parker, K. J. (2011). Psychological stress in childhood and susceptibility to the chronic diseases of aging: Moving toward a model of behavioral and biological mechanisms. *Psychological Bulletin, 137*(6), 959–997.
- Mroczek, D. K., & Spiro, A. III. (2005). Change in life satisfaction during adulthood: Findings from the Veterans Affairs Normative Aging Study. *Journal of Personality and Social Psychology, 88*(1), 189–202.
- Nesselroade, J. R. (1991). The warp and woof of the developmental fabric. In R. Downs, L. Liben, & D. Palermo (Eds.), *Visions of development, the environment, and aesthetics: The legacy of Joachim F. Wohlwill* (pp. 213–240). Hillsdale, NJ: Lawrence Erlbaum.
- Neugarten, B. L., & Hagestad, G. O. (1976). Age and the life course. In R. E. Binstock, & E. Shanas (Eds.), *Handbook of aging and social sciences* (pp. 35–61). New York, NY: Van Nostrand Reinhold.

- Ong, A. D., Bergeman, C. S., & Bisconti, T. L. (2005). Unique effects of daily perceived control on anxiety symptomatology during conjugal bereavement. *Personality and Individual Differences, 38*(5), 1057–1067.
- Rae, M. J., Butler, R. N., Campisi, J., de Grey, A. D. N. J., Finch, C. E., Gough, M., ... Logan, B. J. (2010). The demographic and biomedical case for late-life interventions in aging. *Science Translational Medicine, 2*(40), 40cm21.
- Ram, N., & Gerstorf, D. (2009). Time-structured and net intraindividual variability: Tools for examining the aging of dynamic characteristics and processes. *Psychology and Aging, 24*(4), 778–791.
- Ram, N., Gerstorf, D., Lindenberger, U., & Smith, J. (2011). Developmental change and intraindividual variability: Relating cognitive aging to cognitive plasticity, cardiovascular lability, and emotional diversity. *Psychology and Aging, 26*(2), 363–371.
- Ram, N., & Grimm, K. (2007). Using simple and complex growth models to articulate developmental change: Matching theory to method. *International Journal of Behavioral Development, 31*(4), 303–316.
- Repetti, R. L., Taylor, S. E., & Seeman, T. E. (2002). Risk families: Family social environments and the mental and physical health of offspring. *Psychological Bulletin, 128*(2), 330–366.
- Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology, 66*(5), 688–701.
- Rutter, M. (2007). Proceeding from observed correlation to causal inference: The use of natural experiments. *Perspectives on Psychological Science, 2*(4), 377–395.
- Schulz, R., & Beach, S. R. (1999). Caregiving as a risk factor for mortality: The Caregiver Health Effects Study. *Journal of the American Medical Association, 282*(23), 2215–2219.
- Schulz, R., Mendelsohn, A. B., Haley, W. E., Mahoney, D., Allen, R. S., Zhang, S., ... Belle, S. H. (2003). End-of-life care and the effects of bereavement on family caregivers of persons with dementia. *New England Journal of Medicine, 349*(20), 1936–1942.
- Skinner, E. A. (1995). *Perceived control, motivation, and coping*. Thousand Oaks, CA: Sage.
- Sliwinski, M. (2008). Measurement-burst designs for social health research. *Social and Personality Psychology Compass, 2*(1), 245–261.
- Stroebe, M., Schut, H., & Stroebe, W. (2007). Health outcomes of bereavement. *Lancet, 370*(9603), 1960–1973.
- Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science, 25*(1), 1–21.
- Uglanova, E. A., & Staudinger, U. M. (2012). Zooming in on life events: Is hedonic adaptation sensitive to the temporal distance from the event? *Social Indicators Research, 111*(1), 265–286.
- Wortman, C. B., & Silver, R. C. (1989). The myths of coping with loss. *Journal of Consulting and Clinical Psychology, 57*(3), 349–35.

Author Notes

Frank J. Infurna, Department of Psychology, Arizona State University; Denis Gerstorf, Institute for Psychology at the Humboldt University, Berlin, Germany, German Socio-Economic Panel Study, German Institute for Economic Research, Berlin, Germany and Department of Human Development and Family Studies at the Pennsylvania State University, PA, USA; Nilam Ram, Department of Human Development and Family Studies at the Pennsylvania State University, German Socio-Economic Panel Study, German Institute for Economic Research, Berlin, Germany, and Max Planck Institute for Human Development, Berlin, Germany; Jutta Heckhausen, Department of Psychology and Social Behavior, University of California, Irvine.

Denis Gerstorf and Nilam Ram gratefully acknowledge the support provided by the National Institute on Aging (NIA) RC1-AG035645; NIA R21-AG032379; and NIA R21-AG033109; the DIW Berlin (German Institute for Economic Research); and the Social Science Research Institute at Pennsylvania State University. The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies.

e-mail: frank.infurna@asu.edu; denis.gerstorf@hu-berlin.de; nilam.ram@psu.edu; heckhaus@uci.edu.

About the authors

F. J. Infurna
Arizona State University, Arizona, USA.
e-mail: Frank.Infurna@asu.edu

D. Gerstorf
German Institute for Economic Research (DIW Berlin), Berlin, Germany.
Humboldt University, Berlin, Germany.
The Pennsylvania State University, Pennsylvania, USA.

N. Ram
German Institute for Economic Research (DIW Berlin), Berlin, Germany.
The Pennsylvania State University, Pennsylvania, USA.
Max Planck Institute for Human Development, Berlin, Germany.

J. Heckhausen
University of California, Irvine, USA.

II. Sampling, Recruiting, and Fieldwork Management

Weighting Panel Cohorts in Institutional Contexts

Hans Walter Steinhauer and Sabine Zinn and Christian Aßmann

Abstract

The National Educational Panel Study (NEPS) surveys and tests, next to adults, undergraduates, and newborns, Kindergarten children and students within their institutional contexts. Individuals who decided to participate in the panel study can refuse participation in specific waves or drop out completely. Weighting adjustments are usually applied to account for nonparticipation. Within the institutional cohorts of the NEPS, these adjustments take clustering at the institutional level into account. In NEPS, information on children is enriched by interviews with their parents. Thus, dealing with two distinct but possibly interdependent participation decisions has to be regarded by a joint modeling approach. The results of models analyzing the participation propensity provide insights concerning factors influencing the participation probability. In general, few potential determinants affect participation decisions. These include place of residence, language spoken at home, age, and having missing values in personal or migration-related characteristics. For later waves the participation status of the previous wave has proved to be a good predictor. Moreover, being surveyed and tested within the institutional context positively influences participation decisions.

1 Introduction

Longitudinal studies aim to survey the same individuals over time. In the beginning, an initial sample is drawn. This initial sample reduces in size for different reasons, yielding the set of individuals finally surveyed. Lepkowski and Couper (2002) assign this loss of individuals to different nonresponse processes. First, unit nonresponse is caused by unwillingness to participate in the panel study. Second, among those willing to participate in the panel study there are further processes leading to

unit nonresponse reducing the sample size over time. These include failure to trace persons from one wave to another as well as not being able to contact persons, and finally, refusal to further participate in future waves of a panel study. Because not all of these nonresponse processes reducing the sample size occur at random, there is potential for selection bias. This potential bias can be encountered by weighting adjustments. Weighting adjustments accounting for unit nonresponse are referred to as sample weighting adjustments (Kalton & Kasprzyk, 1986). In panel studies these are applied first, to correct for nonparticipation within the initial sample, and second, for wave-specific unit nonresponse within the panel cohort. Moreover, weights can be adjusted in a way that weighted estimates and distributions confirm with known population parameters and distributions, where this adjustment is referred to as population weighting adjustment (Kalton & Kasprzyk, 1986). Methods used in population weighting adjustments aim to correct for potential bias due to incomplete coverage or noncoverage of the population and sampling error (Brick, 2013). Both, sample and population weighting adjustments—although reducing bias—usually result in an increased variability of weights; thereby lowering the precision of survey estimates (Kalton & Flores-Cervantes, 2003; Valliant, 2004).

Within the National Educational Panel Study (NEPS), among other things, samples of children were drawn for starting cohorts focusing on children in Kindergarten institutions (SC2), on students in Grade 5 (SC3), and on students in Grade 9 also referred to as SC4 (Blossfeld, Roßbach, & von Maurice, 2011). For the initial sample of these three cohorts, panel consent was asked for in advance of the first wave survey. Children willing to participate and providing valid consent forms constitute the panel cohorts of SC2, SC3, and SC4. Detailed information on sample weighting adjustments correcting for the unwillingness to participate in the panel study among the initial sample is given in Steinhauer, Aßmann, Zinn, Goßmann, and Rässler (2015).

The sample weighting adjustments for unit nonresponse among the panel cohorts of SC2, SC3, and SC4 for Wave 1 and Wave 2 are the focus of this chapter. In advance of the survey a parent has to give permission for the child or the student (if not of legal age) to take part in SC2, SC3, or SC4. Together with the permission for their child, the parent is asked to participate him- or herself as well. After initial panel consent has been given, each member of the panel cohorts can either participate in future waves or not. In sum, there are three possible participation statuses, namely: participant, temporary dropout, and final dropout. Children and students taking part in the survey or the test are considered as participants. Children and students explicitly refusing participation in the current and all following waves or their parents withdrawing panel consent are considered as final dropouts.¹ Children and students who, for whatever reason, do not show up at the day of testing and surveying are considered as temporary drops.

1 Besides that, the NEPS basically considers children and students as final dropouts if no information, from whatever source, is available on the children or the students for a period longer than two years.

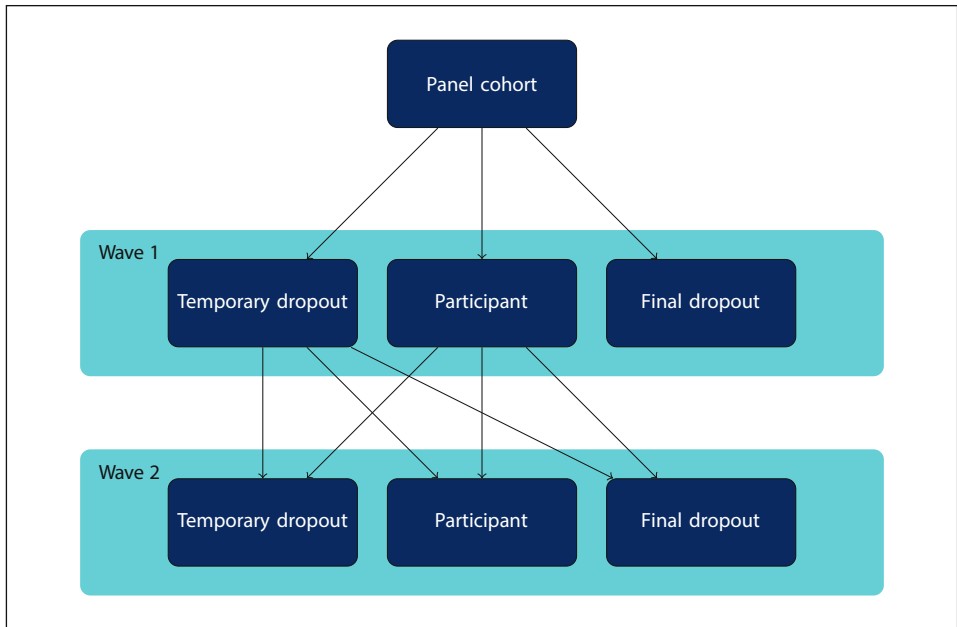
Figure 1 Participation patterns for panel cohort members

Figure 1 shows the possible pathways of panel cohort members through the first two waves. Children and students in the panel cohort can participate in Wave 1 or not. Those that decide to participate form the group of participants in Wave 1. Those that do not participate can be distinguished further into two groups. First, children and students refusing further participation completely, or parents withdrawing their child's panel consent make up the group of final dropouts. Lastly, the group of temporary dropouts consists of children and students not participating in Wave 1 but generally willing to participate in future waves. All those children and students who have not been defined as final dropouts in Wave 1 form the panel cohort for Wave 2. Again, each panel cohort member can participate in the second wave or not. Thus, participants and temporary dropouts from Wave 1 can either remain as such or change their status. Final dropouts are not contacted again in successive waves. Figure 1 also illustrates that the number of combinations of different participation statuses increases rapidly with each wave.

As already mentioned, failure to trace or not being able to contact persons drives up unit nonresponse within the panel cohort. Typically children and students surveyed in an institutional context drop out very rarely, see Table 1. Hence, here nonresponse is not such an issue as it is when children and students have left their institutional context. In the NEPS, these children and students are tracked individually. Students and children also end up in individual tracking if an institution refuses to

further cooperate with the study. Children and students who are individually tracked are surveyed by sending the test instruments and questionnaires to their homes. Clearly, such a process leads to lower response rates than the corresponding surveys conducted in the institutional context.

Moreover, surveying and testing of Kindergarten children in SC2 and students of SC3 and SC4 is accompanied by collecting additional information from other persons. That is, NEPS adopts a multi-informant perspective. Persons additionally surveyed include educators, teachers, institution heads, and one parent. In advance of the Wave 1 survey parents had to provide consent to the participation of their children. At the same time they were asked if they themselves would like to participate in the survey. Parents willing to participate in the panel study together with their child provide information on the family and social background. This information is collected in a computer-assisted telephone interview (CATI). The number of individuals and their parents participating together is usually smaller than the number of participating individuals. This is because not all parents are willing to take part in the survey. The decision process described in Figure 1 applies to all parents who are willing to partake in the panel, too.

2 Data

The numbers corresponding to the different participation patterns of panel cohort members illustrated in Figure 1 are displayed in Table 1. The table gives the numbers of participants, temporary dropouts, and final dropouts for SC2, SC3, and SC4 categorized by the participation status in Wave 1 and Wave 2. The majority of panel cohort members participates in both waves and only a small percentage drops out in one of the two waves. In SC2, 91 % of the children participate in both waves.² Similarly, in SC3 and SC4, 90 % and 93 % of the students participate in both waves.³ The proportion of temporary dropouts is generally very small over all of the three cohorts and is below 1 %. So far, final dropouts have only occurred in Wave 2 of SC3. All together, these figures indicate persistent panel cohorts. To account for unit non-response in the panel cohorts, wave-specific weights for children and students are provided corresponding to the groups of participants displayed in Table 1. We provide two kinds of weights. Cross-sectional weights for individuals participating in a specific wave and longitudinal weights for individuals participating in each wave. For example, for SC3, we provide cross-sectional weights for the 5,774 students participating in Wave 1 (w_{t1}) and for the 5,790 students participating in Wave 2 (w_{t2}).

2 In SC2, Wave 1 was conducted between January and October 2011 and Wave 2 between January and May 2012.

3 In SC3, Wave 1 was conducted between November 2010 and January 2011 and Wave 2 in the same months one year later. Wave 1 of SC4 was conducted between November 2010 and January 2011 and Wave 2 between May and July 2011.

Table 1 Participation status of individuals by starting cohort and wave

Wave 1	Wave 2			Total
	Participant	Temporary dropout	Final dropout	
SC2—Kindergarten children				
Participant	2,739	232	0	2,971
Temporary dropout	24	1	0	25
Final dropout	0	0	0	0
Total	2,763	233	0	2,996
SC3—Grade 5 students				
Participant	5,473	287	14	5,774
Temporary dropout	317	21	0	338
Final dropout	0	0	0	0
Total	5,790	308	14	6,112
SC4—Grade 9 students				
Participant	15,308	321	0	15,629
Temporary dropout	709	87	0	796
Final dropout	0	0	0	0
Total	16,017	408	0	16,425

Note: The data in the table is based on the Scientific Use File versions DOI:10.5157/NEPS:SC2:2.0.0, DOI:10.5157/NEPS:SC3:2.0.0, and DOI:10.5157/NEPS:SC4:1.1.0.

Longitudinal weights are provided for the 5,473 students participating in Wave 1 as well as in Wave 2 of SC3 (*w_t12*).

Because surveying and testing children and students is accompanied by a CATI with one parent there is an additional participation decision; the participation decision of that parent. To account for this additional decision, we provide weights for the group of children or students jointly participating with a parent. For SC2 and SC3, Wave 1 and Wave 2, there is a separate interview with the parents in each wave.⁴ In contrast, there is only one interview with a parent in SC4.⁵ This CATI is carried out between Wave 1 and Wave 2, because the two waves were conducted within one year.

4 The CATI in Wave 1 of SC2 was conducted during April and December 2011. Wave 2 interviews were conducted during February and May 2012. In SC3, the CATI in Wave 1 was conducted between January and July 2011 and in Wave 2 between February and May 2012.

5 This CATI was carried out during January and July 2011.

Thus, this single decision process of a parent to participate is used to provide weights for Wave 1 and Wave 2 for students and parents jointly participating in SC4.

Table 2 gives the numbers for the different joint participation statuses of children and students as well as of their parents. In SC2, most of the parents participate together with their children—in sum 78 % in Wave 1. In comparison, the number of parents and students participating together in SC3 is lower: 65 % of the parents participated together with their children in Wave 1 and 61 % in Wave 2. Compared to this, 54 % of the students and parents in SC4 participated together in Wave 1 and 55 % in Wave 2.⁶

Wave-specific weights provided for individuals and parents are based on the different groups displayed in Table 2. For couples of children and parents, we again provide two sets of weights—cross-sectional and longitudinal weights. For example, we provide cross-sectional weights for the 3,974 students and parents of SC3 participating jointly in Wave 1 (*w_tp1*), and for the 3,727 couples of students and parents jointly participating in Wave 2 (*w_tp2*). Longitudinal weights are provided for the 3,417 students and parents participating jointly in Wave 1 and Wave 2 of SC3 (*w_tp12*).

In order to reduce bias and not to inflate the variance of the estimates of a weighted analysis too much, variables in nonresponse adjustments should be related to the participation propensity as well as to the variables that are of interest for the subject studied (Little & Vartivarian, 2005). It is often hard to meet both criteria at the same time. This is mainly because of two reasons. First, many surveys are multipurpose surveys, which makes it hard to cover all possible variables of interest. Second, the set of variables available for participants and nonparticipants is usually sparse (Kreuter & Olson, 2011). An ongoing panel study has the advantage of generating new information for the panel cohort with additional waves. Hence, we can address the second problem (at least partly) by basing our models on the most current information available. Such processing may lead to different values of variables used for nonresponse adjustments and thus to different values of the same weight for different versions of Scientific Use Files.⁷

To model the participation propensity of children and students (and their parents), we use variables that are available throughout all cohorts and waves. These include gender (*male* and *female*), age group⁸ (*younger half* and *older half* of the cohort), as well as language spoken at home (*German* and *Non-German*). For SC2, we further consider the children's place of residence (*with both parents* and *with one parent or others*). For SC3 and SC4 further variables include migration background (*Turkish*

6 The reported numbers of final dropouts among parents correspond throughout all three starting cohorts to those parents who refused (further) panel participation prior to the survey and to those ones who refused (further) participation during an interview.

7 For example, the values of the weight *w_t1* might slightly differ between SUF version 1.0.0 and 2.0.0 because new information became available and was used to update *w_t1* after Wave 2.

8 The age of an individual is computed using month and year of birth. The cohort sample is then split into a younger and an older half according to the median age of the entire cohort sample.

Table 2 Joint participation statuses of individuals and parents by starting cohort and wave

Students	Parents			
	Participant	Temporary dropout	Final dropout	Total
	SC2—Wave 1			
Participant	2,322	448	201	2,971
Temporary dropout	18	4	3	25
Final dropout	0	0	0	0
Total	2,340	452	204	2,996
	SC3—Wave 1			
Participant	3,974	462	1,338	5,774
Temporary dropout	177	28	133	338
Final dropout	0	0	0	0
Total	4,151	490	1,471	6,112
	SC3—Wave 2			
Participant	3,727	636	1,427	5,790
Temporary dropout	92	104	112	308
Final dropout	1	2	11	4
Total	3,820	742	1,550	6,112
	SC4—Wave 1			
Participant	8,813	1,448	5,368	15,629
Temporary dropout	360	70	366	796
Final dropout	0	0	0	0
Total	9,173	1,518	5,734	16,425
	SC4—Wave 2			
Participant	9,010	1,443	5,564	16,017
Temporary dropout	163	75	170	408
Final dropout	0	0	0	0
Total	9,173	1,518	5,734	16,425

Note: The data in the table is based on the Scientific Use File versions DOI:10.5157/NEPS:SC2:1.0.0, DOI:10.5157/NEPS:SC3:2.0.0, and DOI:10.5157/NEPS:SC4:1.1.0.

Table 3 Sampling strata in SC4

Stratum	School type
h = 1	Gymnasien
h = 2	Hauptschulen
h = 3	Realschulen
h = 4	Integrierte Gesamtschulen, Freie Waldorfschulen
h = 5	Schulen mit mehreren Bildungsgängen
h = 6	Förderschulen

and *Former Soviet Union*), nationality (*German* or *other*), as well as the sampling stratum of the school. In SC4, there are in total six sampling strata as displayed in Table 3.

The school sample of SC3 is made up of three explicit strata. Because schools from SC4 referring to strata $h = 1, \dots, 5$ also provide education to Grade 5 students, these schools were pooled to the first stratum in sampling schools for SC3. The second stratum of SC3 consists of schools providing schooling to Grade 5 students, but not to Grade 9 students, referring mainly to *Grundschulen* and *schulartunabhängigen Orientierungsstufen*. The third stratum of SC3 includes those schools from the stratum $h = 6$ of SC4 that also educate students in Grade 5. SC3 additionally includes a supplement of 214 cases with a migration background related to Turkey or the Former Soviet Union. For more detailed information on the sampling design, see Steinhauer et al. (2015). Besides the variables already listed, for nonresponse adjustments we additionally consider missing indicators for migration characteristics (language spoken at home, nationality, and migration background) and personal characteristics (gender, month and year of birth). For nonresponse adjustments in Wave 2, we also include the participation status of Wave 1 (*participated* or *dropout*). Besides that, we determine whether the individual still is in the institutional context or is in individual tracking (*individual tracking* or *in the institutional context*). Because not all parents participate in the CATI there is little information from the call record available on participating and nonparticipating parents. For joint decision modeling we consider the number of call attempts to the first contact in the CATI as an indicator for the likelihood of being at home as indicated by Durrant and Steele (2009). Besides that, when modeling parents' participation decisions, we also use children's characteristics. Finally, when modeling parents' participation decisions in Wave 2, we include the parents' participation status from Wave 1; analogous to modeling students' participation decisions in Wave 2.

3 Methods

Probit regressions are used to model the binary participation status (participant vs. dropout). The three possible statuses—participant, temporary and final dropout—at first glance suggest the use of multinomial probit models. However, although SC3 covers all three statuses (see Table 1), the small number of final dropouts does not allow for using multinomial probit models. Thus, we model the participation probability of Kindergarten children and students using univariate binary probit models with a random intercept at the institutional level. Likewise, for the same reason, the joint participation decisions of children and their parents are modeled using a bivariate binary probit model. For more details on the model frameworks given below see Greene (2012). All models have been estimated using the software environment for statistical computing R (R Development Core Team, 2015). The univariate probit with random intercept is estimated using the function `glmer()` from the `lme4` package (Bates, Maechler, & Bolker, 2012). The bivariate binary probit model is estimated using the `zelig()` function with a bivariate binary probit link provided by the `ZeligChoice` package (Owen, Imai, Lau, & King, 2012).

3.1 Univariate Binary Probit Framework

The univariate probit model for $i = 1, \dots, n$ individuals with dichotomous participation status y_i is given by

$$y_i = \begin{cases} 1 & \text{if } \tilde{y}_i > 0, \\ 0 & \text{else} \end{cases} \quad \text{with } \tilde{y}_i = X_i\beta + \xi_i, \quad (1)$$

where \tilde{y}_i denotes a latent variable, X_i the regressors, β the coefficients of the model and $\xi_i \sim N(0, \sigma)$ denotes the disturbance, with $\sigma = 1$. The accordant random intercept probit model is defined by

$$y_{ij} = \begin{cases} 1 & \text{if } \tilde{y}_{ij} > 0, \\ 0 & \text{else} \end{cases} \quad \text{with } \tilde{y}_{ij} = X_{ij}\beta + \underbrace{\alpha_j + \varepsilon_j}_{\xi_j}, \quad (2)$$

where $\alpha_j \sim N(0, \omega^2)$ denotes the random intercept and $\varepsilon_j \sim N(0, \sigma)$ is the disturbance, again with $\sigma = 1$. This extension allows to take clustering on a higher level into account. Individuals (denoted by i) are clustered in groups $j = 1, \dots, m$ of size n_j . The model given in Equation (2) is used to estimate participation probabilities for children and students clustered in their institutions.

3.2 Bivariate Binary Probit Framework

The univariate probit model given in Equation (1) can further be extended to allow for modeling two (possibly) correlated participation decisions. Let i denote an individual, k his or her parent, and ρ the correlation parameter. Then the bivariate binary probit can be written as

$$\begin{aligned} y_i &= \begin{cases} 1 & \text{if } \tilde{y}_i > 0, \\ 0 & \text{else} \end{cases} \quad \text{with } \tilde{y}_i = X_i\beta + \xi_i \\ y_k &= \begin{cases} 1 & \text{if } \tilde{y}_k > 0, \\ 0 & \text{else} \end{cases} \quad \text{with } \tilde{y}_k = X_k\gamma + \xi_k \end{aligned} \quad (3)$$

with $k = 1, \dots, l$ and $l = n$. We assume that $(\xi_i, \xi_k) \sim N(0, \Sigma)$, with

$$\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \quad (4)$$

Here, the parameter ρ measures the conditional correlation between the participation decisions of individuals and their parents. That is, in sum, the bivariate probit model consists of the regression coefficients β and γ , the correlation coefficient ρ , the dependant variables y_i and y_k , the latent variables \tilde{y}_i and \tilde{y}_k , and the design matrices X_i and X_k .

3.3 Derivation of Adjusted Weights

For Y_{it} being the participation status of individual i in Wave t and X_{it} a set of available information, the univariate models described above can be used for estimating the probability $P(Y_{it} = 1|X_{it})$ of an individual i participating in Wave t conditional on available information and participation statuses of previous waves. For the case of a panel survey with two waves the probability to participate in Wave 2 can be written as

$$\begin{aligned} P(Y_{i2} = 1|X_{i2}, X_{i1}) &= P(Y_{i2} = 1|X_{i2}, Y_{i1} = 1) \cdot P(Y_{i1} = 1|X_{i1}) + \\ &P(Y_{i2} = 1|X_{i2}, Y_{i1} = 0) \cdot P(Y_{i1} = 0|X_{i1}). \end{aligned}$$

In accordance therewith, the probability to participate in both waves, that is Wave 1 ($Y_{i1} = 1$) and Wave 2 ($Y_{i2} = 1$), is given by

$$P(Y_{i2} = 1, Y_{i1} = 1|X_{i2}, X_{i1}) = P(Y_{i2} = 1|X_{i2}, Y_{i1} = 1) \cdot P(Y_{i1} = 1|X_{i1}).$$

With Y_{kt} being the participation status of a parent k in Wave t and X_{kt} a set of available information, the bivariate model given above can be used for estimating the proba-

bility for a child and a parent to participate jointly $P(Y_{it}=1, Y_{kt}=1|X_{it}, X_{kt})$ in Wave t conditional on the available information and participation statuses of previous waves.

$$\begin{aligned} P\left(\begin{matrix} Y_{i2} = 1|X_{i2}, X_{i1} \\ Y_{k2} = 1|X_{k2}, X_{k1} \end{matrix}\right) &= P\left(\begin{matrix} Y_{i2} = 1|X_{i2}, Y_{i1} = 1 \\ Y_{k2} = 1|X_{k2}, Y_{k1} = 1 \end{matrix}\right) \cdot P\left(\begin{matrix} Y_{i1} = 1|X_{i1} \\ Y_{k1} = 1|X_{k1} \end{matrix}\right) + \\ &P\left(\begin{matrix} Y_{i2} = 1|X_{i2}, Y_{i1} = 0 \\ Y_{k2} = 1|X_{k2}, Y_{k1} = 1 \end{matrix}\right) \cdot P\left(\begin{matrix} Y_{i1} = 0|X_{i1} \\ Y_{k1} = 1|X_{k1} \end{matrix}\right) + \\ &P\left(\begin{matrix} Y_{i2} = 1|X_{i2}, Y_{i1} = 1 \\ Y_{k2} = 1|X_{k2}, Y_{k1} = 0 \end{matrix}\right) \cdot P\left(\begin{matrix} Y_{i1} = 1|X_{i1} \\ Y_{k1} = 0|X_{k1} \end{matrix}\right) + \\ &P\left(\begin{matrix} Y_{i2} = 1|X_{i2}, Y_{i1} = 0 \\ Y_{k2} = 1|X_{k2}, Y_{k1} = 0 \end{matrix}\right) \cdot P\left(\begin{matrix} Y_{i1} = 0|X_{i1} \\ Y_{k1} = 0|X_{k1} \end{matrix}\right). \end{aligned}$$

Thus, the probability for a child or student and a parent to jointly participate in both waves, that is in Wave 1 ($Y_{i1} = 1, Y_{k1} = 1$) and in Wave 2 ($Y_{i2} = 1, Y_{k2} = 1$), is given by

$$P\left(\begin{matrix} Y_{i2} = 1, Y_{i1} = 1|X_{i2}, X_{i1} \\ Y_{k2} = 1, Y_{k1} = 1|X_{k2}, X_{k1} \end{matrix}\right) = P\left(\begin{matrix} Y_{i2} = 1|X_{i2}, Y_{i1} = 1 \\ Y_{k2} = 1|X_{k2}, Y_{k1} = 1 \end{matrix}\right) \cdot P\left(\begin{matrix} Y_{i1} = 1|X_{i1} \\ Y_{k1} = 1|X_{k1} \end{matrix}\right).$$

The inverse of these probabilities form the adjustment factors for the sample weighting adjustment. In detail, given the panel entry weight w_i for individual i the accordant nonresponse adjusted weight can be computed as

$$\omega_i(t = T) = w_i \cdot P(Y_{iT} = 1|X_{iT})^{-1}$$

for cross-sectional weights and as

$$\omega_i(t = T, \dots, 1) = w_i \cdot P(Y_{iT} = 1, Y_{iT-1} = 1, \dots, Y_{i1} = 1|X_{iT}, X_{iT-1}, \dots, X_{i1})^{-1}$$

for longitudinal weights.

4 Results

4.1 Starting Cohort 2—Kindergarten

In Wave 1 there are only 25 temporary dropouts among the children of SC2. These cases are too few for an accordant binary regression model. Thus, they are adjusted for by an unconditional modeling, that is, the related adjustment factor is $2,996 \div (2,996 - 25)$. Note that here we deviate from the approach presented in Subsection 3.3. In order to compute adjustment factors for the joint participation of children and their parents in Wave 1, we multiply the adjustment factor compensating for nonresponse among parents. The latter has been estimated by means of a random intercept bi-

nary probit model. The data used for this purpose have been imputed to cope with item nonresponse. In detail, we used hot deck imputation (Andridge & Little, 2010) to deal with missing values in the variable ‘language spoken at home’ (in total, eight cases) and in the variable ‘place of residence’ (one missing case). The coefficients of the model characterizing parents’ participation propensity are given in Table 4. We see that parents who have a child living with only one parent or with others have a significantly lower participation propensity. In contrast, parents who have children that predominantly speak German at home have a higher propensity to participate.

In Wave 2, there are 233 children that have temporarily dropped out from the sample and in the longitudinal sample of Wave 1 and 2 there are 257 children classified as temporary dropouts. To analyze the longitudinal participation propensity, the dependent variable in the model is operationalized as a dichotomous variable indicating whether a child participated in both waves or not. The corresponding models estimating the participation propensities for Kindergarten children in the different waves (see Table 5) comprise as explanatory variables age and gender of the children, their place of residence, and the language spoken at home. The two-level structure of children within Kindergartens is considered by specifying a random intercept at the Kindergarten level. The age of the child is the only characteristic showing a significant effect on the participation propensity of a child in Wave 2. In contrast, the propensity of children to participate in both waves, that is, Wave 1 and Wave 2, is additionally significantly influenced by whether the child lives with both parents or not.

Table 4 shows the results of models estimating the joint participation propensities of children and parents for Wave 2 (cross-sectional sample) as well as for Wave 1 and Wave 2 (longitudinal sample). For the group of children and parents participating together in the cross-sectional sample, that is in Wave 2, children’s propensity to jointly participate is negatively influenced by being part of the older age group as well as by living with one parent or others. The propensity is positively influenced by German as the language predominantly being spoken at home. For parents the propensity to jointly participate is negatively influenced by having a child living with one parent or others and positively influenced by German being the language spoken at home. Furthermore, we find a significant residual correlation between the participation decisions of children and parents—though it is not strong. For the longitudinal sample of children and parents jointly participating the effects remain stable and change only slightly in magnitude.⁹

9 For this model the parents’ participation status is operationalized analogous to the longitudinal participation of children. That is, the dependent variable is dichotomous and distinguishes between parents participating in both waves (Wave 1 and Wave 2) or not.

4.2 Starting Cohort 3—Grade 5

The models estimating the participation propensity of Grade 5 students in Wave 1 and Wave 2 are displayed in Table 6. The bivariate binary probit models for estimating the joint participation decision of students and parents are given in Table 7. The weights for Grade 5 students are adjusted as described in Subsection 3.3.

For Wave 1 of SC3 the random intercept probit model displayed in Table 6 shows negative significant effects for students being educated in schools sampled in the SC4-strata $h = 4$ (Integrierte Gesamtschulen and Freie Waldorfschulen) and $h = 5$ (Schulen mit mehreren Bildungsgängen).¹⁰ Also, having missing values in personal characteristics (age group and gender) significantly lowers the participation propensity. In contrast, speaking German as a native language influences the participation propensity positively.

The bivariate binary probit model estimating the joint participation propensity for students and parents in SC3 is given in Table 7. We find that parents whose children are educated in schools of SC4-strata $h = 1$ (Gymnasien), $h = 3$ (Realschulen), $h = 4$ (Integrierte Gesamtschulen and Freie Waldorfschulen), and in schools offering education to Grade 5 but not to Grade 9 students (mainly Grundschulen and schulartunabhängige Orientierungsstufen) have a higher participation propensity. Parents that can be contacted for the CATI by less than four phone calls also have a higher participation propensity to take part together with their children. Likewise, parents who have a child with a Turkish migration background and children speaking German at home have a higher propensity to participate than their counterparts. The effect of speaking German at home also positively influences the students' participation propensity. In contrast, the participation propensity of students is lowered if they are educated in schools belonging to SC4-strata $h = 4$ (Integrierte Gesamtschulen and Freie Waldorfschulen) and $h = 5$ (Schulen mit mehreren Bildungsgängen) as well as by having missing values in personal characteristics (age group and gender). Although very weakly positive, we find significant residual correlation in the joint participation decisions of students and parents.

In Wave 2 of SC3, there are 14 final dropouts, see Table 1. To account for the difference between temporary and final dropout, these cases have been accounted for by an unconditional model, that is, their unconditional participation probability is $14 \div 6,112$. For the remaining 6,098 students a random intercept model has been computed. The results are given in Table 6. Compared to the results of the nonresponse models corresponding to Wave 1, the negative effect of having missing values in personal characteristics remains stable. Also, the positive effect of German as a native language remains positive and significant but reduces in magnitude. However, the effects vanish of being educated in schools of the SC4-strata $h = 4$ (Integrierte Gesamtschulen

10 We use the stratification variables of SC4 in SC3, too, because they provide deeper insights than the pooled stratification variable of SC3.

and Freie Waldorfschulen) and $h = 5$ (Schulen mit mehreren Bildungsgängen); while being educated in schools of SC4-stratum $h = 6$ (Förderschulen) lowers participation propensities in Wave 2 significantly. Students who are individually tracked show a very low participation propensity. Furthermore, compared to Wave 1, the standard deviation of the random intercept increases.

Table 7 shows the results of the model for the joint participation decision of students and parents. Here, the 14 children who have dropped out permanently are excluded together with their parents. Parents participation propensities are mainly influenced by the same characteristics as in Wave 1, changing only slightly in magnitude. The effects vanish of having a child with a Turkish migration background or the child speaking German at home. The couple's own participation status (the students' and the parents') in Wave 1 is found to be a strong predictor for Wave 2 participation. For students the decision to jointly participate in the survey of Wave 2 is significantly lower when being educated in Förderschulen referring to SC4-stratum $h = 6$. Having missing values in personal characteristics as well as being in the field of individual tracking further lowers participation propensities of students significantly. What is interesting to note is that, the students' participation status in Wave 1 is not a significant predictor for their participation in Wave 2. In contrast, the participation decision of a student's parents is positively influencing the student's own decision in Wave 2.

4.3 Starting Cohort 4—Grade 9

The models estimating the participation propensity for Grade 9 students in Wave 1 and Wave 2 are displayed in Table 6. The weights for Grade 5 students are adjusted as shown in Subsection 3.3. To derive weighting adjustment for the group of students and parents jointly participating, we deviate from the approach stated in Subsection 3.3, because analysis has not shown any significant correlation. Thus, the propensity of parents participating in the CATI between Waves 1 and 2 is estimated separately. The joint participation propensity of both students and parents is then obtained by multiplying the corresponding estimated probabilities.

The participation propensity of students in Wave 1 of SC4 is (significantly) negatively influenced by being educated in schools referring to the strata $h = 6$ (Förderschulen) and $h = 2$ (Hauptschulen) and by being German, as well as by having missing values in personal characteristics (age group and gender) or migration characteristics (native language or nationality).¹¹ In addition, speaking German as a native language positively influences the participation decision, see Table 6.

11 This is due to the fact that the reference category is being educated in schools of stratum $h = 1$, that is, Gymnasien.

The parents' decision to participate is affected negatively by having children in a school referring to any of the strata relevant for Grade 9 students.¹²

For Wave 2 the effect of being educated in schools of stratum $h = 6$ (Förderschulen) increases (compared to Wave 1) in magnitude, see Table 6. In contrast, the significant effect of being educated in schools of stratum $h = 2$ (Hauptschulen) vanishes. Being part of the younger half of the age group positively influences the participation decision in Wave 2 as well as speaking German as a native language. Regarding the missing indicators, the effects reduce in magnitude and the estimate for missing values in personal characteristics is not significant anymore. We find that the students' participation status of Wave 1 is a strong predictor for the participation propensity in Wave 2. The standard deviation of the random intercept increases from Wave 1 to Wave 2.

5 Conclusion

This chapter has given insights into the derivation of wave-specific nonresponse adjustments within the institutional cohorts of SC2, SC3, and SC4 of the NEPS. In the NEPS we distinguish between three participation statuses, namely: participant, temporary dropout, and final dropout. Up to Wave 2, the number of persons permanently dropping out from the sample is small. Therefore, we mainly differentiate between participation and temporary dropout in the nonresponse adjustments of weights and adjust for final dropout by the inverse of their percentage of the cohort. We use probit regressions to compute participation probabilities of individuals. Their inverse constitutes the adjustment factors of the related contexts. In particular, we use univariate probit models to describe the participation decision of children and students and bivariate probit models to map the joint participation decision of children or students and their parents. The latter allows for modeling possibly correlated decisions. To account for clustering, that is, children and students being nested within institutions, we use random intercept models. If correlation between children or students and their parents' participation decision turns out to be negligible, the bivariate model setting is replaced by a univariate one, that is, children or students and their parents' decisions are modeled separately. When modeling the participation status in Wave 2, we generally condition on the Wave 1 participation status. In this way, cross-sectional as well as longitudinal weights can be provided together in a straightforward way.

The results of analyzing the participation statuses of Kindergarten children shows that German as the predominantly spoken language together with place of residence influences participation decisions significantly. The results for students in Grade 5 and Grade 9 show that speaking German as a native language, having missing values

12 This is due to the fact that the reference category is being educated in schools of stratum $h = 1$, that is, Gymnasien.

in personal and migration characteristics, as well as being individually tracked are the factors influencing participation decisions. Besides that, SC4-strata-specific effects are found. However, these are not stable over time.

The students of SC4 will leave school soon, that is, they will enter either the vocational track or head toward a career in higher education. In other words, they will leave the institutional contexts of schools and will have to be tracked individually. Clearly, later in time, this will also occur to the students of SC3. Generally, the German education system allows students to enter a large variety of educational pathways. Hence, for nonresponse adjustments in future waves the current approach of describing participation propensities will have to be extended accordingly. Besides that, in future waves, we expect a higher number of final dropouts that should be explicitly included into the modeling process. To this end, a multinomial model framework might be used, for example. Finally, the increasing number of users of NEPS data from SC2, SC3, and SC4 might raise the demand for more subgroup-specific weighting adjustments, for example, Grade 5 students participating jointly with their parents in Wave 1 and Wave 3.

References

- Andridge, R. R., & Little, R. J. A. (2010). A Review of Hot Deck Imputation for Survey Non-response. *International Statistical Review*, 78(1), 40–64. doi: 10.1111/j.1751-5823.2010.00103.x
- Bates, D., Maechler, M., & Bolker, B. (2012). *lme4: Linear mixed-effects models using Eigen and Eigenfaces*. Retrieved 03.06.2014, from <http://CRAN.R-project.org/package=lme4>
- Blossfeld, H.-P., Roßbach, H. G., & von Maurice, J. (Eds.). (2011). *Education as a life-long process: The German National Educational Panel Study (NEPS) [Special Issue]: Zeitschrift für Erziehungswissenschaft (Vol. 14)*. Wiesbaden: VS Verlag für Sozialwissenschaften.
- Brick, J. M. (2013). Unit Nonresponse and Weighting Adjustments: A Critical Review. *Journal of Official Statistics*, 29(3), 329–353.
- Durrant, G. B., & Steele, F. (2009). Multilevel modelling of refusal and non-contact in household surveys: evidence from six UK Government surveys. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 172(2), 361–381. doi: 10.1111/j.1467-985X.2008.00565.x
- Greene, W. (2012). *Econometric analysis* (7th ed.). Boston: Pearson.
- Honaker, J., Owen, M., Imai, K., Lau, O., & King, G. (2013). *bprobit: Bivariate Probit Regression for Two Dichotomous Dependent Variables*. Retrieved 21.01.2014, from <http://cran.r-project.org/web/packages/ZeligChoice/vignettes/ZeligChoice-manual.pdf>
- Kalton, G., & Flores-Cervantes, I. (2003). Weighting methods. *Journal of Official Statistics*, 19(2), 81–97.

- Kalton, G., & Kasprzyk, D. (1986). The treatment of missing survey data. *Survey Methodology*, 12, 1–16.
- Kreuter, F., & Olson, K. (2011). Multiple Auxiliary Variables in Nonresponse Adjustment. *Sociological Methods & Research*, 40(2), 311–332. doi: 10.1177/0049124111400042
- Lepkowski, J. M., & Couper, M. P. (2002). Nonresponse in the second wave of longitudinal household surveys. In R. M. Groves (Ed.), *Survey nonresponse* (pp. 259–272). New York: Wiley.
- Little, R., & Vartivarian, S. (2005). Does Weighting for Nonresponse Increase the Variance of Survey Means? *Survey Methodology*, 31(2), 161–168.
- Owen, M., Imai, K., Lau, O., & King, G. (2012). *ZeligChoice: Zelig Choice Models*. Retrieved 03.06.2014, from <http://CRAN.R-project.org/package=ZeligChoice>
- R Development Core Team. (2015). *R: A Language and Environment for Statistical Computing*. Vienna, Austria. Retrieved 26.06.2015, from <http://www.R-project.org/>
- Steinhauer, H. W., Aßmann, C., Zinn, S., Goßmann, S., & Rässler, S. (2015). Sampling and Weighting Cohort Samples in Institutional Contexts. *ASTA Wirtschafts- und Sozialstatistisches Archiv*, 9(2), 131–157. doi: 10.1007/s11943-015-0162-0
- Valliant, R. (2004). The Effect of Multiple Weighting Steps on Variance Estimation. *Journal of Official Statistics*, 20(1), 1–18.

Acknowledgements

This paper uses data from the National Educational Panel Study (NEPS): Starting Cohort 2—Kindergarten, doi:10.5157/NEPS:SC2:1.0.0 and doi:10.5157/NEPS:SC2:2.0.0, Starting Cohort 3—5th Grade, doi:10.5157/NEPS:SC3:2.0.0, and Starting Cohort 4—9th Grade, doi:10.5157/NEPS:SC4:1.1.0. From 2008 to 2013, NEPS data were collected as part of the Framework Programme for the Promotion of Empirical Educational Research funded by the German Federal Ministry of Education and Research (BMBF). As of 2014, the NEPS survey is carried out by the Leibniz Institute for Educational Trajectories (LifBi) at the University of Bamberg in cooperation with a nationwide network.

Tables

Table 4 Parameter estimates of the response propensity models used to derive adjustment factors for weights corresponding to the subgroup of Kindergarten children and parents jointly participating in SC2

	Wave 1		Wave 2		Wave 1 and Wave 2	
	Children ^a	Parents	Children	Parents	Children	Parents
Intercept		0.443*** (0.076)	1.460*** (0.094)	0.023 (0.067)	1.375*** (0.090)	-0.123 (0.067)
Gender		0.077 (0.054)	-0.005 (0.068)	0.029 (0.050)	-0.035 (0.066)	0.055 (0.049)
female						
Age group		-0.037 (0.055)	-0.248*** (0.069)	-0.048 (0.050)	-0.245*** (0.067)	-0.041 (0.049)
older half						
Place of residence		-0.598*** (0.068)	-0.201* (0.085)	-0.741*** (0.064)	-0.208* (0.083)	-0.745*** (0.064)
with one parent or others						
Language spoken at home		0.595*** (0.071)	0.173* (0.087)	0.819*** (0.064)	0.229** (0.083)	0.795*** (0.064)
German						
Random intercept		0.296				
ω Kindergarten level						
Correlation between			0.217** (0.067)		0.175** (0.060)	
children & parents						
Sample Size		2,996	2,996	2,996	2,996	2,996

Notes: The flags ***, **, and * denote significance at the 0.1 %, 1 %, and 5 % level, respectively. Standard errors are given in parentheses. ^a Adjustment factors for Kindergarten children in Wave 1 were computed as $2,996 \div (2,996 - 25)$. To model individual participation, the glmer function with a probit link provided by lme4 package (Bates et al., 2012) in R (R Development Core Team, 2015) was used. To model joint participation decisions, the zelig function with bprobit link provided by ZeligChoice package (Owen et al., 2012) in R (R Development Core Team, 2015) was used. Correlation parameter from the bivariate probit model is transformed according to Honaker, Owen, Imai, Lau, and King (2013). Reference categories are: gender (male), age group (younger half), place of residence (with both parents), language spoken at home (other than German).

Table 5 Parameter estimates of the response propensity models used to derive adjustment factors for weights corresponding to the subgroup of Kindergarten children participating in SC2 Wave 1 as well as in Wave 1 and Wave 2

	Wave 1 ^a	Wave 2	Wave 1 and Wave 2
Intercept		1.725***	1.557***
		(0.122)	(0.112)
Gender		-0.012	-0.045
female		(0.077)	(0.073)
Age group		-0.190*	-0.190*
older half		(0.081)	(0.076)
Place of residence		-0.178	-0.188*
with one parent or others		(0.099)	(0.094)
Language spoken at home		0.112	0.195
German		(0.108)	(0.100)
Random intercept			
ω Kindergarten level			
Sample Size		2,996	2,996

Notes: The flags ***, **, and * denote significance at the 0.1 %, 1 %, and 5 % level, respectively. Standard errors are given in parentheses. ^a Adjustment factors for Kindergarten children in Wave 1 were computed as $2,996 \div (2,996 - 25)$. To model individual participation, the glmer function with a probit link provided by lme4 package (Bates et al., 2012) in R (R Development Core Team, 2015) was used. Reference categories are: gender (male), age group (younger half), place of residence (with both parents), language spoken at home (other than German).

Table 6 Parameter estimates of the response propensity models used to derive adjustment factors for weights corresponding to the subgroup of students participating in SC3 and SC4 Wave 1 and 2, respectively

	Starting Cohort 3		Starting Cohort 4	
	Wave 1	Wave 2	Wave 1	Wave 2
Intercept	1.233*** (0.270)	3.901*** (0.508)	1.814*** (0.091)	2.079*** (0.195)
SC4-stratum $h = 1$	-0.298 (0.278)	-0.542 (0.450)		
Gymnasien				
SC4-stratum $h = 2$	-0.331 (0.288)	-0.560 (0.439)	-0.146* (0.067)	-0.163 (0.172)
Hauptschulen				
SC4-stratum $h = 3$	-0.192 (0.284)	-0.444 (0.450)	-0.070 (0.069)	-0.205 (0.180)
Realschulen				
SC4-stratum $h = 4$	-0.749* (0.310)	-0.194 (0.705)	-0.108 (0.082)	0.152 (0.250)
Integrierte Gesamtschulen Freie Waldorfschulen				
SC4-stratum $h = 5$	-0.636* (0.302)	-0.649 (0.528)	-0.117 (0.094)	0.120 (0.248)
Schulen mit mehreren Bildungsgängen				
SC4-stratum $h = 6$	0.117 (0.302)	-2.129*** (0.450)	-0.207* (0.088)	-1.674*** (0.171)
Förderschulen				
Schools educating students in Grade 5 but not in Grade 9	-0.368 (0.299)	-0.486 (0.524)		
Age group	-0.055 (0.067)	0.189 (0.127)	0.066 (0.045)	0.284*** (0.070)
younger half				
Gender	0.061 (0.063)	0.184 (0.110)	-0.070 (0.038)	0.025 (0.064)
female				
Missing indicator for personal characteristics	-1.148*** (0.116)	-1.143*** (0.196)	-2.259*** (0.329)	-0.056 (0.477)
Native language	1.140*** (0.068)	0.415** (0.148)	0.433*** (0.049)	0.276** (0.088)
German				
Nationality			-0.169* (0.070)	-0.001 (0.111)
German				

Table 6 (continued)

	Starting Cohort 3		Starting Cohort 4	
	Wave 1	Wave 2	Wave 1	Wave 2
Class size			-0.058	-0.094
less than 25			(0.050)	(0.095)
Missing indicator for			-1.323***	-0.705***
migration characteristics			(0.074)	(0.126)
Migration background	-0.169	-0.571		
Turkish	(0.312)	(0.492)		
Student participating in		-0.392		0.566***
wave 1		(0.268)		(0.106)
Individual tracking in		-3.498***		
wave 2		(0.181)		
Random intercept				
ω school level	0.311	0.500	0.276	0.844
Sample Size	6,112	6,098	16,425	16,425

Notes: The flags ***, **, and * denote significance at the 0.1 %, 1 %, and 5 % level, respectively. Standard errors are given in parentheses. To model individual participation, the glmer function with a probit link provided by lme4 package (Bates et al., 2012) in R (R Development Core Team, 2015) was used. Reference categories are: stratum (migrant supplement), age group (older half), gender (male), migration background (other than Turkish), native language (other than German), nationality (other than German), class size (25 or more), missing indicators (no missing values), student participating in Wave 1 (no), individual tracking in Wave 1 (no).

Table 7 Parameter estimates of the response propensity models used to derive adjustment factors for weights corresponding to the subgroup of students and parents jointly participating in SC3 in Wave 1 and Wave 2 as well as in SC4

	Starting Cohort 3				Starting Cohort 4	
	Wave 1		Wave 2		Wave 1	
	Parents	Students	Parents	Students	Parents	Students ^b
Intercept	-0.709*** (0.167)	1.188*** (0.242)	-2.163*** (0.235)	3.018*** (0.395)	-0.689*** (0.079)	
SC4-stratum $h = 1$	0.596*** (0.169)	-0.287 (0.247)	0.661*** (0.219)	-0.439 (0.349)		
Gymnasien						
SC4-stratum $h = 2$	0.239 (0.174)	-0.329 (0.255)	0.186 (0.224)	-0.519 (0.343)	-0.423*** (0.043)	
Hauptschulen						
SC4-stratum $h = 3$	0.390* (0.171)	-0.201 (0.252)	0.499* (0.221)	-0.461 (0.346)	-0.259*** (0.047)	
Realschulen						
SC4-stratum $h = 4$	0.473* (0.185)	-0.768** (0.264)	0.685* (0.238)	-0.332 (0.452)	-0.237*** (0.058)	
Integrierte Gesamtschulen						
Freie Waldorfschulen						
SC4-stratum $h = 5$	0.170 (0.181)	-0.596* (0.265)	0.376 (0.234)	-0.403 (0.398)	-0.502*** (0.062)	
Schulen mit mehreren Bildungsgängen						
SC4-stratum $h = 6$	-0.020 (0.175)	0.105 (0.270)	0.001 (0.226)	-1.590*** (0.347)	-0.559*** (0.058)	
Förderschulen						
Schools educating students in	0.586** (0.179)	-0.352 (0.263)	0.525* (0.230)	-0.262 (0.370)		
Grade 5 but not in Grade 9						
Native language	0.440*** (0.050)	1.099*** (0.064)	0.122 (0.066)	0.243* (0.115)	0.386*** (0.035)	
German						
Migration background	0.409* (0.194)	-0.175 (0.280)	-0.081 (0.246)	-0.376 (0.396)		
Turkish						
Age group		-0.077 (0.063)		0.083 (0.096)	0.120*** (0.027)	
younger half						
Gender		0.061 (0.060)		0.163 (0.087)	-0.043 (0.022)	
female						

Table 7 (continued)

	Starting Cohort 3				Starting Cohort 4	
	Wave 1		Wave 2		Wave 1	
	Parents	Students	Parents	Students	Parents	Students ^b
Missing indicator for		-1.080***		-0.933***	-0.163	
personal characteristics		(0.111)		(0.153)	(0.305)	
Student participating in			0.217*	-0.379	0.032	
wave 1			(0.101)	(0.209)	(0.054)	
Number of calls	1.297***		0.493***		1.437***	
less than 4	(0.043)		(0.048)		(0.027)	
Parent participating in			2.337***	0.308***		
wave 1			(0.051)	(0.087)		
Individual tracking in				-2.589***		
wave 2				(0.099)		
Nationality					0.319***	
German					(0.047)	
Missing indicator					0.171**	
migration characteristics					(0.062)	
Correlation	0.097*		0.415**			
ρ students parents	(0.049)		(0.158)			
Random intercept					0.259	
ω school level						
Sample Size	6,112		6,098		16,425	

Notes: The flags ***, **, and * denote significance at the 0.1 %, 1 %, and 5 % level, respectively. Standard errors are given in parentheses. ^b Because there was no correlation in the participation decisions of students and parents in SC4, decisions were modeled separately. To model individual participation, the glmer function with a probit link provided by lme4 package (Bates et al., 2012) in R (R Development Core Team, 2015) was used. To model joint participation decisions, the zelig function with bprobit link provided by ZeligChoice package (Owen et al., 2012) in R (R Development Core Team, 2015) was used. Correlation parameter from the bivariate probit model is transformed according to Honaker et al. (2013). Reference categories are: stratum (SC3: migrant supplement, SC4: $h = 1$), age group (older half), gender (male), migration background (other than Turkish), native language (other than German), nationality (other than German), number of calls (4 or more), missing indicators (no missing values), student participating in Wave 1 (no), parent participating in Wave 1 (no), individual tracking in Wave 1 (no).

About the authors

H. W. Steinhauer

Leibniz Institute for Educational Trajectories (LifBi), Bamberg.

e-mail: hans-walter.steinhauer@lifbi.de

S. Zinn

Leibniz Institute for Educational Trajectories (LifBi), Bamberg.

Research Group: International Migration, Max Planck Institute for Demographic Research, Rostock.

e-mail: sabine.zinn@lifbi.de

C. Aßmann

Leibniz Institute for Educational Trajectories (LifBi), Bamberg.

Chair of Statistics and Econometrics, University of Bamberg, Bamberg.

e-mail: christian.assmann@uni-bamberg.de

Variance Estimation with Balanced Repeated Replication: An Application to the Fifth and Ninth Grader Cohort Samples of the National Educational Panel Study

Sabine Zinn

Abstract

In order to obtain valid inference, the analysis of survey data requires special approaches to account for sampling design features. This is particularly true when analyzing complex survey data in which inclusion probabilities are not constant, as is the case for the National Educational Panel Study. Here, statistical methods like ordinary least squares estimation might lead to biased conclusions about what social and behavioral processes one might be interested in. One way to achieve proper results even when a statistical method does not explicitly account for survey design features is by using the method of balanced repeated replication. This methodology provides correct assessment of the variances for a wide range of estimators from stratified multistage sampling designs. In this chapter, we describe how to use the method of balanced repeated replication to analyze the data of the first wave of the fifth- and ninth-grader cohort samples of the National Educational Panel Study. We illustrate its capacities by means of two examples: First, we analyze the proportion of migrants in the fifth and ninth grade in German schools. Then, we study the aspiration of students in the ninth grade concerning educational attainment. The results of both applications underline the effectiveness of the method.

1 Introduction

The fifth- and ninth-grader cohort samples of the National Educational Panel Study (NEPS) were established using stratified multi-stage sampling. This is a standard strategy to collect complex survey data by randomly selecting sampling units from clusters at two or more hierarchical levels. In the NEPS, students were sampled from two levels: schools and classes. Before sampling, schools were stratified explicitly according

to school types and implicitly according to the federal states, regional classification, and funding. Then, from school type strata, at the first stage, schools were sampled (primary sampling units), and at the second stage, classes were sampled (secondary sampling units). Subsequently, in each selected class, all students were asked to participate in the survey. Unfortunately, statistical standard estimation techniques have difficulties handling such a design because they are usually applied for simple random sampling. Applying them to a multi-stage sampling design increases the risk of underestimating the variability of survey statistics. The objective of this article is to describe a method that allows for the proper estimation of sampling variances for the NEPS fifth- and ninth-grader cohort samples.

Popular methods that apply to this task are the Taylor series linearization and replication methods (Lee & Forthofer, 2006; Wolter, 2007). Taylor series linearization computes the overall variance estimate as a weighted combination of stratum variance estimates. It is well suited to statistics that have a theoretical derivation of a variance formula, such as the coefficients of generalized linear regression models. However, it cannot be used to compute variance estimates of non-differentiable statistics, such as median and other percentiles. Replication methods are usually used for this purpose. These methods conduct variance estimation by selecting a set of dependent subsamples from the overall sample. The sampling variance of the overall estimate is then derived by computing parameter estimates from each subsample and calculating the variability between the subsample estimates. A prerequisite of the replication methods is that subsamples be formed in such a way that each subsample has the same structure as the parent sample. Replication methods require a sufficiently large number of replicates to yield unbiased statistical inference.

Therefore, pure replication might fail in a stratified design like the NEPS sample design. Here, pseudo-replication methods pose a remedy: The basic idea is to construct subsamples consisting of random groups that represent the sampling units in the different implicit and explicit strata. A systematical formation of these groups allows for computing unbiased variance estimates—even if some strata only comprise a few elements. Jackknife repeated replication, balanced repeated replication, and bootstrapping are common pseudo-replication methods. Jackknife repeated replication works by iteratively removing a single random group from the full sample to create a replicate (Berger & Skinner, 2005; Rao, Wu, & Yue, 1992). In contrast, balanced repeated replication forms a set of replicates by assigning random groups to subsamples in a balanced way (Rust & Rao, 1996; Wolter, 2007). The basic idea of the bootstrap method is to create replicates of the same size and structure as in the parent sample (Efron, 1979). In the past, many research studies have been conducted to assess the quality of each of the three pseudo-replication techniques when used to estimate sampling variances in complex survey designs (see, e. g., Kish and Frankel, 1974; Krewski and Rao, 1981; Rust and Rao, 1996). The main finding is that all three replication techniques show a similar performance for statistics that can be expressed as smooth functions of totals. For statistics that cannot be expressed in this way, such

as sample quantiles, the situation differs. Here, the jackknife method is known to produce inconsistent estimators (Rao & Wu, 1985; Shao & Tu, 1995). Generally, bootstrapping is found to be slightly less effective than balanced repeated replication and jackknife repeated replication because it requires more replicates to reach a comparable precision of the variance estimates (Lee & Forthofer, 2006). In summary, of the three repeated replication methods, the method of balanced repeated replication seems to have the widest application scope and therefore to be the most convenient one for general purposes. multiplying them by the parent sample. Naturally, replication weights comprise all information about the sampling design of a survey. That is, if an analyst cannot access all design features of a survey, replication weights nevertheless allow for regarding the entire sampling design. Legal data security regulations mostly hinder any dissemination of information about non-respondents. This is also the case in the NEPS. Here, information about schools and students who refuse to participate in the study is highly confidential. However, without this kind of information, no nonresponse adjustment can be conducted in order to avoid invalid inference. To make variance estimation possible nevertheless, the NEPS methods group provides replication weights for the method

Jackknife repeated replication, balanced repeated replication, and bootstrapping can be applied without further ado if so-called replication weights are available. Replication weights allow for deriving the set of replicates necessary for variance estimation by simply of balanced repeated replication.¹

The present contribution seeks to describe how these weights can be derived and how they can be used for computing survey statistics for the fifth- and ninth-grader samples of NEPS.² The rest of the article is organized as follows: In Section 2, we detail the structure and the sampling design of the fifth- and ninth-grader cohort samples of the NEPS. In Section 3, we describe the concept of the method of balanced repeated replication and the derivation of accordant replication weights in detail. Furthermore, we detail the adjustment of weights necessary to concord with the design of the parent sample. Section 4 presents the usage of these weights to derive special survey statistics, such as quantiles and population ratios. Section 5 concludes with a critical assessment of the method of balanced repeated replication, revealing its limitations and pointing to multilevel modeling as a powerful alternative.

1 The replication weights are available on request from the author or by writing an email to methods.neps@uni-bamberg.de.

2 All figures presented were computed using the Scientific Use Files with the identification code doi 10.5157/NEPS:SC3:2.1.0 and doi 10.5157/NEPS:SC4:1.1.0.

2 Data and Sampling Design

The fifth- and ninth-grader samples of the NEPS comprise children attending secondary school in the fifth and ninth grade in Germany in the school year 2010/2011. It was built upon a stratified multi-stage sampling design (Aßmann et al., 2011; Aßmann, Steinhauer, & Zinn, 2012): At first, strata were formed; then, schools were sampled from these strata. Classes were selected within the sampled schools. Finally, all students in the selected classes were asked to participate. Schools were selected from the set of officially recognized and state-approved secondary schools in Germany. In this process, six different school types were differentiated: Schools that offer schooling only to children with learning disabilities, *Gymnasien*, *Hauptschulen*, *Realschulen*, *Integrierte Gesamtschulen*, and schools offering all tracks of secondary education except an academic track. We subsequently refer to the latter five school types as regular schools. Special-needs schools and regular schools form the six explicit strata of the ninth-grader sample. The fifth-grader sample consists of three explicit strata that partly overlap with the ninth-grader sample. That is, the first explicit stratum of the fifth-grader sample was established based on five of the six explicit strata of the ninth-grader sample. In more detail, the stratum comprises fifth graders from regular schools that provide schooling to ninth and fifth graders. Special-needs schools make up the second explicit stratum of the fifth-grader sample. Finally, the third explicit stratum of the fifth-grader sample contains children who attend schools that provide schooling only to fifth graders and not to ninth graders. Besides the explicit stratification, an implicit stratification based on the federal states, regional classification, and the organizing institution was used. After sampling schools in the first stage, two classes each (if available) from grade five and nine were sampled within regular schools in a second stage. Thereafter, all children in the selected classes were asked to participate. Students of all classes in special-needs schools were asked to participate in the NEPS.

The first wave of the ninth-grader sample contains information from interviews and tests conducted within two different periods. One took place in autumn 2010 and one in spring 2011. The interviews and tests for the first wave of the fifth-grader sample were conducted in autumn 2010. Overall, the ninth-grader sample comprises information from students from 648 schools: 15,629 students participated in the autumn survey, and 15,308 students participated in the autumn and in the spring surveys. The first wave of the fifth-grader sample comprises 260 schools and contains information on 5,555 students.³ Tables 1 and 2 show the number of schools and students according to the grade sampled within the different strata.

3 For sake of simplicity, we do not consider the additional NEPS sample of fifth-grade students with a Turkish migration background or a migration background related to the former Soviet Union. This is because this sample differs considerably from the basic fifth-grader cohort sample concerning sampling design and structure.

Table 1 Stratum-Specific Numbers of Sampled Schools and Students in the First Wave of the Ninth-Grader Sample

Stratum	Schools	Students in autumn survey 2010	Students in autumn and spring survey 2010 and 2011
Gymnasien	149	5,118	5,069
Hauptschulen	181	3,570	3,515
Realschulen	104	3,108	3,069
Integrierte Gesamtschulen	55	1,617	1,609
Schools offering all tracks of secondary education except an academic track	56	1,127	1,116
Special-needs schools	103	1,089	930

Table 2 Stratum-Specific Numbers of Sampled Schools and Students in the First Wave of the Fifth-Grader Sample

Stratum	Schools	Students
Regular schools offering schooling to fifth graders and not to ninth graders	21	430
Regular schools offering schooling to fifth graders and to ninth graders	182	4,559
Special-needs schools	57	566

3 The Method of Balanced Repeated Replication

The method of balanced repeated replication (BRR) is a widely-used technique for variance estimation in surveys that are subject to stratified multi-stage sampling. It was first introduced by McCarthy (1969) for the case in which only two primary sampling units are sampled with replacement on the first sampling stage. Today, several extensions to the original approach exist that allow the BRR to be applied to a wider scope of tasks (e. g., Rao & Shao, 1999; Shao & Chen, 1999; Shao, Chen, & Chen, 1998; Saigo, Shao, & Sitter, 2001). Before we detail the essentials of the BRR in the following section, we first describe the statistical setting to which the BRR is applied.

3.1 The Setting

Suppose we face a survey sample subject to stratified multi-stage sampling involving H strata. Each stratum h comprises n_h primary sampling units (PSUs), and every primary sampling unit i contains secondary sampling units (SSUs) j . All units k that are part of secondary sampling units are constituted to be fully sampled. If a survey weight is available for each sampled element, an unbiased estimator of a population total Y for a variable y is given (Rao & Shao, 1999)

$$\hat{Y} = \sum_{(h, i, j, k) \in s} w_{hijk} y_{hijk},$$

in which s describes the sample, y_{hijk} is the value of variable y associated to unit (h, i, j, k) , and w_{hijk} is the corresponding sampling weight. In many cases, a survey estimator $\hat{\theta}$ can be written as a function $g(\cdot)$ of a vector of estimated totals:

$$\hat{\theta} = g(\hat{A})$$

with

$$\hat{A} = \sum_{(h, i, j, k)} w_{hijk} a_{hijk},$$

and a_{hijk} is a vector of values corresponding to unit (h, i, j, k) . Examples for such estimators are ratios of two estimated totals, correlation coefficients, and regression coefficients (Shao, 1996). Assuming, for instance, that we are interested in the prevalence of learning disabilities among male students, an estimator for this quantity is the ratio of the number of male students with learning disabilities to the number of male students. Thus, it can be expressed as

$$\hat{\theta} = g(\hat{A}) = \frac{\sum_{(h, i, j, k)} w_{hijk} z_{hijk}}{\sum_{(h, i, j, k)} w_{hijk} x_{hijk}},$$

where x_{hijk} is a dichotomous variable that is coded by 1 if a student is male and 0 otherwise, and z_{hijk} a dichotomous variable that is 1 if a male student suffers from a learning disability and 0 otherwise.

3.2 The Method

The basic idea of the BRR is to construct a set of balanced replicates from random groups in the parent sample. Random groups are commonly only formed from the primary sampling units, disregarding any further sub-sampling. Such proceeding is predicated on the fact that the sampling variance can be approximated adequately from the variation between the totals of the primary sampling units when the first-stage sampling fraction is small (which is usually the case). In survey statistics, this practice is known as the ultimate cluster approximation (Kalton, 1979; Lee & Forthofer, 2006). In its original version, the BRR assumes only two primary sampling units per stratum, namely $n_h = 2$ for all strata h . A single replicate is formed by systematically deleting one PSU from each stratum and then doubling the sampling weights of the primary sampling units remaining. Hence, the replication weight $w_{hijk}^{(r)}$ of entity k located in SSU j and PSU i in stratum h corresponding to the r th replicate is $((h, i, k, k) \in s, r = 1, \dots, R)$:

$$w_{hijk}^{(r)} = \begin{cases} 2w_{hijk}, & \text{if PSU } i \text{ from stratum } h \text{ is part of the } r\text{th} \\ 0, & \text{otherwise.} \end{cases}$$

Because of the practice of neglecting half of the parent sample within each replicate, the BRR is also called the method of balanced half-samples. To promote unbiased variance estimators, the set of replicates has to be *balanced* (Wolter, 2007). That is, each pair of primary sampling units from different strata has to have the same frequency in appearing in the set of replicates. This condition can be formalized to

$$\sum_{r=1}^R \delta_h^{(r)} \delta_k^{(r)} = 0 \text{ for all } h \neq k; h, k = 1, \dots, H,$$

with

$$\delta_h^{(r)} = \begin{cases} +1, & \text{if PSU 1 from stratum } h \text{ is part of the } r\text{th replicate,} \\ -1, & \text{if PSU 2 from stratum } h \text{ is part of the } r\text{th replicate.} \end{cases}$$

A minimal set of balanced replicates can be derived using a Hadamard matrix⁴ of order R in which R is the smallest multiple of four greater than H :

$$H + 1 \leq R \leq H + 4. \tag{1}$$

In more detail, the entries (h, r) of a Hadamard matrix A of order R determine the primary sampling units that have to remain in a half-sample to obtain a balanced set of

4 A Hadamard matrix is a square matrix whose entries are either -1 or $+1$ and whose rows are mutually orthogonal.

replicates.⁵ In other words, $\delta_h^{(r)}$ equals entry (h, r) of matrix A . An approximately unbiased variance estimate is then obtained by

$$\text{var}[\hat{\theta}] = \frac{1}{R} \sum_r (\hat{\theta}^{(r)} - \hat{\theta})(\hat{\theta}^{(r)} - \hat{\theta})',$$

where $\hat{\theta}^{(r)}$ is the survey estimate based on replicate r (i. e., weighted with the replication weights $w_{hijk}^{(r)}$).

One crucial prerequisite for the feasibility of any method of repeated replication is that each single set of replication weights maintain the representation of the population structure in the sample, that is, replication weights have to be adjusted for unit nonresponse at a least condition. At this point, the BRR might encounter severe problems: Because the method implies deleting half of the parent sample, very small sample sizes might result, causing unfeasible adjustment factors for non-response and therefore also nonsensical replication weights. A simple variant of the BRR that allows for overcoming this difficulty is perturbing the replication weights by a factor ε , $\varepsilon \in \{0, 1\}$ (Judkins, 1990):

$$w_{hijk}^{(r)}(\varepsilon) = \begin{cases} (1 + \varepsilon)w_{hijk}, & \text{if PSU } i \text{ from stratum } h \text{ is part of the } r\text{th replicate,} \\ (1 - \varepsilon)w_{hijk}, & \text{otherwise.} \end{cases} \quad (2)$$

The variance estimator that results is

$$\text{var}[\hat{\theta}] = \frac{1}{\varepsilon^2 R} \sum_r (\hat{\theta}^{(r)}(\varepsilon) - \hat{\theta})(\hat{\theta}^{(r)}(\varepsilon) - \hat{\theta})'. \quad (3)$$

For convenience, mostly ε is set to 0.5 (Rust & Rao, 1996).

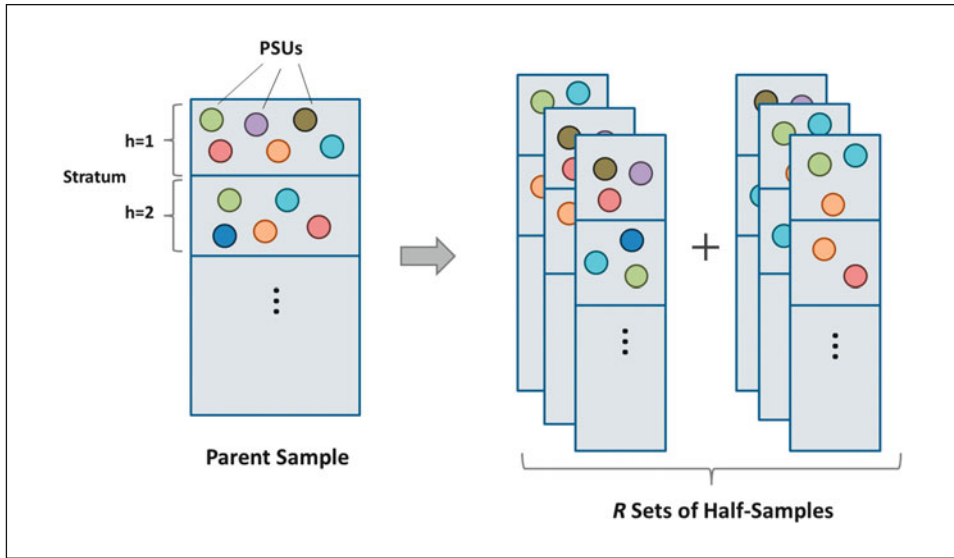
The BRR can easily be extended to cases in which strata comprise more than two primary sampling units. The basic idea is to randomly divide the set of primary sampling units in each stratum h into two groups, g_h^1 and g_h^2 of almost the same sizes, that is,

$$g_h^1 = \left\lceil \frac{n_h}{2} \right\rceil \quad \text{and} \quad g_h^2 = n_h - g_h^1. \quad (4)$$

By means of these groups, a set of balanced replicates can still be constructed using Hadamard matrices (Rao & Shao, 1996): The entry (h, r) of a Hadamard matrix of order R determines whether Group 1 or Group 2 is assigned to a half-sample. Figure 1 illustrates this creation of replicates.

⁵ Here, the row of the Hadamard matrix that consists only of ones is excluded; see Rao & Shao (1996).

Figure 1 Creation of replicates by assigning groups of primary units (PSUs) to half-samples for each stratum separately



For the computation of a survey estimate $\hat{\theta}^{(r)}$ based on replicate r , the replication weights (2) have to be modified in the following way:

$$w_{hijk}^{(r)}(\varepsilon) = \begin{cases} \left(1 + \varepsilon \sqrt{\frac{n_h - g_h^1}{g_h^1}}\right) w_{hijk}, & \text{if entry } (h, r) = +1, \\ \left(1 - \varepsilon \sqrt{\frac{g_h^1}{n_h - g_h^1}}\right) w_{hijk}, & \text{if entry } (h, r) = -1. \end{cases} \quad (5)$$

The variance estimator of this BRR variant does not change and is given by equation (3). In surveys with very large sample sizes, this grouped variant of the BRR might produce asymptotically incorrect results (Valliant, 1987). To overcome this issue, Rao and Shao (1996) suggest repeating the random grouping T times and taking the average of the resulting T BRR variance estimators. To put it more succinctly, random groups are formed T times at first from the primary sampling units in each stratum and always based on the same Hadamard matrix of order R (resulting in R multiplied by T replicates). Then, variance estimators $var^{(t)}(\hat{\theta})$ are computed for the T sets of random groups ($t = 1, \dots, T$). The mean of these estimators constitutes the revised variance estimator:

$$var_T(\hat{\theta}) = \frac{1}{T} \sum_{t=1}^T var^{(t)}(\hat{\theta}). \quad (6)$$

In simulation studies, Rao and Shao (1996) found that $T = 30$ produces unbiased results. The group variant of the BRR fits well to the design of the school samples of NEPS and allows for computing unbiased variance estimators of related survey statistics.

The basic sampling weights might be subject to post-stratification and unit nonresponse adjustment. To capture the possible impact of the weight adjustment on variance estimates, each set of replicates has to be treated with the same adjustment steps as applied to the sampling weights (Rao & Shao, 1999).

4 Application to the Fifth- and Ninth-Grader Samples

To facilitate NEPS data users unbiased variance estimation, the NEPS methods group provides replication weights for the BRR. In this section, we describe how these weights were created and how they can be applied.

4.1 Construction of Replication Weights

The group variant of the BRR is well suited to estimate sample variances from the fifth- and ninth-grader cohort samples of NEPS. Its central element is the stratum-wise formation of random groups of primary sampling units. Here, we have to take into account the fact that the variability of relevant student attributes might not only differ remarkably between the different explicit strata, but also with respect to the variables of implicit stratification (i. e., federal state, regional classification, and funding). To cope with this issue, we followed an approach used in PISA (OECD, 2005): We formed so-called pseudo-strata grouping schools according to explicit and implicit stratification variables. It is important to note that the data at hand comprise only one school for some value combinations of the stratification variables considered in NEPS. For example, the ninth-grader sample comprises only one private special-needs school in Mecklenburg-Vorpommern. However, BRR requires at least two schools per stratum. Thus, it is unfeasible to build pseudo-strata on the basis of all variables of explicit and implicit stratification. For the sake of convenience, we therefore constructed pseudo-strata only according to the variables of explicit stratification and according to a geographical grouping of the federal states. More concretely, we grouped the federal states into northern, southern, western, and eastern states.⁶ In sum, we formed 23 pseudo-strata for the ninth-grader sample and 9 pseudo-strata

6 The northern states are *Schleswig-Holstein*, *Hamburg*, *Mecklenburg-Vorpommern*, *Bremen*, and *Niedersachsen*. The group of southern states contains *Bayern* and *Baden-Württemberg*, and the group of western states consists of *Nordrhein-Westfalen*, *Hessen*, *Rheinland-Pfalz*, and *Saarland*. Finally, the eastern states are *Berlin*, *Brandenburg*, *Sachsen*, *Sachsen-Anhalt*, and *Thüringen*.

Table 3 Numbers of Sampled Schools in the First Wave of the Ninth-Grader Sample According to the Pseudo-Strata Formed

Pseudo-Stratum	Northern States	Southern States	Western States	Eastern States
Gymnasien	30	42	57	20
Hauptschulen	28	92	58	3
Realschulen	20	43	39	2
Integrierte Gesamtschulen	17	2	29	7
Schools offering all tracks of secondary education except an academic track	12	0	9	35
Special schools	23	16	42	22

Table 4 Numbers of Sampled Schools in the First Wave of the Fifth-Grader Sample According to the Pseudo-Strata Formed

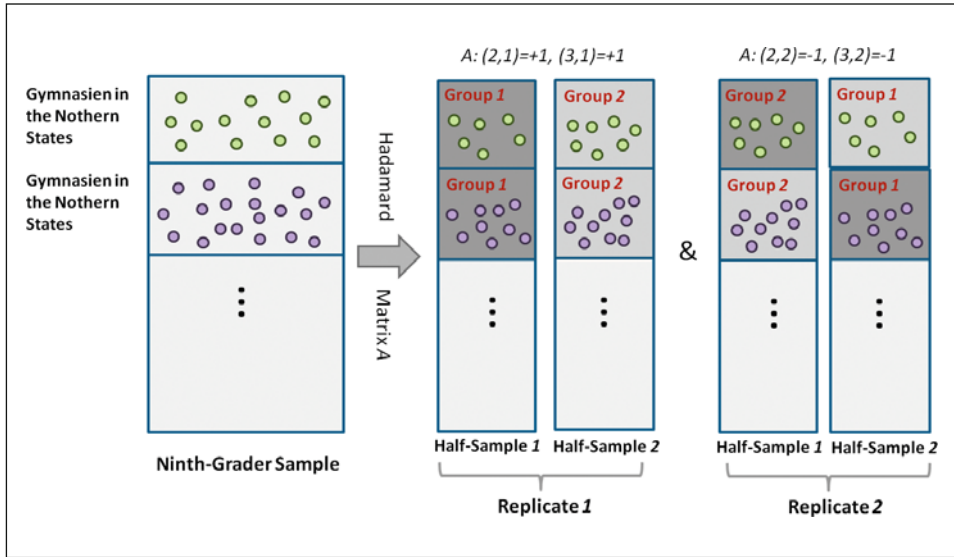
Pseudo-Stratum	Northern States	Southern States	Western States	Eastern States
Regular schools offering schooling to fifth graders and not to ninth graders	40	62	64	16
Special schools	13	10	23	11
Regular schools offering schooling to fifth graders and to ninth graders			21	

for the fifth-grader sample.⁷ The distribution of schools according to distinct pseudo-strata is given in Table 3 and Table 4.

Furthermore, we randomly divided the set of schools in each pseudo-stratum into two groups of almost equivalent size. To determine the group sizes, we used formula (4). Once the groups had been created, we assigned them to one of the two half-

7 Except for the academic track in the southern states, the ninth-grader sample does not contain any school that offers all tracks of secondary education. Therefore, the sum of 24 possible pseudo-strata reduces to 23. The fifth-grader sample comprises three explicit strata: special-needs schools, regular schools offering schooling to fifth graders and not to ninth graders, and regular schools offering schooling to fifth graders and to ninth graders. The latter contain mainly schools in Berlin and Brandenburg. Therefore, schools in this stratum have not been further subclassified according to the four federal state groups defined. Hence, we yield nine pseudo-strata in total for the fifth-grader sample.

Figure 2 The schools in the pseudo-strata “Gymnasien in the northern states” and “Gymnasien in the southern states” of the ninth-grader sample are divided into two random groups: Group 1 and Group 2. According to the entries of the Hadamard matrix A , these groups are assigned to the half-samples of the replicates.



have not been corrected for unit nonresponse. In the NEPS, the data on nonresponse among schools and students is highly confidential due to legal data security regulations. Studies conducted by the NEPS methods group revealed that school and student nonresponse is systematic in the fifth- and ninth-grader samples (Steinhauer, Aßmann, Zinn, Goßmann, & Rässler, 2015). Neglecting this fact when analyzing the NEPS data might lead to bias in survey estimates. To nevertheless allow for the application of the BRR method, the NEPS method group provides replication weights that are adjusted for institutional and individual nonresponse. For the nonresponse adjustment of the replication weights, we employed the same methods and models as were applied to the sampling weights: We used cell weighting to adjust for nonresponse at the school level and response propensity modeling to correct for nonresponse at the individual level. Both approaches are described in great detail in Steinhauer, Aßmann, Zinn, Goßmann, & Rässler (2015).

For the construction of half-samples, we applied a BRR variant that uses a perturbation term (see equation (5)), that is, the distinct half-samples of the replicates schools were weighted differently. Thus, to ensure a reasonable weight adjustment, each school had to enter the nonresponse model accordingly weighted. The according weighting factors $k_{hi}^{(r)}(\varepsilon)$, $h = 1, \dots, H$; $i = 1, \dots, n_i$; $r = 1, \dots, R$, can easily be derived from equation (5):

$$k_{hi}^{(r)}(\varepsilon) = \begin{cases} 1 + \varepsilon \sqrt{\frac{n_h - g_h^1}{g_h^1}}, & \text{if entry } (h, r) \text{ of Hadamard matrix } A \text{ is } +1, \\ 1 - \varepsilon \sqrt{\frac{g_h^1}{n_h - g_h^1}}, & \text{if entry } (h, r) \text{ of Hadamard matrix } A \text{ is } -1. \end{cases}$$

For the fifth- and ninth-grader samples, any post-stratification to external population distributions data was not deemed necessary (Steinhauer et al., 2015). Therefore, the BRR replication weights were not subject to post-stratification either.

4.2 Variance Estimators of Selected Survey Statistics

We subsequently illustrate the application of the BRR replication weights provided by the NEPS methods group. We show how to obtain reasonable variance estimates for two selected survey statistics, which might be of interest when analyzing the NEPS fifth- and ninth-grader sample. First, we compute the proportion of migrants in the fifth and ninth grade in German schools. Second, we study the aspiration of students in the ninth grade concerning educational attainment.

Proportion of migrants in the fifth and ninth grade in German schools

The NEPS fifth- and ninth-grader sample comprises information regarding the migration background of students. Based on this information, the NEPS provides a variable that describes the migration generation status of a student up to the 3.5th generation (Olczyk, Will, & Kristen, 2014). Here, a student is assigned to belong to the group of migrants of the 3.5th generation if at least two of the student's grandparents were born abroad. We use this variable to quantify the proportion of fifth- and ninth-graders with migration background in German schools. In total, 5,555 fifth graders and 15,308 ninth graders took part in the first and the second wave of the NEPS study. The migration background of 5,487 fifth graders and 15,288 ninth graders could be identified. Thus, the migration background of 68 fifth graders and 20 ninth graders is unknown. To determine the proportion of fifth- and ninth-graders with a migration background in German schools, we assume two scenarios: In the first scenario, we assume that none of the students with unknown migration background has a migration background, and in the second scenario, we assume that all students with an unknown migration background have a migration background. Hence, we yield a minimum and a maximum value for the proportion of students with a migration background. To compute the proportions, we account for unequal sampling probabilities by attaching the corresponding sampling weight to each student. The respective values are given in Table 5. Apart from special-needs schools and schools that offer schooling only to fifth graders and not to ninth graders, we find almost the same values for scenario one and two. The proportion of migrants in the fifth grade ranges from 14.1 % to 50.7 %, and in the ninth grade from 19.8 % to 42.4 %. For convenience,

Table 5 Proportion of Students With Migration Background

School type	Fifth graders 1st scenario	Fifth graders 2nd scenario	Ninth graders 1st scenario	Ninth graders 2nd scenario
Grundschulen	0,215	0,242	–	–
Orientierungsstufen	0,433	0,433	–	–
Hauptschulen	0,374	0,374	0,423	0,424
Realschulen	0,279	0,286	0,273	0,274
Gymnasien	0,272	0,274	0,226	0,226
Integrierte Gesamtschulen	0,363	0,376	0,377	0,383
Schools offering all tracks of secondary education except an academic track	0,141	0,147	0,198	0,201
Special schools	0,331	0,507	0,348	0,349

we assume subsequently that all students with an unknown migration background have a migration background, that is, we restrict computations to those that follow to Scenario 2.

To check the significance of the results, we compute confidence intervals for the derived proportions. To do this, we use two approaches: a naïve approach that assumes simple random sampling and the BRR approach. Figures 3 and 4 show the corresponding results. Overall, the sampling variance (and hence, the confidence intervals) achieved by the both approaches do not significantly differ. Nevertheless, we find that in the fifth-grader sample, the naïve approach leads to a slight overestimation in the variability of the proportion of migrants in Gymnasien, special schools, schools offering all tracks of secondary education except an academic track, Grundschulen, and Orientierungsstufen. For Hauptschulen, the approach results in a slight underestimation. Likewise, in the ninth-grader sample, the naïve approach causes a slight underestimation in the variability of the proportion of migrants in Gesamtschulen and a slight overestimation of the migrant proportion in special schools.

Figure 3 Confidence intervals of the proportion of fifth graders with migration background computed without and with using sampling and replication weights (SRW); MB: Schools offering all tracks of secondary education except an academic track

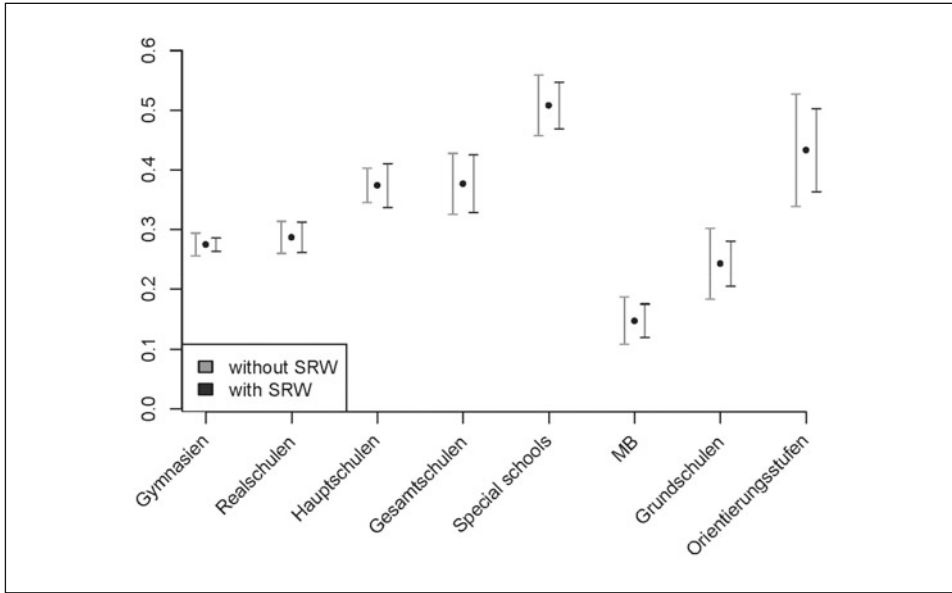
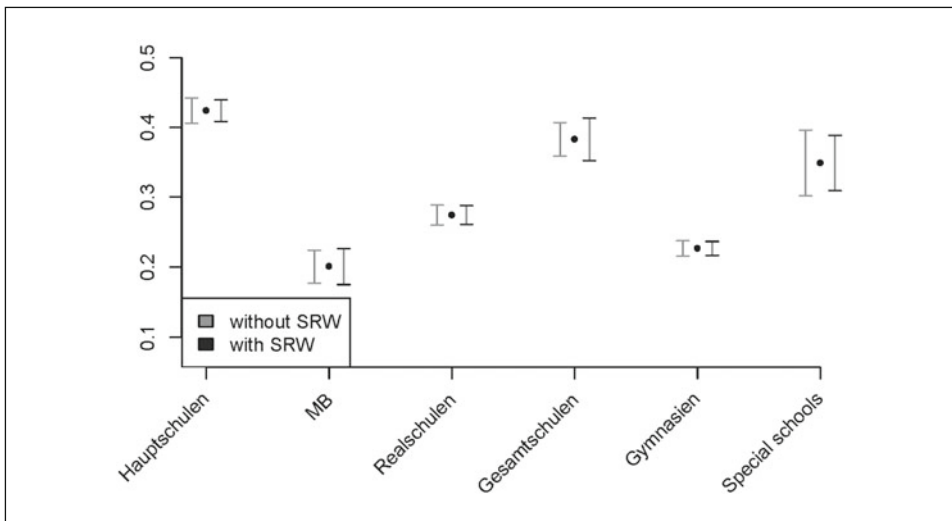


Figure 4 Confidence intervals of the proportion of ninth graders with migration background computed without and with using sampling and replication weights (SRW); MB: Schools offering all tracks of secondary education except an academic track



Aspiration of ninth graders concerning educational attainment

In the ninth-grader sample, students were asked to realistically assess the highest educational degree they might be able to attain. They could choose between leaving school without graduation, lower secondary school with graduation from *Hauptschule* or *Realschule*, and graduation from secondary school qualifying for university admission (*Abitur*). In this simple example, we employ a logistic regression to study the circumstances that drive the educational aspiration of students. The model fits, regardless of whether or not a student aspires to graduate from secondary school qualifying for university admission. We consider six explanatory variables (their values are given after the colon):

- gender: female and male,
- the type of school a student attends: *Gymnasium*, *Hauptschule*, *Realschule*, *Integrierte Gesamtschule*, schools offering all tracks of secondary education except an academic track (MB),⁸
- migration background (at least one parent was born abroad): yes or no,
- the grade point average of a student based on his/her grades in mathematics and German: ranges from 1 to 6,⁹
- the educational attainment of the mother (whether the mother of a student graduated from secondary school qualifying for university admission or not): yes or no,
- the educational attainment of the father (whether the father graduated from secondary school qualifying for university admission or not): yes or no, and
- the socioeconomic status of a student (mapped by the highest value of the ISEI index of both parents): ranges from 10 to 89.

All these variables are available in the ninth-grader sample of the NEPS.¹⁰ In this example, we face a high number of missing values. Complete cases only exist for 49% of all considered cases. A sophisticated way to cope with this problem is to impute the incomplete data by chained equations (*mice*). This approach specifies a multivariate imputation model by a set of conditional densities, one for each incomplete variable (van Buuren & Groothuis-Oudshoorn, 2001). The approach is easily manageable, and associate software exists (van Buuren & Groothuis-Oudshoorn, 2011). To produce consistent variance estimates, we apply the BRR method, with replication weights adjusted for nonresponse among schools and students. Using *mice* and the BRR method in combination does not pose a problem: Consistent results are ensured if the data at hand is (multiply) imputed as often as replicates exist—applying the respective set of replication weights each time. Thereafter, the results are combined as

8 In this analysis, we omit students from special-needs schools because they hardly aspire towards a graduation from secondary schools that would qualify them for university admission.

9 In the German schooling system, grade “1” indicates the best achievement and “6” the worst.

10 The respective Scientific Use Files are available at the NEPS data center, doi: 10.5157/NEPS:SC4: 1.1.0.

Table 6 Coefficients of the Estimated Model and the Associated 95 % Confidence Intervals (CI) With and Without Applying Sampling and Replication Weights (SRW)

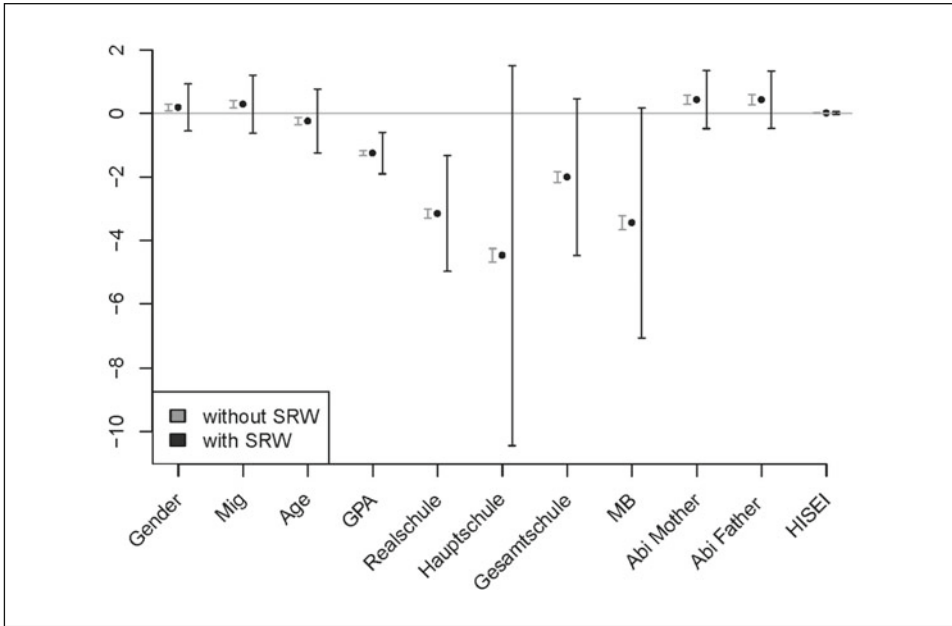
Variable	Coefficient	CI without SRW	CI with SRW
Age			
Older than the median age	-0.241	(-0.356, -0.125)	(-1.246, 0.765)
Gender			
Female	0.188	(0.087, 0.290)	(-0.553, 0.931)
School type			
Hauptschule	-4.469	(-4.683, -4.255)	(-10.443, 1.505)
Realschule	-3.153	(-3.290, -3.015)	(-4.976, -1.329)
Integrierte Gesamtschule	-2.004	(-2.167, -1.841)	(-4.468, 0.460)
MB	-3.441	(-3.662, -3.219)	(-7.050, 0.169)
Migration background			
Yes	0.291	(0.170, 0.411)	(-0.620, 1.203)
Mother has Abitur			
Yes	0.434	(0.295, 0.573)	(-0.478, 1.346)
Father has Abitur			
Yes	0.431	(0.281, 0.581)	(-0.474, 1.337)
Grade point average	-1.249	(-1.327, -1.170)	(-1.901, -0.596)
Highest value of ISEI	0.014	(0.011, 0.017)	(-0.029, 0.056)
Number of cases:* 14,373.			

* After subtracting all special school students from the sample of ninth-grade students attending the first and the second wave of NEPS, the sample comprises 14,378 students. Five of these students did not participate in the survey. For them, only data on parental interviews are available. Hence, 14,737 students remain for analysis.

described in Section 3.2. While the computational burden of such processing is high, the hardware is less of a problem today. We estimate all logistic regression models with sampling weights using the method of weighted least squares.¹¹ Table 6 shows the coefficients of the estimated model and the associated 95 % confidence intervals with and without applying sampling and replication weights for variance estimation; see also Figure 5.

11 In order to fit the model, we used the *lrm* function of the *rms* package of the statistical software R.

Figure 5 The coefficients of the estimated logistic regression model (black dots) and the associated 95 % confidence intervals (vertical lines in grey) computed with and without using weights (SRW); Mig: migration background, GPA: grade point average, MB: schools offering all tracks of secondary education except an academic track, Abi Mother: mother has Abitur, Abi Father: father has Abitur, HISEI: highest value of ISEI



The results demonstrate that neglecting the sampling design when estimating the variance of the regression coefficients leads to a clear underestimation. When applying sampling weights and replication weights for variance estimation, that is, when accounting for the sampling design, most of the effects that were significant before become insignificant. However, considering the fact that the sample at hand is not a self-weighted one but is subject to rather variable sampling weights,¹² this is not a surprising outcome. In such cases, sampling weights almost always increase the standard errors of regression estimates (Gelman, 2007). The reason is that sampling weights are derived to allow for making inferences on the population level, and if sampling weights vary notably, any analytical inference reflects the accordant uncertainty. In conclusion, the example shown underlines the feasibility of using an approach accounting for the sampling design when estimating sampling variances.

A proper way to circumvent sampling weights and variance corrections like the BRR method in a regression model is by including the sampling design in the mod-

12 The sampling weights range from 0.114 to 2.454, with a median value of 0.928.

eling process. In other words, the modeler uses a modeling approach that maps the structure of the sample and includes all design-specific variables as explanatory variables. In the considered example, this could be achieved by using a multi-level model that accounts for the fact that students are nested within schools and that regards federal state differences in the German educational system.

5 Conclusion

In this article, we have described how the BRR method can be applied to obtain (approximately) unbiased variance estimators for survey statistics computed from the NEPS fifth- and ninth-grader samples. For this purpose, we first detailed the principles of the BRR method and then presented its application to the NEPS data. In this context, we elaborated the derivation of replication weights necessary to conduct the method. Furthermore, we described how these weights were adjusted to cope with unit nonresponse among schools and students. Finally, we illustrated the BRR method by means of two examples: First, we computed estimates and confidence intervals for the proportion of migrants in the fifth and ninth grade in German schools. Then, we employed a logistic regression to study the aspiration of students in the ninth grade concerning educational attainment. The results of both applications underline the importance of using an approach that regards the sampling design of the NEPS data for variance estimation.

Currently, replication weights are built using pseudo-strata formed according to school types and a geographical grouping of federal states. For many objectives, such a classification is absolutely sufficient. However, depending on the subject being studied, further school characteristics might also have explanatory power on the variance of the subject of interest. For example, concerning an offering of additional educational courses, schools that are financed by public money might considerably differ from schools that are financed by private money. Likewise, whether a school is located in a city or the countryside might have an effect on the variety of the cultural activities of the students. The fifth- and ninth-grader samples of NEPS are not rich enough to facilitate a meaningful sub-classification according to a large set of stratification variables. That is, in any case, the feasibility of the variance estimates computed by using the replication weights provided depends on the research object.

In view of the sampling design of the considered samples, multilevel modeling offers a general alternative to simple regression modeling. It allows for a direct consideration of the design-based features. Multilevel event history models, in particular, lend themselves to the description of longitudinal data sets such as the NEPS fifth- and ninth-grader samples.

References

- Aßmann, C., Steinhauer, H., Kiesl, H., Koch, S., Schönberger, B., Müller-Kuller, ... Blossfeld, H.-P. (2011). Sampling design of the National Educational Panel Study: Challenges and solutions. In H.-P. Blossfeld, H.-G. Roßbach, & J. von Maurice (Eds.), *Zeitschrift für Erziehungswissenschaft, 14. Education as a lifelong process: The German National Educational Panel Study (NEPS)* (pp. 51–65). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Aßmann, C., Steinhauer, H. W., & Zinn, S. (2012). *Weighting the fifth and ninth grader cohort samples of the National Educational Panel Study, panel cohorts*. (NEPS Technical Report). Bamberg: University of Bamberg, National Educational Panel Study.
- Berger, Y. G., & Skinner, C. J. (2005). A jackknife variance estimator for unequal probability sampling. *Journal of the Royal Statistical Society B*, 67(1), 79–89.
- Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1), 1–26.
- Gelman, A. (2007). Struggles with survey weighting and regression modeling. *Statistical Science*, 22(2), 153–164.
- Judkins, D. R. (1990). Fay's method for variance estimation. *Journal of Official Statistics*, 6(6), 223–239.
- Kalton, G. (1979). Ultimate cluster sampling. *Journal of The Royal Statistical Society*, 142(2), 210–222.
- Kish, L., & Frankel, M. (1974). Inference from complex samples. *Journal of the Royal Statistical Society B (Methodological)*, 36(1), 1–22.
- Krewski, D., & Rao, J. N. (1981). Inference from stratified samples: Properties of linearization, jackknife and balanced repeated replication methods. *Annals of Statistics*, 9(5), 1010–1019.
- Lee, E. S., & Forthofer, R. N. (2006). *Analyzing complex survey data* (2nd ed.). London: SAGE Publications.
- McCarthy, P. J. (1969). Pseudo-replication: Half samples. *Review of the International Statistical Institute*, 37(3), 239–264.
- Olczyk, M., Will, G., & Kristen, C. (2013). *Personen mit Zuwanderungshintergrund im NEPS: Wege zur Identifizierung von Generationenstatus und Herkunftsgruppe*. (NEPS Working Paper 41a). Bamberg: University of Bamberg, National Educational Panel Study.
- OECD. (2005). *PISA 2003 Data Analysis Manual, SAS Users* (Report No. 54093 2005). Paris: OECD.
- Rao, J. N., & Shao, J. (1996). On balanced half-sample variance estimation in stratified random sampling. *Journal of the American Statistical Association*, 91(433), 343–348.
- Rao, J. N., & Shao, J. (1999). Modified balanced repeated replication for complex survey data. *Biometrika*, 86(2), 403–415.

- Rao, J. N., & Wu, C. F. (1985). Inference from stratified samples: Second-order analysis of three methods for nonlinear statistics. *Journal of American Statistical Association*, 80(391), 620–630.
- Rao, J. N., Wu, C. F., & Yue, K. (1992). Some recent work on resampling methods for complex surveys. *Survey Methodology*, 18(2), 209–217.
- Rust, K. F., & Rao, J. N. (1996). Variance estimation for complex surveys using replication techniques. *Statistical Methods in Medical Research*, 5(3), 283–310.
- Saigo, H., Shao, J., & Sitter, R. R. (2001). A repeated half-sample bootstrap and balanced repeated replications for randomly imputed data. *Survey Methodology*, 27(2), 189–196.
- Shao, J. (1996). Resampling methods in sample surveys. *Statistics: A Journal of Theoretical and Applied Statistics*, 27(3-4), 203–237.
- Shao, J., & Chen, Y. (1999). Approximate balanced half sample and related replication methods for imputed survey data. *The Indian Journal of Statistics*, 61(1), 187–201.
- Shao, J., & Tu, D. (1995). *The jackknife and bootstrap*. New York: Springer.
- Shao, J., Chen, Y., & Chen, Y. (1998). Balanced repeated replication for stratified multi-stage survey data under imputation. *Journal of the American Statistical Association*, 93(442), 819–831.
- Steinhauer, H., Aßmann, C., Zinn, S., Goßmann, S., & Rässler, S. (2015). Sampling and weighting panel cohorts in institutional contexts. *AStA Wirtschafts- und Sozialstatistisches Archiv*, 9(2), 131–157.
- Valliant, R. (1987). Some prediction properties of balanced half-sample variance estimators in single-stage sampling. *Journal of The Royal Statistical Society Series B*, 49(1), 68–81.
- van Buuren, S., & Groothuis-Oudshoorn, K. (2001). Mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45(3), 1–67.
- van Buuren, S., & Groothuis-Oudshoorn, K. (2011). *Mice: Multivariate imputation by chained equations. R package version 2.9*. Retrieved from <http://CRAN.R-project.org/package=mice>
- Wolter, K. M. (2007). *Introduction to variance estimation* (2nd ed.). New York: Springer.

About the author

S. Zinn
Leibniz Institute for Educational Trajectories (LifBi), Bamberg.
Max Planck Institute for Demographic Research, Rostock.
e-mail: sabine.zinn@lifbi.de

Challenges in Gaining Access: The School Cohorts of the National Educational Panel Study

André Müller-Kuller, Sonja Meixner and Michaela Sixt

Abstract

One of the German National Educational Panel Study's (NEPS) main responsibilities is to collect data on the educational processes and competence development of students in schools. Access to the NEPS school samples is administered on an institutional basis. With regard to the multilevel-/multi-informant perspective, parents, teachers, and school principals are also requested to participate in order to receive important context information about our target persons. In addition to the special requirements of this study design, major challenges to the administration of school studies arise mainly due to the federal sovereignty and responsibility of the German educational system and to the fact that participation in all NEPS surveys is voluntary. Furthermore, the special design of the NEPS and its claim of maintaining cross-cohortal coherence provide challenging tasks, particularly with regard to obtaining access to schools and requesting the authorization of surveys in each wave. During the negotiation and administration processes within a multidimensional system with multiple players (scientists, ministries, schools, targets, etc.), various—sometimes competing and changing—interests need to be brought in line. In addition, dealing with and operationalizing a multicohort sequence design requires specific strategies that have had to be implemented to cope with the inherent complexity of the study. This article points out some central aspects, developments, and efforts in dealing with these challenges after three years of NEPS fieldwork.

1 The School Cohorts in the NEPS Multicohort Sequence Design

Education has become one of the most important key factors not only for societies and economics, but also for individual life chances. However, compared with the importance of education, there is rather little knowledge about how competencies develop and education is acquired over the life course in Germany. There are no adequate panel studies, and there is a lack of longitudinal data (Blossfeld & Schneider, 2011). The National Educational Panel Study (NEPS), with its multicohort sequence design, was set up to find out more about how education is acquired, to understand how it impacts individual biographies, and to describe and analyze the major educational processes and trajectories across the life span (Blossfeld, von Maurice, & Schneider, 2011). Not only is the design challenging, but so, too, is getting access to and recruiting the participants as well as the administration of the surveys in the different cohorts of the NEPS, which is described for the school cohorts, in particular, in the following section.

In the following section, the current text provides a deeper insight into the design and realization of the school cohorts located in the NEPS multicohort sequence design. Section 2 documents the challenges arising from legal regulations, and Section 3 focuses on aspects relevant in accessing and administering school surveys.

The multicohort sequence design is a very innovative and also complex design for collecting comparative data on competence development and educational pathways over the whole life course. The idea is to start with six panels at the same time and follow the participants on their educational careers through manifold educational stages and transitions. By representing the life course throughout these six cohorts and surveying them parallel to one another, the NEPS is able to provide data for the whole life course after only a few years.

Implementing the multicohort sequence design in the first funding phase means designing and administering 72 main surveys in all cohorts, which together represent about 60,000 target persons and 40,000 context persons. Because of our standardized process of developing, evaluating, and optimizing the instruments and survey procedures, there are about another 90 development, pilot, and linking studies that accompany our main studies. The preparation of the instruments for a main study begins with an extensive review of literature, expert interviews, and/or cognitive pretests. Then, several development studies follow, particularly those for the competence tests. Furthermore, there is always a pilot study one year before a main study goes into the field. In the pilot studies, instruments and survey procedures identical to those used in the main survey are employed to test if everything works well. There are some studies, especially in the school cohorts, to investigate mode effects, for example, if a change from testing with paper and pencil to computer-based assessment makes a difference. Additionally, there are linking studies after the main surveys to assure a comparable measurement of competencies over the life course. Each sub-study follows standardized procedures and is accompanied by the Survey Coordination De-

partment to implement the multicohort sequence design effectively and to collect data all in one piece (see Ristau & Beyer, this volume). Furthermore, implementing the multicohort sequence design and producing comparable data does not mean that the same techniques are used in every sample. Rather, this means that it is important to think very carefully about the target persons and their specific situations, especially when designing surveys with cohort-specific instruments, survey modes, and motivation- and incentive strategies. This is especially the case when looking at the two school cohorts because there are various possibilities of transitions to other educational trajectories.

In most of the Federal States of Germany, students enter secondary education after Grade 4.¹ Upon the transition, the lower secondary school system splits into different tracks or types of school systems, principally the *Hauptschule*, *Realschule*, *Gesamtschule*, and *Gymnasium*. The *Gymnasium* and the *Gesamtschule* are the tracks that lead directly to upper secondary education and a certificate for entering higher education, whereas the other tracks prepare students for vocational training. Lower secondary education ends with Grades 9 or 10, depending on the federal state. Students may then enter upper secondary school (*gymnasiale Oberstufe*), which is situated essentially in two school types, namely the *Gymnasium* and the *Gesamtschule*. Alternatively, students may enter the vocational educational system or the labor market. To cover all these transitions in detail, the NEPS contains two panel studies in schools, with cohorts starting in Grades 5 and 9 (see von Maurice, Blossfeld, & Roßbach, this volume).

Although it would be efficient to do the same survey procedures in both panels, the participants' different educational situations, which lead to differences in the cohort- and survey design, have to be kept in mind. For both school cohorts, we draw a class-based sample (see Section 3) and follow the students on their way through school beginning with Grades 5 and 9 in the fall of 2010. Because very little is known about students with special educational needs in the area of learning (SEN-L), the NEPS also draws an additional panel of this population to answer the question of whether and how students with SEN-L can be meaningfully included in the large-scale assessment. This question is examined by a series of feasibility studies (see Nusser, Heydrich, Carstensen, Artelt, & Weinert, this volume, and Heydrich, Weinert, Nusser, Artelt, & Carstensen, 2012).

The surveys of the first waves are based on group testing in school with several paper-and-pencil tests and a paper-and-pencil questionnaire for the students, paper-and-pencil questionnaires for the teachers, as well as paper-and-pencil questionnaires for the principals of the schools. Furthermore, there is a computer-assisted telephone interview with one parent to obtain even more background information. Surveying context persons completes the picture of the social and learning environments of

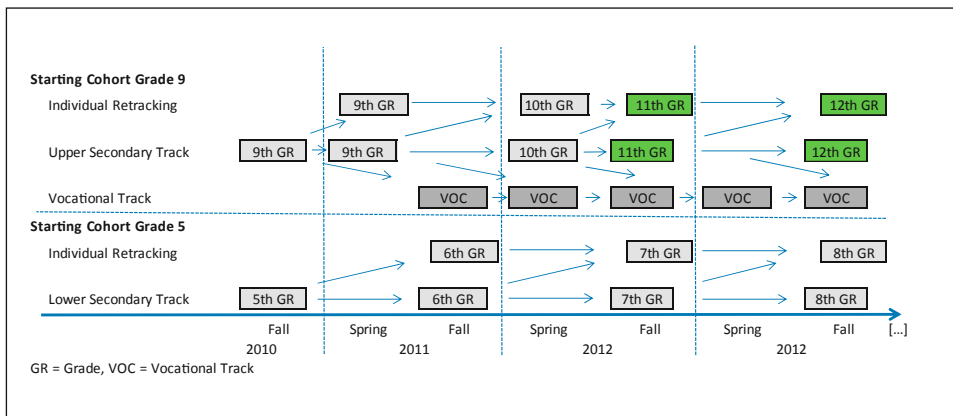
1 There are two exceptions: In Berlin and Brandenburg, there are another 2 years of elementary school before students enter lower secondary school after Grade 6.

the students. Furthermore, information about the regional context, are available for the schools and the homes so that metadata on regional and local levels down to the families' neighborhoods can be merged. In both cohorts, a multi-informant (student, parent, and teacher) as well as a multilevel perspective (student, class, and school) is of central importance to get as much information as possible and to comprehensively map the participants' contexts.

In both NEPS school cohorts, tests and questionnaires for the students are administered in groups at school, and contact with the target persons is organized via the school. As long as the respondents visit the schools where the NEPS is conducted, it is comparatively easy to reach the respondents and to keep them in the panel. However, if a respondent leaves the NEPS school because he or she has changed schools, or if his or her school cancels its participation in the study, other methods must be found to stay in contact with this special group of respondents and to collect data in a way that is comparable with the main field survey. Therefore, a concept of surveying these respondents in an individualized way has been developed by NEPS: the field of individual retracking. With this individual field in the school cohorts, the NEPS is able to survey not only the mainstream and the standard paths through school, but also nonstandard careers and individual pathways over the life course (see Sixt, Goy, & Besuch, this volume). As a result, the first transitions that the NEPS has to handle in its school cohorts are the transitions from the main field in school into the field of individual retracking (see Figure 1).

In addition to this adjustment of the design in our school cohorts, we are also interested in the transition to the vocational educational system. For the cohort of the ninth graders, it is possible to change from upper secondary education to the vocational educational system. Because respondents leave school and spread out over the whole country, a group-based survey would be impossible. Therefore, in this educa-

Figure 1 Transitions in the school cohorts



tional stage, we switch the survey mode and administer a computer-based telephone interview as well as face-to-face testing and interviewing at home with a survey program related to the instruments for the participants who stay in school. Furthermore, we increase the survey cycle to semi-annual surveys to stay in contact and get information about transitions (Ludwig-Mayerhofer, Solga, Leuze, & Dombrowski, 2011).

As shown, the designs of the two school cohorts are rather complex due to the manifold individual pathways NEPS wants to cover. Besides the challenges connected with the design of the study, there are special challenges in getting access to the school cohorts.

2 Legal Challenges

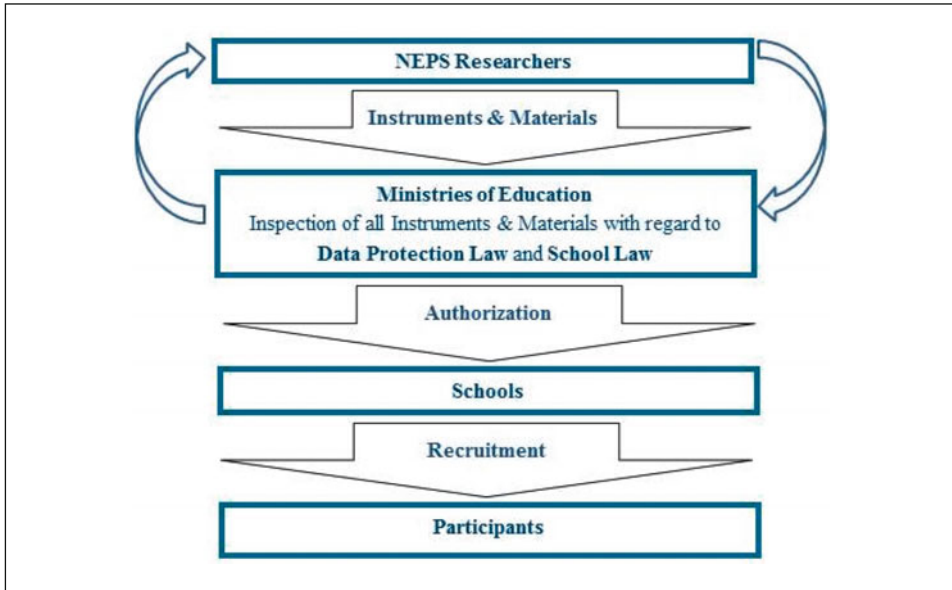
First of all, it is important to note that the participation in all NEPS surveys is—in contrast to other representative large-scale assessments—voluntary for all participants. Schools, school principals, teachers, students, and students' parents are not obligated to respond; rather, they can freely decide to do so. Therefore, it is the responsibility of NEPS researchers to make participation in the study popular for every target institution and person and also to manage the legal challenges that emerge from the structural condition (for a total of about 70 sub-studies in the context of the two starting cohorts from 2009 to 2013). The challenges mainly arise due to the federal sovereignty and responsibility of the German educational system. Conducting a national school survey in Germany, as well as gaining access to schools and recruiting respondents within the school context, in particular, is primarily a special case of jurisdiction for this reason.

In Germany, national school surveys require an explicit permit by the ministry of education of each federal state.² In general, the ministerial authorization process includes the control of compliance of all survey procedures and documents with the school law and the data protection law. The intended sampling procedures and data collection procedures, in particular, as well as all instruments and materials from each single sub-study, are inspected and need ministerial consent before they may be used in the field (see Figure 2).

Primarily, the respective educational act of the federal state is taken by the ministries as a basis for their review process. The ministries' task here is to maintain the interests of their subordinated institutions and the associated persons, such as school principals, teachers, students, and students' parents, as well as to assume direct responsibility for these institutions concerning scientific research projects within the school context. For example, the ministries check the questionnaires to see if there are any questions on them that may lead the students to self-accusation, such as ques-

2 North-Rhine-Westphalia represents an exception in which explicit ministerial permission is not necessary.

Figure 2 Negotiation levels within the legal authorization process



tions about criminal behavior. Such questions are not allowed and are eliminated from the questionnaires by the ministries, even if the scientific value of these questions may be high and researchers expect new and interesting findings. Another objective of the ministerial review process is the reduction—as far as possible—of the cognitive and temporal burden of the students, which comes along with the surveys. If the questioning and testing takes too much time in the ministries' opinion, they impose reductions on the instruments. Furthermore, the ministries critically analyze the time and effort that schools need to spend on the coordination and organization of the surveys running in these schools. They pay attention to the preservation of the daily routine in schools and therefore place a high value on effective survey procedures.

As already mentioned above, the ministries also take care of the compliance of the survey with data protection regulations. Priority here is given to the respective educational act of the federal state that regulates data protection issues for research projects in schools to some extent. If the educational act does not comprise appropriate data protection regulations, the ministries of education refer to the data protection act of their federal state or to the German Federal Data Protection Act (*Bundesdatenschutzgesetz*) as a guiding framework of data protection issues (Meixner, Schiller, von Maurice, & Engelhardt-Wölfler, 2011).

The goal of this part of the ministerial review process is to take responsibility for the right to privacy of all survey participants within the school context. The right

to privacy includes the right to informational self-determination. This derives from the Basic Constitutional Law of the Federal Republic of Germany (*Grundgesetz für die Bundesrepublik Deutschland*) and simultaneously from the German Federal Data Protection Act. According to this act, all survey participants are to be protected from unregulated disclosure and utilization of personal data. Therefore, the ministries of education examine whether respondents put themselves at a disadvantage through their participation in the NEPS surveys and the disclosure of data that comes along with these surveys, especially because the participants' contact data are collected due to the panel design of the NEPS, which requires questioning and testing the same students several times. In general, contact data are very sensitive data because they allow for a clear re-identification of the participants; therefore, the data protection laws provide a strict handling of this sort of data (see also Arbeitskreis Deutscher Markt- und Sozialforschungsinstitute e. V. (ADM), Arbeitsgemeinschaft Sozialwissenschaftlicher Institute e. V. (ASI), Berufsverband Deutscher Markt- und Sozialforscher e. V. (BVM), & Deutsche Gesellschaft für Online-Forschung e. V. (DGOF), 2011; Meixner et al., 2011). In this context, the ministries of education therefore focus strongly on the NEPS' compliance with the data collection-, data processing-, and data dissemination procedures in terms of the current and general data protection standards. Moreover, they attach high importance to the method of collecting the respondents' consent to participate in the NEPS surveys and, in addition, to the kind of information about the survey with which they are provided.

Ultimately, the ministerial review process results in a professional statement about the planned school survey by the ministry of education of each federal state. When a state claims modifications, negotiations between the ministry and the researchers about possible adjustments follow (see Figure 2). In general, these negotiations are very complex and resource-intensive.

In general, the negotiations focus on finding appropriate solutions for all federal states in order to avoid any variations due to federal-state-specific adjustments as far as possible. In this context, it is also essential to take account of the longitudinal design of the NEPS and the claim of cross-cohortal coherence in questioning and testing. Altogether, this is quite a difficult task considering the 16 different school laws as basis for the negotiations. Furthermore, despite the existence of a general German Federal Data Protection Act and of quite similar core elements of the accompanying 16 data protection acts of the federal states, slight differences between the formulations of the data protection laws leave room for interpretation. This contributes to active discussions with the ministries, and endeavors are made to come to a mutual agreement.

Within the negotiations between the researchers and the ministries, scientific interests and requirements sometimes compete with ministerial interests. Metschke and Wellbrock (2000) and as well Häder (2009) point out that the freedom of science, in particular, which is guaranteed by the Basic Constitutional Law of the Federal Republic of Germany, may collide with general personal rights, mainly the right to in-

formational self-determination. The deriving challenge for social science is therefore to find an acceptable compromise between both claims and their realization. This requires that the researchers, especially the survey management, have a solid knowledge about both sides to bring the different interests in line and to get ministerial authorization from each federal state in time. As a result, not only do divergent interpretations and motivations have to be coordinated, but so, too, do the different intensities, foci, and lengths of authorization procedures. For instance, the negotiation processes in the first surveys took 10 to 22 weeks before reaching the federal states, and the requirements of the ministries varied greatly. To make matters worse, ministerial requirements also change over time just as negotiators change. In the meantime, these interactions with the ministries of education are routinely incorporated within the NEPS survey processes but steadily call for interaction and cooperation.

In addition to the authorization of the school surveys by the ministries of education, the sampled schools also need to show a willingness to cooperate, particularly in recruiting the participants on-site (see Figure 2). At the school level, in some federal states, for example, in Hesse and North-Rhine-Westphalia, the consultation of the respective school conference and the approval of the parents' association is necessary before carrying out a scientific research project at school. After the agreement of the school to the participation in the NEPS surveys, students are sampled and asked about participation. With regard to the multi-informant perspective of the NEPS, students' parents, teachers, and school principals are also requested to participate. As mentioned above, there is no law obligating people to participate in the NEPS. Thus, according to the data protection laws, the respondents' written consent is needed unless special circumstances require another form, for example, telephone interviews (see also Iraschko-Luscher, 2006). The written form is intended to protect people against overly hasty or thoughtless consent and to force them to think about the possible consequences of this consent beforehand. However, this written form of consent might increase the nonresponse rate in certain social groups and endanger the representativeness of the sample (Arbeitskreis Deutscher Markt- und Sozialforschungsinstitute e. V. (ADM), n. d.; Metschke & Wellbrock, 2000).

A precondition of the students' participation in the NEPS school surveys is the (written) consent by both the student and one of his or her parents. Because parents are responsible for their minor children (under 18 years) in school issues, they need to agree to the participation of their children in the NEPS and sign the corresponding consent form. Since the German data protection laws do not define an age limit for the individual capacity of the discernment of minors, and since it is impossible to verify this for each participant in the context of the study, we rely upon German Criminal Law and ask for written consent in Grade 9 when the majority of the participants have reached age 14.

Generally, the consent is only valid and effective if the individual has the ability to form a rational judgment about the issue and to make a free decision on his or her participation or non-participation in the NEPS. Giving adequate and sufficient

information to the participants is therefore an essential precondition to anticipate the possible consequences of the participants' consent. This information includes information about the sort of collected data (survey and contact data), the purpose of data collection, the data processing, and the data utilization, as well as about the voluntariness of the participation in the NEPS and the possibility of withdrawing consent at any time. In this context, it is not easy (and often not allowed) to avoid juristic wording in the letters for the participants. Nevertheless, it is important that everyone understand the information about the study, regardless of any individual's educational background. We therefore also distribute easy-to-understand, target-specific flyers to the participants, which sum up the most important information about the NEPS.

3 Challenges in Accessing and Administering School Surveys

In addition to the legal challenges in surveying students and the corresponding context persons, special challenges with respect to accessing and administering school surveys must be coped with. As described in Section 1, the sampling of the NEPS school cohorts, and therefore, the access to and the surveying of participants, is based on institutions. We chose this sampling strategy because institutions play a central role in educational mediation. The strategy offers the possibility of the direct involvement of educationally relevant mediators and informers, such as teachers and school principals. Furthermore, this access and survey strategy (i. e., group testing in relatively stable settings) is expected to be highly standardized and relatively cost- and resource-efficient. The management of the challenging access and administration of the NEPS school cohorts is described in the following section.

- 1) The *size* of the NEPS school samples and their *composition* (i. e., the stratification and the desire of representativeness of the samples; Aßmann et al., 2011) requires particular efforts in the processes of sampling, recruitment, and survey administration. More than 22,000 students in more than 800 institutions in 16 different educational systems have to be handled in both of these cohorts. To produce comparable data for Germany, consistency and comparability of the survey instruments is just as important as the need for equal survey procedures across all strata and federal states. At all, this has to be done with respect to the structural peculiarities of each country. Additionally, specific conditions relating to special sample populations, such as adjusted instruments for students with special needs and motivation letters in foreign languages for migrants, must be taken into account.
- 2) As demonstrated above, the implementation of a *panel design* in school surveys requires special legal procedures as well as measures to ensure the maintenance of coherency and the comparability of the instruments and survey procedures over the waves.

Furthermore, additional requirements arise from the extensive survey program resulting from the theoretical framework of the NEPS (see Blossfeld, von Maurice, & Schneider, 2011) that cannot be realized in one survey taken during one school day. Therefore, the survey program must be split up into several waves with the same participants, and a variety of different instruments have to be administered in one survey (2–4 school hours).

In addition, the organization of the surveys must consider the fact that the survey time (i. e., the length of each survey instrument as well as the whole time, including administration time) is not identical to the duration of a common lesson or a school day as such. Moreover, in each interview and test setting, additional supervisors (i. e., teachers) have to be organized for legal reasons. Additionally, surveying the same institutions and participants repeatedly and/or on several occasions has to take into account the limited capacity of the heterogeneous stakeholders (e. g., with respect to different processing times and the appropriateness of test difficulty).

Moreover, since panel studies are related and situated in a space-time continuum, structural changes in the school systems of the federal states must be taken into account, such as changing transitions between educational stages as well as different durations of school trajectories over time.

Beyond this, there are some real operative challenges that arise from the uniqueness of the NEPS panel design. For instance, the NEPS survey plan provides annual or bi-annual surveys in the school cohorts, which always have to be synchronized with the school calendars. One complication in this context is the fact that the German Federal States have different holiday schedules. This not only requires compliance with the surveys themselves, but also with the entire preparation of the surveys (i. e., the timing of status and contact information updates, tracking processes, approval procedures, etc.). Furthermore, the aim of surveying the same students repeatedly, organized in group settings, different circumstances, and conditions such as class repeating and class- and course divisions, has to be considered. This leads to a great effort in gathering the participating students on the test day.

- 1) To get panel data as soon as possible, the studies of *two school panels* start nearly at the same time. For organizational reasons, they are partially managed in the same schools. A total of 178 schools with either a fifth and/or a ninth grade were recruited. The class-based sampling used in the main studies implies less effort for schools than age-based sampling, which is—as seen in the pilot studies—a major reason for schools not to participate. Nevertheless, the design with mostly four classes of two grades at one school entails particular challenges for the instrument developers, the data collecting institutes, and not least for schools and participants.
- 2) Another feature of the NEPS is its *multilevel and multi-informant approach* to the study of the effects of different school structures and school reforms as well as the interactions between the individual and changing learning environments

(Blossfeld et al., 2011). This is done in a manifold way. On the one hand, annual surveys of the subject teachers and school principals of the participating students are taken. Therefore, the link between students and their teachers has to be verified with each wave, and the respective actors have to be recruited and administered. In addition to the corresponding survey instruments, appropriate procedures and referrals have to be developed and administered. On the other hand, context information is collected from the parents of the participating students. This is done by computer-administered telephone interviews (CATI), usually during each wave (see Section 1). Due to the peculiarities of the German school system, the survey instruments and procedures are also reviewed by the ministries of the federal states. With respect to the different requirements of the federal states, complex survey procedures as well as survey tools (e. g., federal-state-specific administration of the interview contents) have to be implemented.

- 3) As mentioned above, participation in all NEPS surveys is *voluntary at all times and for all levels*. Due to a greatly increased number of school surveys in recent years with the ongoing empirization of educational research, willingness to participate has continued to decrease. This effect was amplified by the negative evaluation reports of federal state-specific school systems in national and international assessment studies in recent decades. Therefore and due to the fact that extrinsic incentives (e. g., feedback concerning individual teaching and performance diagnostics, monetary incentivization, etc.) are limited for several reasons, the intrinsic motivation of all actors must be acquired as far as possible. Minimizing burdens and barriers with regard to participation is necessary to assure sufficient participation rates.

Due to these unique features, the NEPS design calls for a high flow of information and close cooperation and support at different levels, for example, with the ministries, schools, and participants. The following section provides a short overview of these processes and how all this affects and interacts with the accessing and recruitment of the NEPS school cohorts.

After the project was positively evaluated by the *German Research Foundation* (DFG) and was funded by the *Federal Ministry of Education and Research* (BMBF), the *Standing Conference of the Ministers of Education and Cultural Affairs of the Länder in the Federal Republic of Germany* (KMK) announced its strong support of the NEPS project. Besides this general statement, the KMK addressed a letter of support to schools in which it appealed for support for the project. Furthermore, the NEPS Project Management reports annually on the current project development, the survey plans, and upcoming data releases to the KMK.

Due to the federalism and as shown above, not only is the legal authority of the educational systems located in the German federal states, but so, too, is its structural configuration. For this reason, it is essential to integrate the ministries of education of the federal states in several processes. Starting with the sampling process of schools,

the ministries provide and update the sampling frames and offer data to frame features that are otherwise not accessible. In the process of school recruitment, in particular, the ministries offer assistance by sending letters of support and motivation, by preparing and advertising information events for schools, by sending recruitment and cover letters to schools, and by conducting school follow-ups.

One interesting element in this context is that the probability of school participation varies depending on the support of the ministries in the recruitment process. We see that compared with a missing or lesser engagement of the ministries, the participation rate of the schools is up to 30 % higher in the federal states in which the ministries send out the motivation letter to the schools and follow up on the sampled schools by themselves. A similar phenomenon can be identified for the liability of the feedback of the schools in the survey process whenever the ministries are involved. On the other hand, one can expect that the capacity of the ministries to actively foster school participation decreases with the number of schools that have to be recruited in their federal state. For this reason, special replacement strategies and parallel recruitment for the realization of the targeted sample were implemented. For each initially sampled school, four replacement schools were derived that could replace the possible dropout of the original school in a specified order (Aßmann et al., 2011). In the cohort of fifth graders, a total of 683 schools were contacted, 246 of which agreed to participate. Of these schools, 37 withdrew their initial commitment and were replaced. In the cohort of ninth graders, a total of 1,741 schools were contacted, 584 of which initially agreed to participate. 35 of these schools withdrew their initial commitment and were also replaced.

Besides the KMK, the ministries of education and the schools have to be integrated directly because their commitment and support is more-or-less directly crucial to the success of recruitment and gaining access. A central function, of course, lies with the school principals, coordinators, and teachers. They are indispensable gatekeepers at the meta- and individual levels in convincing relevant mediators, such as school conferences, parents' associations, and of course, the participants themselves. Furthermore, they are directly involved in various administrative processes of the surveys, such as organization and the provision of human and physical capacity (e.g., supervisors and rooms), when updating the status and contact information of the targets. Finally, their own participation is important as a basic source of context information. Therefore, involved school principals and teachers have to be thoroughly briefed as well as incentivized as far as possible to keep up their motivation and support of the NEPS.

In summary, the chosen access and survey strategies offer great possibilities for studying the developments and trajectories of the targets in an innovative and comprehensive manner. However, these strategies are also associated with special challenges and requirements that should be kept in mind. Therefore, the central coordination of the project in general, as well as the processes of accessing and administering the surveys, in particular, is both an essential and major task.

References

- Arbeitskreis Deutscher Markt- und Sozialforschungsinstitute e. V. (ADM). (1994). *Kernprobleme im Datenschutzrecht und Standesrecht der demoskopischen Umfrageforschung*. Retrieved from http://www.adm-ev.de/fileadmin/user_upload/PDFS/Kernprobleme_D.pdf
- Arbeitskreis Deutscher Markt- und Sozialforschungsinstitute e.V. (ADM), Arbeitsgemeinschaft Sozialwissenschaftlicher Institute e.V. (ASI), Berufsverband Deutscher Markt- und Sozialforscher e.V. (BVM), Deutsche Gesellschaft für Online-Forschung e.V. (DGOF). (2011). *Richtlinie zum Umgang mit Adressen in der Markt- und Sozialforschung*. Retrieved from http://rat-marktforschung.de/fileadmin/user_upload/pdf/R07_RDMS.pdf
- Aßmann, C., Steinhauer, H. W., Kiesl, H., Koch, S., Schönberger, B., Müller-Kuller, ... Blossfeld, H.-P. (2011). Sampling designs of the National Educational Panel Study: Challenges and solutions. In H.-P. Blossfeld, H.-G. Roßbach, & J. von Maurice (Eds.), *Zeitschrift für Erziehungswissenschaft, 14. Education as a lifelong process: The German National Educational Panel Study (NEPS)* (pp. 51–65). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Blossfeld, H.-P., Roßbach, H.-G., & von Maurice, J. (Eds.). (2011). Education as a lifelong process: The German National Educational Panel Study (NEPS) [Special issue]. *Zeitschrift für Erziehungswissenschaft, 14*. Wiesbaden: VS Verlag für Sozialwissenschaften.
- Blossfeld, H.-P., & Schneider, T. (2011). Data on educational processes: National and international comparisons. In H.-P. Blossfeld, H.-G. Roßbach, & J. von Maurice (Eds.), *Zeitschrift für Erziehungswissenschaft, 14. Education as a lifelong process: The German National Educational Panel Study (NEPS)* (pp. 35–50). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Blossfeld, H.-P., von Maurice, J., & Schneider, T. (2011). The National Educational Panel Study: Need, main features, and research potential. In H.-P. Blossfeld, H.-G. Roßbach, & J. von Maurice (Eds.), *Zeitschrift für Erziehungswissenschaft, 14. Education as a lifelong process: The German National Educational Panel Study (NEPS)* (pp. 5–17). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Häder, M. (2009). *Der Datenschutz in den Sozialwissenschaften: Anmerkungen zur Praxis sozialwissenschaftlicher Erhebungen und Datenverarbeitung in Deutschland*. (Working Paper No. 90). Berlin: Rat für Sozial- und Wirtschaftsdaten.
- Heydrich, J., Weinert, S., Nusser, L., Artelt, C., & Carstensen, C. H. (2013, September). Including students with special educational needs into large-scale assessments of competencies: Challenges and approaches within the German National Educational Panel Study (NEPS). *Journal for Educational Research Online, 5*(2). Retrieved from <http://www.j-e-r-o.com/index.php/jero/article/view/367>
- Iraschko-Luscher, S. (2006). Einwilligung—Ein stumpfes Schwert des Datenschutzes? *Datenschutz und Datensicherheit, 30*, 706–710. doi: 10.1007/s11623-006-0196-0

- Ludwig-Mayerhofer, W., Solga, H., Leuze, K., & Dombrowski, R. (2011). Data on educational processes: National and international comparison. In H.-P. Blossfeld, H.-G. Roßbach, & J. von Maurice (Eds.), *Zeitschrift für Erziehungswissenschaft, 14. Education as a lifelong process: The German National Educational Panel Study (NEPS)* (pp. 251–266). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Meixner, S., Schiller, D., von Maurice, J., & Engelhardt-Wölfler, H. (2011). Data protection issues in the National Educational Panel Study. In H.-P. Blossfeld, H.-G. Roßbach, & J. von Maurice (Eds.), *Zeitschrift für Erziehungswissenschaft, 14. Education as a lifelong process: The German National Educational Panel Study (NEPS)* (pp. 301–313). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Metschke, R., & Wellbrock, R. (2000). *Datenschutz in Wissenschaft und Forschung*. Berlin: Verwaltungsdruckerei Berlin.

About the authors

A. Müller-Kuller
Leibniz Institute for Educational Trajectories (LifBi), Bamberg.
e-mail: andre.mueller-kuller@lifbi.de

S. Meixner
University of Würzburg, Würzburg.

M. Sixt
Leibniz Institute for Educational Trajectories (LifBi), Bamberg.
e-mail: michaela.sixt@lifbi.de

Cooperation and Communication Within Scientific Organizations: The Role of Survey Coordination

Ina-Sophie Ristau and Stephanie Beyer

Abstract

This article describes challenges and practical implementations that occur as a result of the multi-cohort sequence design and the complex structure of the National Educational Panel Study (NEPS). Due to the different starting cohorts within the NEPS, there is a main differentiation in the sampling methods. Two of the six starting cohorts are samples in individual contexts (newborns, adults), whereas the other four are gathered through an institution-based approach (Kindergarten, school Grades 5 and 9, and university). Because of the heterogeneity and complexity of the samples, two surveying institutes are in charge of data collection. The main goal of the Survey Coordination Department is to link all involved sections and to keep the communication processes transparent. Therefore, we point out which practical solutions have been generated within the NEPS to face the challenges emerging from its aim to describe educational processes over the individual lifespan in a longitudinal design. The problem of complexity is discussed theoretically but is primarily considered from a practical perspective.

1 Introduction

In institutions of a certain size, the question of how to solve problems due to complexity arises. In general, this question is discussed in terms of issues of organizations within the field of organizational theory as well as in subfields of business administration. These approaches cover organizations in a very broad range, such as for-profit companies, international corporations, and non-profit organizations. However, complexity is rarely seen as an issue of scientific organizations. This fact is surprising because there has been an increase in the need for new and complex scientific networks in recent years, especially in the context of complex data collection and data-dissemination facilities. The increasing complexity of data structures poses new and diverse challenges to scientific organizations. Within this framework, scientists have

to cope with interdisciplinarity as well as cooperation with for-profit institutions (e. g., data-collection institutes) that often take part in scientific processes since these institutions have expertise in practical fieldwork. Therefore, communication and the reduction of complexity are important factors for a functioning and successful work process.

With this in mind, the following question arises: How can complexity within a scientific organization be managed? To answer this question, we must provide insight into a complex scientific institution: the National Educational Panel Study (NEPS). In order to shed light on this issue, (1) we present common definitions of coordination and discuss the problem of complexity within organizations. In a second step, (2) we demonstrate that this problem also occurs in the context of scientific institutions concerning multiple dimensions such as design, structure, sampling, and communication processes. In a third step, (3) the practical solutions to the problems of complexity in scientific organizations are examined using the example of the NEPS. We demonstrate how the Department of Survey Coordination has implemented various instruments and processes to solve problems arising from complex structures.

2 Complexity, Coordination, and Communication Within (Scientific) Organizations

There is a certain degree of complexity within organizations resulting from a range of alternatives of action and an increasing number of actors or employees. The issue of complexity is predominantly discussed within organizational studies and business administration but has never been treated as an issue of scientific organizations. Since science has to cope with issues of interdisciplinarity, projects often consist of scientists from different disciplinary cultures and backgrounds. Additionally, modern scientific research often requires the involvement of for-profit institutions (e. g., data-collection institutes), which operate under economic principles. Taking these different aspects into account, there is a growing need for coordination and communication management within scientific organizations. In this context, communication has been discussed as the essential and integrating element in organizations that secures their existence (cf. Oelert, 2003; Wolf, 2010). Without organized communication, organizations of a certain size would have to deal with the same problems over and over again as well as the loss of information and competencies. In the worst case, an uncoordinated organization might quickly collapse if it does not find a way to reduce complexity (cf. Luhmann, 1997). This is also an issue within scientific organizations. Cooperating with many institutions set in different locations is not only necessary but also valuable and productive and makes coordination even more important.

Following Malone and Crowston (1994), coordination can emerge in any kind of system, be it categorized as biological, human, or computational. Organizational theory suggests a very broad definition of this phenomenon and defines coordination as

“managing dependencies between activities” (Malone & Crowston, 1994, p. 90). Additionally, recent research has “provided insight into the microprocesses involved in coordinating [...] by shifting the analytic focus from coordinating mechanisms as reified standards, rules, and procedures to coordinating as a dynamic social practice” (Jarzabkowski, Lê, & Feldman, 2012, p. 907).

As shown above, the issues of coordination and communication processes have been discussed in other contexts but have been neglected when it comes to scientific organizations. Hence, we aim to shed light on this issue, especially in the context of the administration of longitudinal surveys in scientific organizations, and we therefore provide insight into the current work of the Survey Coordination Department of the NEPS to exemplify our coordination processes.

3 Complexity Within the NEPS and the Need for Coordination Mechanisms

There are several factors that contribute to the enormous organizational complexity and heterogeneity of a panel study. Parallelism of studies, heterogeneity of starting cohorts and pathways, interdisciplinarity, and the cooperation with institutes located all over Germany have to be taken into account, especially within the NEPS the longitudinal design.¹ The need for coordination management arises from three main points in particular: design, structure, and sampling.

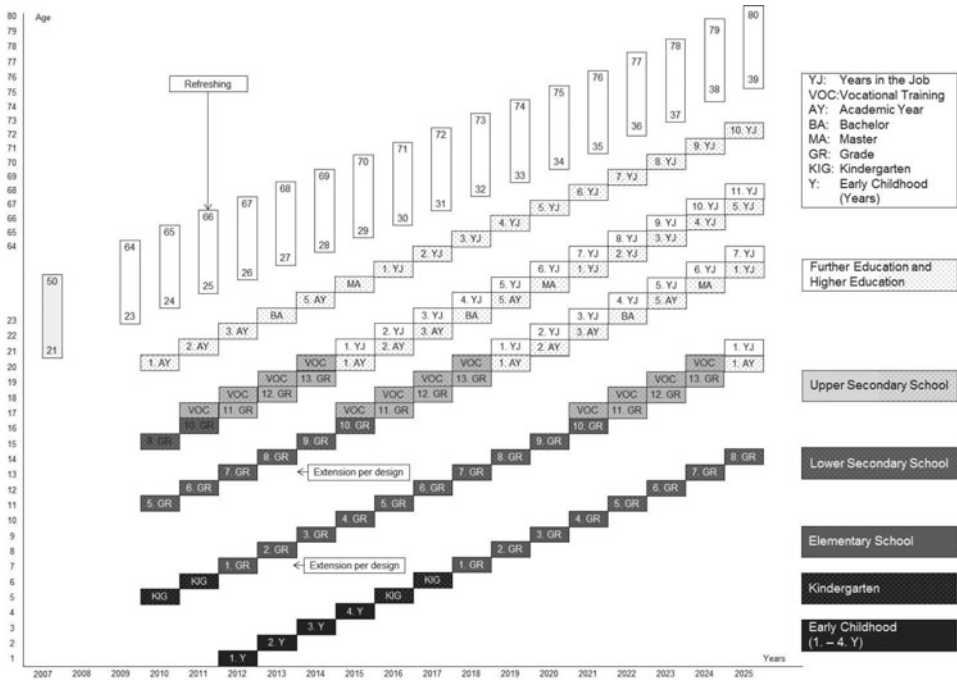
3.1 Design

The NEPS has been set up to find out more about educational processes and trajectories across the life span. It aims to explore how competencies develop over the life course and how family, peer-group, and institutional contexts influence competence development (cf. Blossfeld, Roßbach, & von Maurice, 2011). To answer this wide range of questions, the NEPS uses a longitudinal design with six starting cohorts (six panel studies, each one starting at a different stage in the life course). These cohorts cover the whole life span from birth until retirement and have been realized through a multi-cohort sequence design (cf. Figure 1), which allows for collecting and analyzing data in a life-course perspective. The different cohorts are connected to each other, especially if target persons move from one cohort to another during their participation within the panel. Therefore, it is crucially important that the processes and survey materials be consistent.

In this context, a large number of sub-studies are needed. On the whole, the NEPS realizes 165 quantitative sub-studies over a period of five years within its current fi-

¹ For further information, compare von Maurice, Blossfeld, & Roßbach in this volume.

Figure 1 The multi-cohort sequence design (2007–2025)



anced period from 2009–2013. Due to design aspects, these sub-studies are split into 72 main studies, 54 pilot studies, and 39 developmental studies that are subdivided among the six starting cohorts. Data resulting from main studies are provided to the scientific community as a Scientific Use File no more than 18 months after the fieldwork has taken place.²

3.2 Structure

Due to the heterogeneity of the design as well as the numerous research questions, it is essential that different disciplines and experts be combined in one team. To realize the variety of sub-studies in form and content, more than 200 scientists from all over Germany work within the NEPS. More than 90 researchers work at the Institute of Educational Research Bamberg (INBIL), where the project management and the Survey Coordination Department are also located. An interdisciplinary consortium

² For further information, see www.neps-data.de.

Figure 2 Geographical distribution of institutes and universities participating in the NEPS

of research institutes, research groups, and leading researchers round out the NEPS network (cf. Figure 2).

All in all, there are currently over 30 different research units, which are integrated in a framing concept following two underlying principles:

- 1) Educational biographies should be divided into eight stages with a particular focus on transitions from one stage to another.
- 2) Five major theoretical dimensions, the so-called “pillars,” should be captured.³

³ For further information, compare Blossfeld et al. (2011).

Embedded in this framework is the Survey Coordination Department as well as the Methods Department (cf. Blossfeld, von Maurice, & Schneider, 2011). For the realization of each sub-study, it is essential to work together with experienced data-collection institutes and an interdisciplinary team. Since the sampling strategy comprises individual and institutional sampling (Section 3.3), two data-collection institutes are appropriate to cover these two different settings. One of the involved institutes within the NEPS is the Institute for Applied Social Sciences (infas), which is an independent institute for social research. The institute's main expertise lies in the collection of data based on individual interviews and computer-assisted telephone interviews (CATI), in particular. The second institute is the International Association for the Evaluation of Educational Achievement Data Processing and Research Center (IEA DPC), which is specialized in research located in institutional settings, such as schools and Kindergartens.

3.3 Sampling

Apart from the design and structure, the sampling of a longitudinal study also has an effect on the degree of complexity. Within the sampling of the NEPS, a distinction is drawn between institutional and individual samples according to the context in which the target population is located. Therefore, the starting cohorts of the Kindergarten children, the fifth and ninth graders, and the first-year students are sampled in institutional contexts (Kindergartens, schools, and universities). The newborn and adult sample is drawn in an individual sampling frame with the admission of the registration office.⁴ As shown in Figure 1, the first sub-study of the NEPS took place in 2009 within the adult starting cohort, followed by the Kindergarten, school, and university starting cohorts in 2010. The newborns started as the last cohort in 2012. In each starting cohort, a plurality of sub-studies, a variable number of developmental studies, and usually one pilot study followed by one or two main studies are conducted per year.

3.4 Resulting Tasks for the Survey Coordination Department

Based on the above-mentioned complexity, a need for survey coordination and defined processes is obvious. The Survey Coordination Department faces the challenge of reducing the complexity of processes due to the multiplicity of cohorts and the parallelism of sub-studies. Therefore, it is of fundamental importance to create transparency as well as to link the involved parties of the NEPS-network, especially the experts of each starting cohort, the experts of the five pillars, and the data-collection

4 For further information, compare Assmann et al. (2011).

institutes. All interests, expertise, and participants can thus be led in a transparent communication process and can finally produce high-quality data, which are then made available to the scientific community.

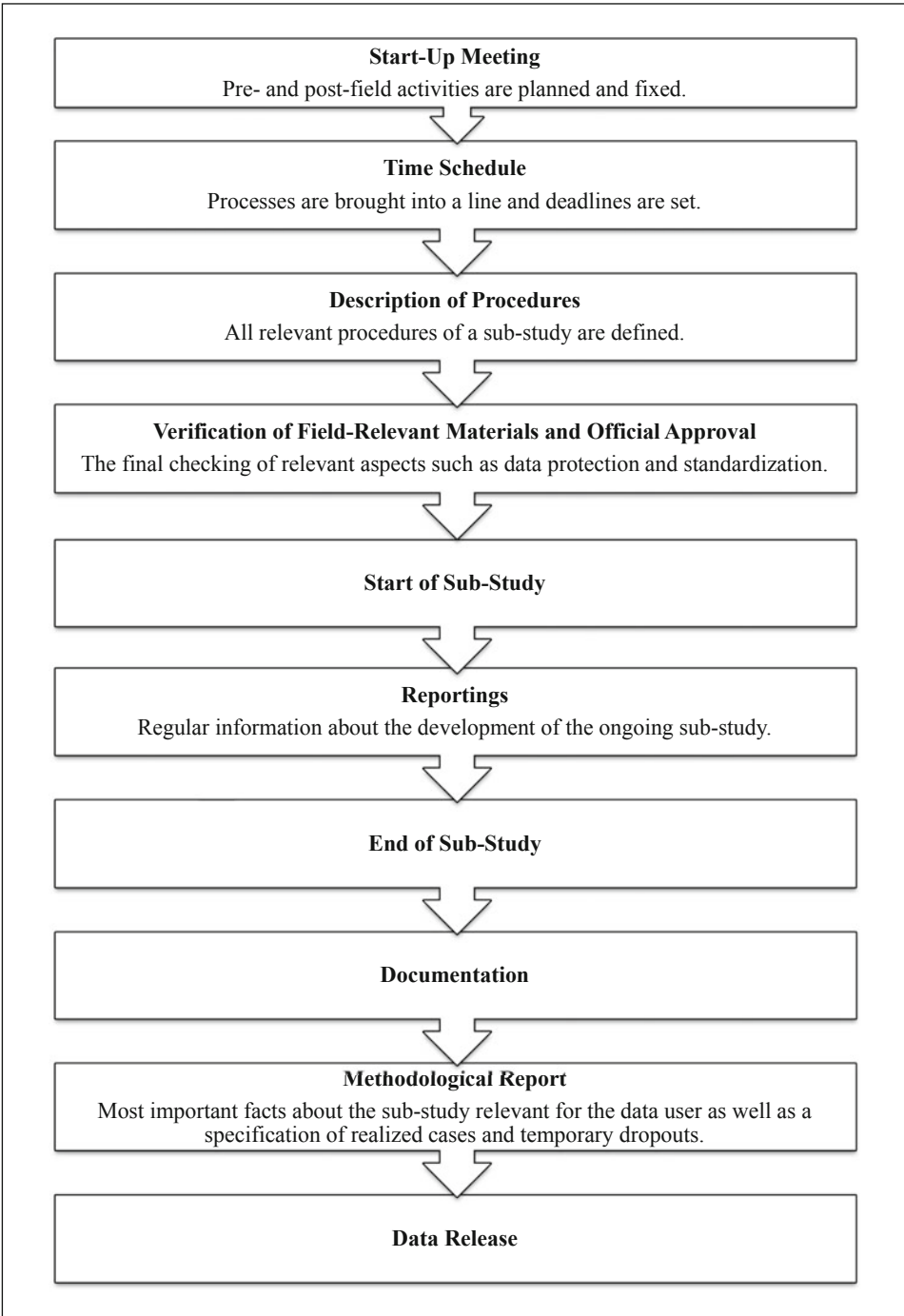
4 Coordinating a Longitudinal Survey—The Practical Experience of Scientific Coordination Within the NEPS

The Survey Coordination Department was implemented to handle the complexity discussed above, to avoid redundant working procedures, and to make certain processes transparent. Based on the description in Chapter Two, the department's task is to manage dependencies among activities as well as to function as an interface of different agencies, such as researchers, the Finance Department, and data-collection institutes. So far, the Survey Coordination Department has developed several procedures (cf. Figure 3) as practical solutions to reduce complexity within the NEPS. These solutions can be seen as a dynamic social practice constructed and reconstructed through the activities of coordinating as described by Jarzabkowski et al. (2012).

To begin with, one of the central documents of the Survey Coordination Department is the survey overview. This document comprises all planned sub-studies and includes information about the basic parameters of each study (e.g., sample size, starting cohort affiliation, field time, design aspects, etc.). It guarantees the coherence and successful realization of all sub-studies in adequate succession. At a more specific level, a *start-up meeting* is scheduled for each sub-study. It is held either as a face-to-face meeting or as a telephone conference and is to be scheduled ten months in advance with all relevant participants: the Survey Coordination Department, the responsible persons in charge of the sub-study administration, and the data-collection institute. Responsibilities, tasks, and fields of activities are assigned in this context. For instance, the head of the sub-study is chosen and given a fixed position. One of the main tasks of the head of a sub-study is to create a time schedule that includes every process needed to realize the sub-study and to bring this schedule into agreement with all other parties. Moreover, the communication of changes and delays concerning the timeline is important. Furthermore, the start-up meeting makes it possible to define, specify, or adjust the sample size and the design of the forthcoming sub-study⁵ (e.g., testing domains, rotation of instruments). Previous methodological reports serve as a basis for the meeting to provide sufficient orientation since studies are not independent but rather synchronized with one another. Furthermore, incentives and further materials (e.g., stopwatches, calculators) are planned. All discussed and approved issues are specified in a start-up form, which undergoes a review pro-

5 In most cases, the design and sample size are defined in the survey overview but need to be expatiated since there might have been changes in previous sub-studies as well as other major changes due to unintended effects.

Figure 3 Chain of main processes



cess by all participants after the meeting has taken place. After this process, the document is binding for all parties.

In a second step, the most important deadlines and follow-up processes are transferred into a *time schedule*, for which the head of the study takes responsibility. In cooperation with the Survey Coordination Department, this document is supervised and made accessible to all personnel within the NEPS in order to guarantee transparency. To keep track of everything and to keep everyone updated, both a time-schedule meeting with one representative from each stage and pillar and the Data Protection Department take place on a bi-weekly basis. Changes, delays, and different interests and information may be discussed under the supervision of the Survey Coordination Department.

As a next step, *descriptions of the procedures* and the sample are required before each sub-study is ready to begin. The description of the procedures contains the most important elements of the sub-study (e. g., sample size, realization in the field) as well as a subsumption of the sub-studies of the previous and following waves of the starting cohort. The longitudinal planning of the cohort is secured with this document.

Since the instruments (questionnaires) are supposed to go to different authorities before finalization, we implemented a final *verification of field-relevant materials*, which is followed by an *official approval*. The instruments are checked in different departments for data protection verification, test of scales, and test of programming. The data protection verification takes place at different stages of the development of an instrument. It is first checked at the beginning of the developmental process, at which point the Data Protection Center comments on all variables that might be used in the questionnaire (cf. Meixner, Schiller, von Maurice, & Engelhardt-Wölfler, 2011).

In two of the six starting cohorts, one of the most significant steps before the field period can begin is the approval of the ministries of education of all 16 Federal States of Germany.⁶ Since the sub-studies within these starting cohorts take place in German schools, the 16 federal states, which are responsible for their school systems (due to cultural sovereignty), are involved in the process (cf. von Maurice, Sixt, & Blossfeld, 2011). As a result, all school-context-based sub-studies underlie an additional review in each federal state.⁷ All questionnaires and documents that are intended to be used in the sub-study have to be submitted to the ministries of education. However, before this submission takes place, the Survey Coordination Department has to check all documents according to corporate design aspects (paper-and-pencil-based interviews and all materials that are sent to the target persons, e. g., cover letters), verification of standardization of scales, and the observance of the instrument length. This

6 This step is only required for studies taking place in schools because schools in Germany are governed by each federal state. Therefore, an approval of instruments and procedures is needed by every federal state to gain access to the institutions. The materials for the other cohorts (early childhood, Kindergarten, students, and adults) do not need to be handed in.

7 For further information, compare Müller-Kuller, Meixner, & Sixt in this volume.

means that the submission of all documents to the ministries can take place after the approval of the Survey Coordination Department. After the approval of the Ministries (only in the school cohorts), the instruments and documents are finalized. In the context of the verification of field-relevant factors, the *official approval* of the final instruments and documents is given by the head of the sub-study as well as by the Survey Coordination Department. Once the materials have been approved, everything is set and the field phase can begin.

During the field phase, *reportings* are sent regularly by the data-collection institutes in order to inform these institutes about the field progress, such as the number of realized interviews. Furthermore, the Survey Coordination Department is responsible for the *documentation* of the material of each sub-study. Therefore, all documents that have been disposed in the field are systemized and archived. This is done electronically as well as on paper to make an accessible filing system available.

The next step in the timeline is the submission of a *methodological report* sent by the responsible data-collection institute several weeks after the survey-phase has ended. Within this report, the data-collection institute gives an overview of the conducted fieldwork and important facts to be able to understand the data-collection process ahead of and within the field. This report provides information about the design of the study, the sample, field instruments, realization in the field, the interviewers, selectivity, and weighting. Problems are to be disclosed and discussed to give the data user the optimal background to be able to understand and use the data accurately. The information in the methodological report is also used for the preparation of the subsequent sub-study within the specific cohort. At this point, the process ends with a final meeting with all involved parties in which the results and experiences of the sub-study are discussed. These outcomes are to be considered in the start-up meeting of the follow-up study. Finally, the data are released no later than 18 months after the completion of the field phase.

5 Summary

With this article, we are able to demonstrate that complexity within scientific organizations such as the NEPS is an important topic and that it is obvious that reducing complexity with coordination and communication mechanisms is essential. Due to the structure, design, and sampling frame, different kinds of dependencies occur and need to be managed by coordination processes as discussed by Malone and Crowston (1994). Within the NEPS, the Survey Coordination Department manages these complex structures with various implemented processes and structures that we have described in order to provide a first example of scientific organization coordination. Our practical experience has shown that the management of dependencies and the identification of processes result in a transparent, well-organized, and successful process of work in an interdisciplinary network. Therefore, every participating party is

informed and involved with all embedded procedures due to the timeline. With a centrally managed Survey Coordination Department, crucial processes can be over-viewed and complexity can be reduced. Consistency is the important factor to successfully run a panel study with six parallel starting cohorts. All in all, we have been able to demonstrate that within a scientific organization such as the NEPS, structured and transparent processes lead to a more effective realization of sub-studies and support scientific work by reducing this work's affiliation with bureaucratic red tape.

References

- Assmann, C., Steinhauer, H. W., Kiesl, H., Koch, S., Schönberger, B., Müller-Kuller, A., & Blossfeld, H.-P. (2011). Sampling designs of the National Educational Panel Study: Challenges and solutions. In H.-P. Blossfeld, H.-G. Roßbach, & J. von Maurice (Eds.), *Zeitschrift für Erziehungswissenschaft, 14. Education as a lifelong process: The German National Educational Panel Study (NEPS)* (pp. 51–65). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Blossfeld, H.-P., Roßbach, H.-G., & von Maurice, J. (2011). Education as a Lifelong Process: The German National Educational Panel Study (NEPS) [Special Issue]. *Zeitschrift für Erziehungswissenschaft, 14*. Wiesbaden: VS Verlag für Sozialwissenschaften.
- Blossfeld, H.-P., von Maurice, J., & Schneider, T. (2011). The National Educational Panel Study: Need, main features, and research potential. In H.-P. Blossfeld, H.-G. Roßbach, & J. von Maurice (Eds.), *Zeitschrift für Erziehungswissenschaft, 14. Education as a lifelong process: The German National Educational Panel Study (NEPS)* (pp. 5–17). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Jarzabkowski, P. A., Lê, J. K., & Feldman, M. S. (2012). Toward a theory of coordinating: Creating coordinating mechanisms in practice. *Organization Science, 23*, 907–927. doi:10.1287/orsc.1110.0693
- Luhmann, N. (1997). *Die Gesellschaft der Gesellschaft*. Frankfurt am Main: Suhrkamp.
- Malone, T. W., & Crowston K. (1994). The interdisciplinary study of coordination. *ACM Computing Surveys, 25*, 87–119. doi: 10.1145/174666.174668
- Meixner, S., Schiller, D., von Maurice J., & Engelhardt-Wölfler, H. (2011). Data protection issues in the National Educational Panel Study. In H.-P. Blossfeld, H.-G. Roßbach, & J. von Maurice (Eds.), *Zeitschrift für Erziehungswissenschaft, 14. Education as a lifelong process: The German National Educational Panel Study (NEPS)* (pp. 283–299). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Oelert, J. (2003). *Internes Kommunikationsmanagement: Rahmenfaktoren, Gestaltungsansätze und Aufgabenfelder*. Wiesbaden: Deutscher Universitätsverlag.
- von Maurice, J., Sixt, M., & Blossfeld, H.-P. (2011). *The German National Educational Panel Study: Surveying a cohort of 9th graders in Germany*. (NEPS Working Paper No. 3). Bamberg: University of Bamberg, National Educational Panel Study. Retrieved from https://www.neps-data.de/Portals/0/Working%20Papers/WP_III.pdf

Wolf, G. (2010). *Der Business Discourse. Effizienz und Effektivität der unternehmensinternen Kommunikation*. Wiesbaden: Gabler.

About the authors

I. S. Ristau
Leibniz Institute for Educational Trajectories (LifBi),
Wilhelmsplatz 3, 96047 Bamberg, Germany.
e-mail: ina.ristau@lifbi.de

S. Beyer
Chair of Sociology, esp. Sociological Theory,
University of Bamberg,
Feldkirchenstr. 21, 96045 Bamberg, Germany.

The Concept of Individual Retracking in NEPS— Approach, Practice, and First Empirical Evidence From Starting Cohorts 3 and 4

Michaela Sixt, Martin Goy and Georg Besuch

Abstract

For panel studies like the National Educational Panel Study (NEPS), it is of vital importance to keep the respondents on board and gather information over the life course in a consistent way. In the school cohorts of the NEPS, tests and questionnaires are administered in groups at school. As long as the respondents visit the schools where the NEPS is conducted, it is comparatively easy to reach these respondents and to keep them in the panel. However, if a respondent leaves an NEPS school due to changing schools, or if a school cancels its participation in the study, a different approach must be found to maintain contact with this special group of respondents and to continue collecting data from this group in a way that is comparable with the main field survey. For this reason, a concept of surveying these respondents in an individualized way was developed by the NEPS. In this article, we introduce the concept of individual retracking applied in and planned for the school cohorts of the NEPS, and we provide insight into the practice and challenges of this kind of data collection. We begin by introducing individual retracking as part of the aims and scope of the NEPS to survey not only mainstream but also nonstandard careers and individual pathways over the life course. Based on a review of the research literature on the designs and applications of individual retracking in longitudinal studies on educational processes, we introduce the approach taken by the NEPS in terms of its theoretical concept and its survey practice and present first empirical evidence on the basic sample structure and the response rates in the individually retracked survey as compared with the main field survey. We conclude the article with an outlook on the next steps to also introduce individual retracking in the NEPS in the contexts of elementary education and higher secondary education.

1 Introduction

The main aim of the National Educational Panel Study (NEPS) is to collect high-quality and comparable data on competence development and educational pathways in Germany over the whole life course and to make this data accessible to the scientific community. To do so, the NEPS develops theoretically and empirically based test and survey instruments built on two conceptual principles: (1) Educational biographies are divided into eight educational stages to allow for a stage-specific view of the particular situations and trajectories within a specific stage as well as the crucial transitions between them. (2) To assure a consistent measurement of theoretical constructs of high importance in educational research over the life course, the NEPS is based on five pillars, which focus on competence development, learning environments, educational decisions, migration, and returns to education. To be able to offer information on educational pathways over the life course as opposed to only at the end of the life course, the NEPS consists of six panel studies arranged in a multicohort sequence design (Blossfeld, von Maurice, & Schneider 2011, pp. 13 f.).

Two of these panels represent populations of students in schools: the starting cohort of fifth graders and the starting cohort of ninth graders. For both cohorts, the NEPS sampled randomly selected schools in all Federal States of Germany and requested the schools' participation and consecutively the participation of all students in two randomly selected classes in Grade 5 and Grade 9 (Aßmann et al., 2011). NEPS Stage 4 follows the students on their way through lower secondary education (Grades 5 to 10) up to and including their transition to upper secondary education, at which point NEPS Stage 5 takes over those students transferring into the academic track (Grade 11 to 12 or 13), and NEPS Stage 6 takes over those transferring into the vocational track.

In both cohorts, the main surveys started in fall 2010 with testing competencies and surveying the students with paper-and-pencil instruments in school. Furthermore, questionnaires were administered to class teachers to gain information on the class as well as on the quality of instruction for the German and mathematics teachers. In addition, the school principals were asked via a paper-and-pencil questionnaire to detail information about the school context (Frahm et al., 2011). To complement the survey with information on the home contexts of the students, the NEPS surveyed the students' parents (one parent per student) via telephone interviews about the context at home including, for instance, the social origin of the families.¹ Surveying context persons assures drawing a fuller picture of the social and learning environments of the students. Furthermore, regional information is available for the schools and the homes so that metadata on regional and local levels down to the fam-

1 The surveys in the institutional context were administered by IEA Data Processing and Research Center (DPC), Hamburg. The surveys in the individual field of the parents were administered by the Institute for Applied Social Sciences (infas), Bonn.

ilies' neighborhoods can be merged. Thus, not only a multicohort, but also a multi-informant and multilevel perspective can be realized with the design of the NEPS school cohorts.

In contrast to the younger NEPS cohorts, however, the participation of the students and their parents is decoupled in the school cohorts. As pilot studies reveal, coupling the participation leads to lower recruitment rates in these cohorts. As participation in the NEPS is voluntary, the parents and, in the case of the ninth graders, the students themselves have to agree to their continued participation in the panel. Of course, the schools are the first to agree to participate and stay in the panel.

For panel studies like the NEPS and its school cohorts with voluntary participation, incentive strategies and a proper concept of panel care are central to recruiting respondents and ensuring their continued participation in the study. In the NEPS, the students in school receive monetary incentives (€5 until Grade 8, €10 in higher grades). The teachers and principals who fill out a questionnaire and the teachers who coordinate the survey at school and cooperate with the data-collecting institute receive small presents show appreciation for their engagement. Furthermore, the NEPS puts a great deal of effort into writing motivation letters and providing information material, such as newsletters, informational brochures, and flyers for schools, parents, and students.

In both NEPS school cohorts, tests and questionnaires are administered in groups at school, and the contact to the target persons is organized via the school. As long as the respondents visit schools that participate in the NEPS, it is comparatively easy to reach the respondents and to keep them in the panel. However, if a respondent leaves the NEPS school because he or she has changed schools or if his or her school cancels its participation in the study, a different approach must be found to stay in contact with this special group of respondents and to collect data in a way that is comparable with the main field survey. Therefore, a concept for surveying these respondents in an individualized manner was developed by the NEPS: the field of individual retracking. With this individual field in the school cohorts, the NEPS is able to survey nonstandard careers and individual pathways over the life course in addition to the more mainstream or standard ways through schools surveyed in the main field of the panel study.

Based on a review of research literature and applications on designs and results of individual retracking in previous longitudinal studies on educational processes, in the following section, we introduce the approach taken by the NEPS in terms of its theoretical concept and its survey practice and present first empirical evidence on the basic structure and the response rates in the individually retracked survey in comparison with the main field survey. We conclude the article with an outlook on the next steps to also introduce individual retracking in the contexts of elementary education and higher secondary education.

2 Individual Retracking in Longitudinal Studies on Educational Processes

In the following section, we present a brief overview of a review of research conducted to identify designs and applications of individual retracking in longitudinal studies on educational processes. This review was conducted to investigate if and how other longitudinal studies apply individual retracking with regard to concepts, methods, and feasibility.

In line with the focus of this article on the school cohorts of the NEPS, we limited the scope of our investigation to longitudinal studies in elementary, lower secondary, and/or upper secondary school. We chose to include only studies conducted in Germany for reasons of comparability and access to the study documentation. Regarding the concept of individual retracking, we consider only those studies to apply this approach that survey respondents longitudinally in an individualized way *alongside* a longitudinal main field survey in school context.

To identify relevant longitudinal studies in Germany, we used the overview provided by Blossfeld and Schneider (2011) as a vantage point for this review of literature. In their synopsis of national and international longitudinal studies on education, Blossfeld and Schneider list 29 available longitudinal studies conducted in school contexts in Germany: four studies covering education from preschool age onward, 15 studies focusing on development and decisions in general schools, as well as 10 studies focusing on transitions from school to vocational training, to university, or to the labor market (Blossfeld & Schneider, 2011, pp. 38–43).

We complemented this list with a database search on research literature for the timespan from 2009 to 2014 to investigate if additional longitudinal studies in Germany that potentially apply individual retracking could be identified. Additionally, we contacted researchers involved in conducting current longitudinal studies in school contexts for which documentation might not yet be available.

Of the studies listed by Blossfeld and Schneider (2011), only one study could be identified that employs individual retracking according to our definition: the Study *Bildungsprozesse, Kompetenzentwicklung und Selektionsentscheidungen im Vorschul- und Schulalter* (BiKS, von Maurice, Artelt, Blossfeld, Faust, Roßbach, & Weinert, 2007; Mudiappa & Artelt, 2014) conducted in the Federal States of Bavaria and Hesse. In the first of the two longitudinal BiKS studies, which follows students from three years onward from Kindergarten into elementary school, those students who attend elementary schools in the initial sample, in which there were too few other study participants who attended the same schools or where the schools then cease their participation, are individually retracked and tested individually at home (Faust, Kratzmann, & Wehner, 2013, p. 36; Homuth, Mann, Schmitt, & Mudiappa, 2014, p. 21; Schmidt, Schmitt, & Smidt, 2009, p. 7). In the second BiKS study, which follows students from 8 years onwards in elementary school and in their transition from elementary to secondary school, individual retracking is also applied. If, after the transition from el-

elementary to secondary school in the course of the study, students of the second BiKS sample were not surveyed within the secondary schools for four reasons (that is, the target persons started attending schools in a region not covered by the study; there were less than three study participants in total in their schools; there was no information available on the schools they attended after transition from primary school; or if their school ceased participation in the study), students were then followed with individual retracking, and were surveyed with questionnaires sent to their homes, but they no longer participated in any competence tests (Homuth et al., p. 22; Lorenz, Schmitt, Lehl, Mudiappa, & Roßbach, 2013, p. 27; Schmidt et al., 2009, pp. 9–10).

In addition to BiKS, our review identified only one further study that applies individual retracking: the BERLIN study (Maaz, Baumert, Neumann, Becker, & Dumont, 2013). This study uses a research perspective to follow the transformation in school structure in the Federal State of Berlin, where the former four-tier school system of lower secondary education was switched to a two-tier system beginning in the school year 2010/2011. The study has two levels: Level I of the BERLIN study focuses on the impact that the change in system conditions has on the transition of students from elementary to lower secondary school and their pathways through lower secondary school (Module 1). Level II investigates the implications that the restructuring of secondary school has by comparing two student cohorts starting in Grade 9 in their transition to Grade 12: One cohort (Module 2) continues tracking the students of Module 1 as part of a larger, representative cohort that fully traverses the reformed secondary education; a second cohort (Module 3) serves as a control group—these students traverse through secondary education in the school system prior to restructuring. Individual retracking is used in Level II of the study. Those students who have been assessed at the first point of measurement in Grade 9 in either Module 2 or Module 3 and who left school after Grade 9 prior to the second point of measurement in Grade 10 are individually retracked and surveyed with a questionnaire sent to their homes. All students who left school prior to the third and final point of measurement in Grade 12 are also individually retracked and surveyed with a final questionnaire sent to their homes (Maaz, Baumert, Neumann, Becker, Kropf, & Dumont, 2013, p. 39–42).

3 NEPS Concept of Individual Retracking

As the overview in Section 2 shows, individual retracking is sparsely used in German longitudinal studies in school contexts. Consequently, there was not much empirical evidence from other studies available when the NEPS started to implement individual retracking. As mentioned above, the objective of individual retracking in the NEPS is to stay in contact with students who left NEPS schools or whose school quit participation in the NEPS. By this measure, the NEPS maintains the possibility of surveying and testing these persons in later surveys. Furthermore, keeping the individual

contact serves the purpose of collecting current data that are comparable with data from the main field. In this section, we outline the concept of individual retracking applied in the NEPS in the starting cohorts of the fifth and ninth graders, we present the materials implemented, and we detail the standard procedure of an individual survey and the experiences with this concept in the surveys for which data is already available.

The present design of individual retracking that has been developed in the last years satisfies the requirements of common mail surveys (Porst, 2001). In short, these are sincere and informative cover letters, which underline the importance and confidentiality of survey participation; preaddressed and freepost return envelopes; short survey instruments; multiple contacts; as well as incentives offered to show appreciation of participation.

To meet the requirements of a panel study, the concept of individual retracking in school cohorts comprises standardized procedures and several instruments applied in every individual survey. Essential for the NEPS, and especially for individual retracking, is the availability of unambiguous status information regarding the target person, that is, the information of whether a student still attends an NEPS school or if he or she has left this institution. The assignment of the group to which a student belongs is surveyed by so-called "school update lists" sent out in advance of each field start. According to these lists, each student is classified as a part of either the main field or the individual field. Besides the correct classification, continuous contact with the target person is of vital importance. Cohort-specific surveys in the field of individual retracking parallel to the main field are intended to gather comparable data. These surveys include the send-out of three different survey instruments. Two of these instruments are also applied in the main field: first, a short questionnaire to track the current address of the respondent, and second, a questionnaire for students that provides comparable data to that of the main field. The third instrument is a very short questionnaire developed especially for the requirements of individual retracking and is exclusively applied in this field because this information is available from the schools in the main field. This update questionnaire tracks the current status of the respondent, for example, whether the respondent still attends school or has already left school for some kind of vocational training, what kind of school or training he or she attends, the location of the school, and the class the student is visiting. The update questionnaires of the cohorts are very similar but have cohort-specific adjustments regarding the status range. Therefore, the update questionnaire pursues the same task as the school update list in the main field: classifying the status of the respondents.

In conclusion, the transition of a student from the main track to the individual track is not just a transition in administrative terms but also a transition of the survey context. These students need to know that they are still part of the NEPS sample even though their mode of participation has changed. They need information about their new status, especially in case of the students' first individual survey, and the im-

plications of this transition for future NEPS surveys. Moreover, the parents of these students need to receive this information as well. To address these aspects, each individual survey contains a cover letter for students and an additional cover letter for parents. Furthermore, a short informational brochure with general study information is included in analogy to the main field.

Before the field work of the individual field begins, the students in this field have to be identified. This information is provided by the above-mentioned school update lists.² Based on these, a list of student IDs in the individual field is processed. For these IDs, the corresponding student and parent addresses are provided. Students with valid addresses are contacted two weeks after the corresponding main field phase has begun. Every student gets a student questionnaire, a status update questionnaire, a short address questionnaire, and an information letter for parents and students. Cases with missing or invalid addresses get the status “temporary dropout.” Ideally, these students will be contacted in the next survey at their new address. If material could not be forwarded due to the relocation of the target persons, the questionnaires are resent if the postal service imprints the new address on the envelope. For target persons without an address memo, an address tracing procedure is installed.³ If new addresses can be investigated, the send-out process starts with a delay of several weeks for these cases. If this tracing is not successful, the students are allocated the status of temporary dropout and will be contacted again in the following survey.

We know from other studies that the application of a reminder increases the response rate substantially. Hence, we decided to send out a mail reminder if there is no response two weeks after the first contact. The reminder consists of a modified cover letter for students, a copy of the short address questionnaire, and the update questionnaire. We decided not to send the questionnaire a second time to lower the burden for the respondent (providing the same incentive). There are no multiple-reminder send-outs. Reminder nonrespondents are treated as temporary dropouts and are be contacted in the next survey.

Analogous to the main field, the respondents obtain a monetary incentive and a letter of thanks if they send back one of the requested instruments. Afterwards, the short address questionnaires are forwarded to the Institute for Applied Social Sciences (infas).⁴ The data of the returned student and update questionnaires are processed by the IEA Data Processing and Research Center (DPC) and transferred to the NEPS Data Center.

2 On these lists, each target person in the school has an identification number and a status code for the survey context.

3 The new addresses are acquired through telephone interviews with the parents or via address tracing.

4 Due to data protection obligations, nonanonymous data and survey administration are institutionally separated. Names and addresses of target persons are administrated and provided by the infas.

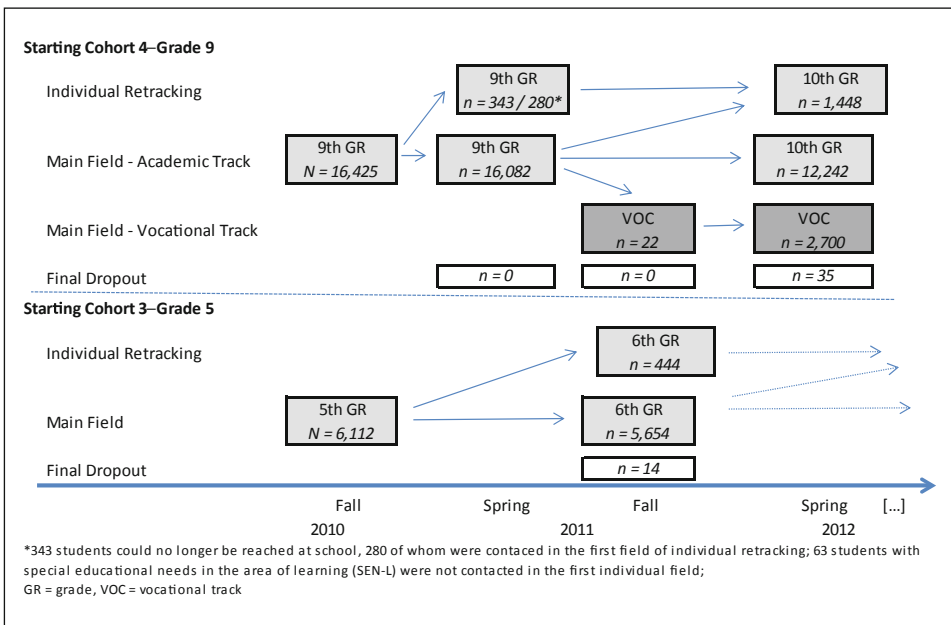
4 Empirical Evidence

In this section, we present first empirical evidence on the basic structure and the response rates in the individually retracked survey and the main field survey. First, we detail the design of the two school cohorts, their panel structure, and the different survey fields. Then, we shed a light on the reasons for the change to the field of individual retracking and present selected, basic information on the structural make-up of the subsamples. On this basis, we compare response rates in the field of individual retracking of both NEPS school cohorts with the respective main field.

4.1 Panel Structure and Survey Fields

The first field of individual retracking began in spring 2011 (cf. Figure 1). At that point, the starting cohort of the ninth graders was surveyed a second time. As Figure 1 shows, the sample of starting cohort Grade 9 consists of 16,425 students, 16,082 of which could still be contacted in the main field in school in spring 2010. 343 students (2 %) dropped out of the main field and thus switched to the field of individual retracking. As expected, this is a rather small group because only six months passed

Figure 1 Survey fields in Starting Cohort 4—Grade 9 and Starting Cohort 3—Grade 5



by and the survey began in the same school year so that only minor changes in school career were to be expected.

The second wave of individual retracking in the cohort of the ninth graders (by then, the target persons were attending Grade 10) started in spring 2012. Unfortunately, we cannot use data from the second field of individual retracking in spring 2012 because the edition of this data has not been finished at time of writing this article. Preliminary data show that there were 1,448 students in the field of individual retracking in spring 2012.⁵

For the starting cohort of fifth graders, the first wave with individual retracking was in fall 2011. Out of the 6,112 students in this cohort, 444 students (7%) attending the sixth grade then could not be reached in an NEPS school.

4.2 Reasons for Individual Retracking

As described in the beginning of this chapter, there could be individual and school-based reasons why we could no longer reach the participants at school. Individual reasons could be a removal or a planned change of school if the tracks offered by the school do not fit with individual interests.⁶ School-based reasons appear if the school withdraws its willingness to participate in the NEPS, if a school ceases offering the respective grade level, or if a school is closed. Furthermore, it could be that there are too few participants for continued participation of the school in the panel study.

As Table 1 shows, in spring 2011 for Starting Cohort 4, we find that 47% of the participants belonging to the field of individual retracking had changed schools. More than half of the students in this group changed schools because of school-based reasons: 53% of the individually contacted students left the main field because their schools quit their participation in the NEPS.

In Starting Cohort 3, the reasons for a change into the field of individual retracking are comparably distributed for the ninth graders: 48% of the individually contacted respondents had changed schools, and for 46% of them, the school cancelled its participation. In another 1% of cases, the school closed down, and in the case of 5%, there were too few participants at the school level, which meant that the NEPS was no longer testing at this school.

5 As described in Figure 1 and already mentioned in Section 1, it is possible in some Federal States to leave school and change to a vocational track after Grade 9. In this case, the NEPS starts a complete individual field in which the participants are contacted by telephone interviews and tested every two years at home (Ludwig-Mayerhofer et al., 2011).

6 In single cases, it is also possible that a child has to leave school because of insufficient grades or inadmissible behavior.

Table 1 Reasons for Changing Into the Field of Individual Retracking

Starting Cohort	4—Grade 9		3—Grade 5	
	Spring 2011		Fall 2011	
Change of school	160	47 %	213	48 %
School withdraws willingness to participate	183	53 %	204	46 %
School was closed	–	–	5	1 %
School ceases participation (number of participants at school level too low)	–	–	22	5 %
Total	343	100 %	444	100 %

4.3 Basic Structure of the Subsamples in Starting Cohort 4—Grade 9

To describe the basic structure of the subsample in Starting Cohort 4—Grade 9, this paper uses data from the National Educational Panel Study (NEPS): Starting Cohort 4—9th Grade, doi:10.5157/NEPS:SC4:1.0.0. Regarding basic socio-demographic information (cf. Table 2), such as sex and age, we have nearly an equal distribution and find no significant differences between the main (51 % male; 49 % female, average year of birth: 1995 [standard deviation = 0.7]) and the individual field (53 % male; 47 % female, average year of birth: 1995 [standard deviation = 0.8]).

Regarding migration background, we find significant ($p < 0.001$) differences between students in the main field and those in the individual field: In addition to the fact that we have less information available in the individual field, the proportion of participants who have migrated themselves (9 %) or who have at least one parent who migrated (36 %) is higher than in the main field (6 % and 26 %, respectively). We also find interesting differences between the two groups in the field of the individual retracking: While the group with individual reasons seems to be very similar to the group in the main field (6 % and 26 %, respectively), in the group with school-based reasons, the proportion of participants with a migration background is nearly twice as high (11 % and 44 %, respectively; $p < 0.001$).

Looking at the educational background, we first find significant ($p < 0.01$) differences in the proportion of parental information between the main and the individual field. First, in the main field, we have parent interviews for 56 % of the participants, and in the individual field, we only have parent interviews for 46 % of the participants. The two fields in individual retracking do not differ in this respect (46 % and 45 % with parent interviews). Second, we find that the educational background of target persons in the main field differs from those in the individual field ($p < 0.05$): In the field of individual retracking, less information is available (6 % compared with 4 % in the main field), and with 10 % of higher-educated parents, this share is lower than in the main field (18 %). We also find that those in the individual field are more similar

Table 2 Starting Cohort 4—Grade 9: Basic Structure of the Samples (Spring 2011)

Variables	Main field	Individual retracking		
	Total (<i>n</i> = 16,060)	Individual reasons (<i>n</i> = 160)	School-based reasons (<i>n</i> = 183)	Total (<i>n</i> = 343)
Sex				
Male	51 %	48 %	57 %	53 %
Female	49 %	52 %	43 %	47 %
Year of birth				
No information	0 %	1 %	1 %	1 %
Valid information	100 %	99 %	99 %	99 %
Mean (std. dev.)	1995 (0.7)	1995 (0.7)	1995 (0.8)	1995 (0.8)
Median	1995	1995	1995	1995
Min	1990	1993	1993	1993
Max	1999	1997	1996	1996
Migration background (first generation)				
No information	2 %	6 %	7 %	6 %
No	92 %	88 %	82 %	85 %
Yes	6 %	6 %	11 %	9 %
Migration background (second generation)				
No information	2 %	7 %	6 %	6 %
No	72 %	67 %	50 %	58 %
Yes	26 %	26 %	44 %	36 %
Parent interview				
No parent interview	44 %	54 %	55 %	54 %
Parent interview	56 %	46 %	45 %	46 %
Education of the parents				
No information	4 %	8 %	3 %	6 %
No higher education	78 %	74 %	93 %	84 %
Higher education	18 %	18 %	4 %	10 %
School track (first wave fall 2010)				
Hauptschule	23 %	18 %	23 %	21 %
Realschule/Gesamtschule	38 %	33 %	30 %	31 %
Gymnasium	32 %	40 %	20 %	29 %
Förderschule	7 %	9 %	27 %	19 %

Note. The difference to the total sample of *N* = 16,425 can be explained by *n* = 22 students who changed to the vocational educational system.

to the main field because of individual reasons (8% no information; 18% higher education; no significant difference to the main field) than are those with school-based reasons (3% no information, 4% higher education; $p < 0.001$ in comparison to the main field).

With regard to the visited school track in fall 2010, Table 2 reveals that most of the participants in the main field (38%) visited a type of middle school (e.g., Realschule, Gesamtschule, Schulen mit mehreren Bildungsgängen), about one third (32%) visited a Gymnasium, and about one fourth visited a Hauptschule (23%). An additional 7% visited a Förderschule, which is a school for students with special educational needs in the area of learning (SEN-L).⁷ For the individual field, this distribution differs significantly ($p < 0.001$), especially when regarding the 19% proportion of participants in Förderschule. Furthermore, the proportion of students in middle schools, Hauptschule, and Gymnasium is 7 (31%), 2 (21%), and 3 percentage points (29%) lower than in the main field, respectively.

Taking a closer look, we find that the proportion of students who attend a Förderschule is, with a share of 27%, considerably higher than in the group with individual reasons (9%) and in the main field (7%), especially in the group with school-based reasons for switching to the individual field. Furthermore, this differentiated picture shows that the proportion of students who attend a Gymnasium is higher in the field of individual retracking with individual reasons (40%) than with school-based reasons (20%) or in the main field (32%).

4.4 Basic Structure of the Subsamples in Starting Cohort 3—Grade 5

As the data edition for the Second Wave in fall 2011 was not finished at the time of composing this article, we can unfortunately not use data from an SUF for Starting Cohort 3—Grade 5.⁸ However, as these methodological analyses are important to assure a high quality of the data, we could take a look at the respective data the NEPS received from the data-collecting institutes. It is important to notice, though, that the following findings are preliminary and need to be confirmed by future analyses with the respective SUF.

As Table 3 shows, we find no relevant differences with regard to sex and year of birth between the main field and the individual field. In both fields, sex is nearly equally distributed (52% male students in the main field, 54% male students in the individual field), and the average year of birth is 1999 (std. dev. 0.6 resp. 0.8).

7 The NEPS is conducting a feasible study to investigate whether students with special educational needs in the area of learning can be tested and surveyed in the same way as students who attend regular schools (Heydrich, Weinert, Nusser, Artelt, & Carstensen, 2013). Therefore, these students are integrated in the samples of both school cohorts.

8 The first SUF for NEPS Starting Cohort 3—Grade 5 was released in September 2010 with data from the first surveys in fall 2010 (doi:10.5157/NEPS:SC3:1.0.0).

Table 3 Starting Cohort 3—Grade 5: Basic Structure of the Samples

Variables	Main field	Individual retracking		
	Total (<i>n</i> = 5,654)	Individual reasons (<i>n</i> = 213)	School-based reasons (<i>n</i> = 231)	Total (<i>n</i> = 444)
Sex				
No information	0%	0%	1%	1%
Male	52%	54%	55%	54%
Female	48%	46%	44%	45%
Year of birth				
No information	0%	0%	1%	1%
Valid information	100%	100%	99%	99%
Mean (std. dev.)	1999 (0.6)	1999 (0.8)	1999 (0.7)	1999 (0.8)
Median	2000	1999	1999	1999
Min	1995	1994	1997	1994
Max	2002	2001	2000	2001
Migration background (first generation)				
No information	5%	6%	5%	5%
No	91%	89%	85%	87%
Yes	4%	5%	10%	8%
Migration background (second generation)				
No information	4%	5%	4%	5%
No	72%	62%	64%	63%
Yes	24%	33%	32%	32%
Parent interview				
No parent interview	31%	39%	45%	42%
Parent interview	69%	61%	55%	58%
Education of the parents				
No information	0%	0%	0%	0%
No higher education	79%	85%	89%	87%
Higher education	21%	15%	11%	13%
School track (first wave fall 2010)				
Elementary school	5%	4%	16%	11%
Hauptschule	10%	24%	44%	34%
Realschule/Gesamtschule	34%	32%	20%	25%
Gymnasium	42%	20%	0%	10%
Förderschule	9%	20%	20%	20%

Note. The difference to the total sample of $N = 6,112$ can be explained by $n = 14$ students who withdrew their willingness to participate in the study.

Regarding migration background, we find significant differences ($p < 0.01$ for the first generation, as compared with $p < 0.001$ for the second generation) between students in the main field and those in the individual field: In the latter, the proportion of participants who have migrated themselves (8 %) or who have at least one parent who migrated (32 %) is higher than in the main field (4 % and 24 %, respectively). In contrast to the evidence from Starting Cohort 4, we find no clear hint that those with individual reasons (5 % and 33 %, respectively) are more similar to the main field than those with school-based reasons (10 % and 32 %, respectively).

In the main field, we find a slightly higher proportion of participants with a parent interview (69 %; $p < 0.05$) than in the field of individual retracking (58 %). Regarding the two groups in the field of individual retracking, we also find only slight and no significant differences: We have a parent interview for 61 % of those with individual reasons and for 55 % of those with school-based reasons. Similar to Starting Cohort 4, we find a clear difference between the educational backgrounds of those in the individual field compared with the main field: While the proportion with highly educated parents reaches 21 % in the main field, it is 13 % in the individual field ($p < 0.01$). Again similar to Starting Cohort 4, the group with individual reasons in the field of individual retracking (15 % with parents with higher education) is slightly more similar to the main field than those with school-based reasons (11 % with parents with higher education).

Looking at the school track, it is important to add that in Starting Cohort 3, students might also still be in elementary schools because in two of the Federal States in Germany, elementary school ends after Grade 6. As NEPS Starting Cohort 3 starts with Grade 5, we find a small proportion of 5 % of our participants in the main field in elementary school. Furthermore, 10 % of the participants are in Hauptschule, 34 % are in a kind of middle school, 42 % are in Gymnasium, and 9 % are in Förderschule. The field of individual retracking differs again significantly ($p < 0.001$) from the main field. Comparable with the ninth graders, we find a higher proportion of students in Förderschule (20 %) and Hauptschule (34 %) in the individual field and a clearly lower proportion of students in Gymnasium (10 %).

Looking at the different reasons for the change to the field of individual retracking, we find the same tendencies as in Starting Cohort 4: Those with individual reasons originate more often from a Gymnasium (20 % vs. 0 %), and those with school-based reasons more often from a Hauptschule (24 % vs. 44 %), although the proportion originating from a Förderschule is the same (both 20 %). In addition, the proportion coming from elementary school is 11 percentage points higher in the group with school-based (16 %) compared with individual reasons (4 %; main field: 5 %).

4.5 Response Rates

As we did not send out the materials for students with special educational needs in the field of learning in special schools for this first field of individual retracking in Starting Cohort 9, we only contacted 280 students individually. At that time, the participants were sent a motivation letter (as were their parents to inform them), the paper-and-pencil questionnaire of the main field, and the short update questionnaire for the address to their homes. The short questionnaire to update the status of the student mentioned in Section 3 had not been developed at that time. Furthermore, there was—also differing from the current concept—no reminder for this group. These two instruments were introduced for the first time for the field of individual retracking in the starting cohort of the fifth graders after empirical evidence from the starting cohort of the ninth graders (see below).

As shown in Table 4, after the survey material was sent out, 9% of the addresses of the students turned out to be incorrect so that the materials were returned. Regarding only those with valid addresses ($n = 249$), we received information from 51%; unfortunately, 49% ($n = 123$) did not send back any information. In the main field, 94% of all students participated in the survey in spring 2011. When comparing those proportions, we have to consider that the setting in the main field is completely different from the setting of the individual retracking. In the former, the students are in school and spend nearly one complete school day on NEPS testing and surveying. In the individual field, they have to fill out a paper-and-pencil questionnaire on their own in their leisure time. We also have to keep in mind that the field of individual retracking was started to avoid losing participants completely. In this respect, the fact that we could maintain the contact and collect information from half of the participants we otherwise would have lost for good represents a success.

Regarding the response rates in the two groups in the field of individual retracking, those with individual reasons, and those with school-based reasons, we can see a slight difference: While only 48% of those with individual reasons answered our questions, this proportion is 6 percentage points higher for the participants in the field with school-based reasons (54%, n. s.).

In the survey in fall 2011 for Starting Cohort 3, the concept of individual retracking was adjusted for the first time by adding a status questionnaire to the survey material. Furthermore, a reminder was sent out if there was no response to the first posting.

Finally, as shown in Table 5, 58% of the target persons in the individual retracking field with valid addresses returned their survey material. We received no answer at all from 42%. Comparable with Starting Cohort 4, we also have a problem with invalid addresses: The materials could not be sent out to 18% of the participants. In comparison, in the group with individual reasons, we have a response rate of 52%, and in the group with school-based reasons, we have a response rate of 63% ($p < 0.05$).

Table 4 Starting Cohort 4—Grade 9: Response in the Main Field and the Field of Individual Retracking (Spring 2011)

	Main field		Individual retracking					
	Total	Individual reasons	School-based reasons	Total	Individual reasons	School-based reasons		
Total	16,060	100%	343	100%	160	100%	183	100%
Not contacted at all	–		63	18%	14	9%	49	27%
No valid address	–		31	9%	22	14%	9	5%
Contacted (total)	16,060	100%	249	73%	124	78%	125	68%
No response			971	6%	123	49%	65	52%
Response			15,089	94%	126	51%	59	48%
							58	46%
							67	54%

Note: The difference to the total sample of $N = 16,425$ can be explained by $n = 22$ students who switched to the vocational educational system.

Table 5 Starting Cohort 3—Grade 5: Response in the Main Field and the Field of Individual Retracking (Fall 2011)

	Main field		Individual retracking					
	Total	Individual reasons	School-based reasons	Total	Individual reasons	School-based reasons		
Total	5,654	100%	444	100%	213	100%	231	100%
No valid address	–		78	18%	47	22%	31	13%
Contacted (total)	5,654	100%	366	82%	166	78%	200	87%
No response			327	6%	154	42%	80	48%
Response			5,327	94%	212	58%	86	52%
							74	37%
							126	63%

Note: The difference to the total sample of $N = 6,112$ can be explained by $n = 14$ students who withdrew their willingness to participate in the study.

5 Summary and Outlook

Panel studies based on surveys in institutional contexts run the risk of losing their participants if they leave these institutions or if the institution withdraws its willingness to participate in the study. To be able to follow the participants over their life course independent of the institutional context, the NEPS established a field of individual retracking. In this field, nonstandard educational careers are surveyed by a postal paper-and-pencil questionnaire, an address update analogous to the main field in school, and an additional short paper-and-pencil questionnaire on the current status of the students. Individual testing at home is planned before crucial transitions in the educational biography take place.

Summarizing the results from the analysis, the comparison of the basic sample structure of the main field and the individual field of both starting cohorts leads to some tentative conclusions: First, the proportion of participants with a migration background is nominally higher than in the main field, and the educational background in the individual field is nominally lower. Regarding the reasons for participants changing to the field of individual retracking, we find a more differentiated picture: While the group with school-based reasons is more likely to switch to individual retracking from lower school tracks, those students with individual reasons more likely originate from a Gymnasium. In other words, it seems that lower school tracks are more likely to cancel their participation than higher school tracks and that individual changes to other schools appear more likely in higher school tracks. Taking into account that students without migration and higher educational background (or rather, a socioeconomic background that is highly correlated with education) are more likely to attend the Gymnasium track and that students with migration and lower educational background are more likely to attend lower school tracks (cf., e.g., *Autorengruppe Bildungsberichterstattung*, 2010, p. 65), the aforementioned tendency could be explained by social disparities in school choice or selection.

Comparing the response rates between the two groups in the field of individual retracking, we find a lower participation rate in the group with individual reasons than in the group with school-based reasons in both cohorts. Against the background of the basic structure of the subsamples, we would have initially expected the opposite. Based on the thesis of “education bias” known from survey research (Hartmann & Schimpl-Neimanns, 1992) and the assumption that migration background coincides with a lower participation rate in education (Blohm & Diehl, 2001), a possible expectation could be that the response rate in the group with school-based reasons is lower than in the group with individual reasons. This is a question that should be investigated in detail in further research.

The challenge of keeping the participants in the panel, tracking their current status correctly, and collecting data that are comparable with the main frame increases with the number of alternatives for leaving the institutional context of the NEPS schools. This is especially the case at the transition from lower to upper secondary

school, which is also the point in time when the starting cohort splits up by default: In most Federal States, the educational pathways after Grade 10 split into a vocational track (leading to an occupation) and an academic track (leading to higher education) (cf. Figure 1).

Those students who leave school and transfer to the vocational track are followed by NEPS Stage 6 by way of telephone interviews twice a year (in fall and spring). They are tested every two years at home (Ludwig-Mayerhofer et al., 2011). The switch in survey mode is necessary for this group because the target persons become distributed over realms of possibilities in the vocational track so that an institutional perspective can no longer be upheld for the sample.

For the second group in this cohort, that is, students who continue their school education in the academic track, NEPS Stage 4 hands over the responsibility to NEPS Stage 5, which focuses on the pathway through the academic track to higher education (Wagner et al., 2011). Analogous to the surveys in Stage 4, the students in Grades 11 to 13 at NEPS schools are further tested and surveyed in the institutional context (including gathering information from the context persons). Also analogous to the former waves, there are students who cannot be reached at the NEPS schools anymore. At that stage, this group is especially large because two kinds of school tracks of the lower secondary school system, namely the Realschule and Gesamtschule, end after Grade 10, and many of the students from these tracks change to Gymnasium to attend higher education. We also know from pilot studies that it is very difficult for schools to differentiate whether the students change to vocational or academic tracks if they leave NEPS schools after Grade 10. Therefore, we decided to change the mode for the field of individual retracking after Grade 10, integrating it into the fall surveys of Stage 6.

This survey starts with a screening module to identify whether the respondent belongs to Stage 5 (academic track) or Stage 6 (vocational track). Afterwards, all stage-comprehensive questions are asked, and then the interview splits up: If the participant belongs to Stage 6, the survey program of Stage 6 is conducted; if he or she belongs to Stage 5, the telephone interview ends with an address update and the acquisition of at least one email address. For this group, the second part of the survey with the stage-specific program of Stage 5 is administered as an online questionnaire. Immediately after the end of the telephone interview, the target person is sent a link and a password to take part. It was explicitly decided that the target person has to participate in both surveys to receive the incentive.⁹ The online questionnaire is equiva-

9 The incentive is adjusted to the incentive in Stage 6, in which a higher incentive of €30 is administered to participants at risk of dropping out (those originating from lower secondary education), and a lower incentive of €15 is administered to those participants with a high probability of participation (those originating from middle or higher secondary education). Therefore, our special group receives an incentive of €15 as it is a low-risk group. This is also the reason why the decision was made to provide the incentive only after both the telephone and the online questionnaire have been completed.

lent to the paper-and-pencil questionnaire that the main frame answered in school (excluding the stage-comprehensive questions asked in the telephone interview at the beginning). For this group, we also set up the parent interview to receive comparable information on the context at home. At time of writing this article, we are waiting for first data to check whether this strategy is working.

In the NEPS, there is also a third starting cohort with a field of individual retracking: the Kindergarten cohort. In this cohort, administered in Stage 2, children are individually tested in Kindergarten two years prior to school enrollment. The children's educators are requested to provide some information on the children and the group the children attend. The principals of the Kindergartens are asked about context information of the Kindergarten. In this cohort, we have a coupling of the participation of parents and their children because the children cannot give us enough context information. Analogous to the school cohorts, the children can leave an NEPS Kindergarten, or an NEPS Kindergarten can withdraw its participation from the study. In these cases, individual retracking is organized via a parent interview (the current status of the child and address update; in Kindergarten, there is no questionnaire for the children). At the transition to elementary school in the year 2012, Stage 2 handed over the responsibility for this cohort to Stage 3. At this point, the sample in this starting cohort was refreshed by surveying the entire first grade (Aßmann et al., 2011). In order to keep these children and their parents in the panel, which we cannot track at the NEPS elementary schools, we are currently building a field of individual retracking analogous to the school cohorts with a status update, an address update, and a parent interview.

Up to now, the strategies applied in the NEPS to keep the panel participants in the school cohorts seem to have been working quite efficiently. In general, the panel participation rates are even higher than expected. It remains to be seen whether the strategies currently implemented in elementary school and upper secondary school are effective or whether new strategies need to be developed.

References

- Aßmann, C., Steinhauer, H. W., Kiesl, H., Koch, S., Schönberger, B., Müller-Kuller, A., ... Blossfeld, H.-P. (2011). Sampling designs of the National Educational Panel Study: Challenges and solutions. In H.-P. Blossfeld, H.-G. Roßbach, & J. von Maurice (Eds.), *Zeitschrift für Erziehungswissenschaft, 14. Education as a lifelong process: The German National Educational Panel Study (NEPS)* (pp. 51–65). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Autorengruppe Bildungsberichterstattung. (2010). *Bildung in Deutschland 2010. Ein indikatorengestützter Bericht mit einer Analyse zu Perspektiven des Bildungswesens im demografischen Wandel*. Bielefeld: W. Bertelsmann Verlag.

- Blohm, M., & Diehl, C. (2001). Wenn Migranten Migranten befragen. Zum Teilnahmeverhalten von Einwanderern bei Bevölkerungsbefragungen. *Zeitschrift für Soziologie*, 30(3), 223–242.
- Blossfeld, H.-P., & Schneider, T. (2011). Data on educational processes. National and international comparisons. In H.-P. Blossfeld, H.-G. Roßbach, & J. von Maurice (Eds.), *Zeitschrift für Erziehungswissenschaft*, 14. *Education as a lifelong process: The German National Educational Panel Study (NEPS)* (pp. 35–50). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Blossfeld, H.-P., von Maurice, J., & Schneider, T. (2011). The National Educational Panel Study: Need, main features, and research potential. In H.-P. Blossfeld, H.-G. Roßbach, & J. von Maurice (Eds.), *Zeitschrift für Erziehungswissenschaft*, 14. *Education as a lifelong process: The German National Educational Panel Study (NEPS)* (pp. 5–17). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Faust, G., Kratzmann, J., & Wehner, F. (2013). Methodische Anlage der BiKS-Einschulungsuntersuchungen. In G. Faust (Ed.), *Einschulung: Ergebnisse aus der Studie "Bildungsprozesse, Kompetenzentwicklung und Selektionsentscheidungen im Vorschul- und Schulalter (BiKS)"* (pp. 33–50). Münster: Waxmann.
- Frahm, S., Goy, M., Kowalski, K., Sixt, M., Strietholt, R., Blatt, I., ... Kandera, M. (2011). Transition and development from lower secondary to upper secondary school. In H.-P. Blossfeld, H.-G. Roßbach, & J. von Maurice (Eds.), *Zeitschrift für Erziehungswissenschaft*, 14. *Education as a lifelong process: The German National Educational Panel Study (NEPS)* (pp. 217–232). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Hartmann, P.H., & Schimpl-Neimanns, B. (1992). Sind Sozialstrukturanalysen mit Umfragedaten möglich? Analysen zur Repräsentativität einer Sozialforschungsumfrage. *Kölner Zeitschrift für Soziologie und Sozialpsychologie*, 44(2), 315–140.
- Heydrich, J., Weinert, S., Nusser, L., Artelt, C., & Carstensen, C. H. (2013). Including students with special educational needs into large-scale assessments of competencies: Challenges and approaches within the German National Educational Panel Study (NEPS). *Journal for Educational Research Online*, 5(2), 217–240.
- Homuth, C., Mann, D., Schmitt, M., & Mudiappa, M. (2014). Eine Forschergruppe, zwei Studien: BiKS-3-10 und BiKS-8-14. In M. Mudiappa, & C. Artelt (Eds.), *BiKS—Ergebnisse aus den Längsschnittstudien: Praxisrelevante Befunde aus dem Primar- und Sekundarschulbereich* (pp. 15–28). Bamberg: University of Bamberg Press.
- Lorenz, C., Schmitt, M., Lehl, S., Mudiappa, M., & Roßbach, H.-G. (2013). The Bamberg BiKS Research Group. In M. Pfof, C. Artelt, & S. Weinert (Eds.), *The development of reading literacy from early childhood to adolescence: Empirical findings from the Bamberg BiKS longitudinal studies* (pp. 15–34). Bamberg: University of Bamberg Press.
- Ludwig-Mayerhofer, W., Solga, H., Leuze, K., Dombrowski, R., Künster, R., Ebralidze, E., ... Kühn, S. (2011). Vocational education and training and transitions into the labor market. In H.-P. Blossfeld, H.-G. Roßbach, & J. von Maurice (Eds.), *Zeitschrift für Erziehungswissenschaft*, 14. *Education as a lifelong process: The German National Edu-*

- ational Panel Study (NEPS)* (pp. 251–266). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Maaz, K., Baumert, J., Neumann, M., Becker, M., & Dumont, H. (Eds.). (2013). *Die Berliner Schulstrukturreform: Bewertung durch die beteiligten Akteure und Konsequenzen des neuen Übergangsverfahrens von der Grundschule in die weiterführenden Schulen*. Münster: Waxmann.
- Maaz, K., Baumert, J., Neumann, M., Becker, M., Kropf, M., & Dumont, H. (2013). Anlage und Zielsetzung der BERLIN-Studie. In K. Maaz, J. Baumert, M. Neumann, M. Becker, & H. Dumont (Eds.), *Die Berliner Schulstrukturreform: Bewertung durch die beteiligten Akteure und Konsequenzen des neuen Übergangsverfahrens von der Grundschule in die weiterführenden Schulen* (pp. 35–48). Münster: Waxmann.
- Mudiappa, M., & Artelt, C. (Eds.). (2014). *BiKS—Ergebnisse aus den Längsschnittstudien: Praxisrelevante Befunde aus dem Primar- und Sekundarschulbereich*. Bamberg: University of Bamberg Press.
- Porst, R. (2001). *Wie man die Rücklaufquoten bei postalischen Befragungen erhöht*. In *ZUMA How-to-Reihe* (Vol. 9, pp. 1–12). Mannheim: Zentrum für Umfragen, Methoden und Analysen.
- Schmidt, S., Schmitt, M., & Smidt, W. (2009). *Die BiKS-Studie: Methodenbericht zur zweiten Projektphase*. Retrieved from http://psydok.sulb.uni-saarland.de/volltexte/2009/2534/pdf/Methodenbericht_2009.pdf
- von Maurice, J., Artelt, C., Blossfeld, H.-P., Faust, G., Roßbach, H.-G., & Weinert, S. (2007). *Bildungsprozesse, Kompetenzentwicklung und Formation von Selektionsentscheidungen im Vor- und Grundschulalter: Überblick über die Erhebungen in den Längsschnitten BiKS-3-8 und BiKS-8-12 in den ersten beiden Projektjahren*. Retrieved from http://psydok.sulb.uni-saarland.de/volltexte/2007/1008/pdf/online_version.pdf
- Wagner, W., Kramer, J., Trautwein, U., Lüdtke, O., Nagy, G., Jonkmann, K., ... Schilling, J. (2011). Upper secondary education in academic school tracks and the transition from school to postsecondary education and the job market. In H.-P. Blossfeld, H.-G. Roßbach, & J. von Maurice (Eds.), *Zeitschrift für Erziehungswissenschaft, 14. Education as a lifelong process: The German National Educational Panel Study (NEPS)* (pp. 233–249). Wiesbaden: VS Verlag für Sozialwissenschaften.

Acknowledgement

The authors wish to thank Christoph Ruthenfranz for his assistance in the database search and in screening the corpus of research literature on longitudinal educational studies as reported in Section 2 of this article.

About the authors

G. Besuch

Formerly employed at the IEA Data Processing and Research Center (DPC),
Überseering 27, 22297 Hamburg, Germany.

M. Goy

Institute for School Development Research (IFS),
TU Dortmund University, Vogelpothsweg 78, 44227 Dortmund, Germany.

M. Sixt

Leibniz Institute for Educational Trajectories (LifBi),
University of Bamberg, Wilhelmsplatz 3, 96047 Bamberg, Germany.
e-mail: michaela.sixt@lifbi.de

Challenges and Intentions of Target-Specific Public Relations Work

Götz Lechner, Julia Göpel and Anna Passmann

Abstract

This contribution develops a framework of public relations activities for the National Educational Panel Study (NEPS) carried out by the Leibniz Institute for Educational Trajectories (LifBi) at the University of Bamberg. Public relations (PR) work to promote the success of large-scale assessments addresses at least three kinds of target groups in the NEPS case: participants in the NEPS study, the scientific community, and the stakeholders of the institute. According to different sociological approaches, particularly the functional theory of Niklas Luhmann and the structuration theory by Anthony Giddens, this contribution offers two modes of access to the problem of how to handle the challenges of public relations that support large-scale-assessments like the NEPS in modern Western societies. Based on these preliminary considerations, this paper identifies the relevant target groups and describes the variety of the tailored content and media provided for their communicational purposes by the operational unit Public Relations and Respondent Communications of the LifBi. For these purposes, specific codes of functional systems in the spheres of science, politics, and the educational system are used to deepen functional requirements regarding the systems that the NEPS and LifBi are part of as symbolic codes of appreciation. These codes esteem to foster acceptance and commitment of the survey participants. The media and public relations work for the NEPS is an ongoing process that is constantly being improved upon and expanded in reaction to changes and in response to new demands of the heterogeneous target groups.

1 Introduction

The cultural, economic, and social changes over the last 50 years have profoundly altered the method of access to target persons in the survey process. Dillman, Smyth, and Christian (2008) condense the impact of this change in surveying to six dimensions shown by the following figure from one of the most recommended survey handbooks, “The Tailored Design Method”. These dimensions describe the target persons’ involvement in and control of the survey setting over time in a (at first sight) comprehensive yet (from a sociological point of view) insufficient perspective (see Figure 1).

Over the past decades, human interaction has become media-based or -assisted interaction. This process is connected to what Giddens (1990) calls the disembedding of the individual self, of values, and of action: The distancing of time and space and the overwhelmingly growing importance of abstract systems in modern societies’ everyday life have weakened traditional institutions not only by technical means, but by the institutional change in shape. Abstract systems such as the modes of exchange condensed in the idea of “money”, as well as highly technical expert systems such as airplanes and the World Wide Web, deprive nearly everyone of direct experience and

Figure 1 Seventy-Five Years of Change in Respondent Involvement and Control Over the Survey Process. Dillman et al. (2008, p. 2)

Characteristic	Through the 1960s	1970s through 1980s	1990s to the Present
Human interaction	High: Face-to-face through in-person visits to respondents’ homes	Medium: Remote through a telephone connection	Low: Encounter is more likely to be with a machine or its products
Trust that the survey is legitimate	High: Encouraged by interviewer presence, appearance, and sincerity	Medium: Encouraged through voice inflection and ability to listen to and request additional information	Low: Because of possibility that survey is fake and potentially harmful to respondent
Time involvement with each respondent	High: Interviewer goes to respondent and obtains information one-on-one	Medium: One-on-one, but contact effort is minimal	Low: Minimal to no time with individual respondents
Attention given to each respondent	High: Because of time to find and interview each respondent	Medium: Because of placing calls one after another	Low: Mass emails
Respondent control over access	Low: Households generally accessible	Medium: Unlisted numbers, voice mail, and call monitoring	High: Caller ID, call blocking, email filters
Respondent control over whether to respond	Low: Required breaking off of human interaction	Medium: Ease of hanging up telephone	High: Increased disclosures required to be communicated, social support for refusing

the sense of knowing how something happens or works. These cases are about the sensation of how and why financial markets work or why planes fly.

Institutions—conceived as the way things *are done* and based on the normative requirements of persisting everyday life, the way they *have to be done* as a representation of the power behind daily routines—depend on structure and agency. The material aspect of structure in modern societies mostly appears in a kind of alienating fabric that connects the noncommittal over spreading space by using technical time measurements beyond a common sense of time.

From Giddens' point of view, these gaps and empty spaces in life, recognition, and what we expect to be the *material world* must be filled with “ontological security” or “trust”. Ontological security stands for “confidence or trust that the natural and social worlds are as they appear to be, including the basic existential parameters of self and social identity” (Giddens 1984, p. 375).

Trust is growing in the sense of positive everyday experiences, and ontological security arises from collecting good experiences on the control over fear: “The psychological origins of ontological security are to be found in basic anxiety-controlling mechanisms. ... The generation of feelings of trust in others, as the deepest-lying element of the basic security system, depends substantially upon predictable and caring routines established by parental figures... . Ontological security is ... maintained ... by the very predictability of routine, something which is radically disrupted in critical situations. The swamping of habitual modes of activity by anxiety which cannot be adequately contained by the basic security system is specifically a feature of critical situations” (Giddens, 1984, pp. 50). We betray this system of anxiety-control, which has been shaken through the last decades by social and technological change, by intruding into the everyday life of our respondent in an extremely intimate way with the curiosity of social surveying and the potential danger of leaks that might spread their personal data in illicit use.

Dillman et al. (2008) in Figure 1 paint a picture of the 1960s, which emphasizes face-to-face communication over the importance of abstract systems, shared traditional values over the consequences of individualization, and the order of tidiness over rule-less anomie. These shared values ground institutions of good behavior in interaction. It is the picture of 1960s suburbia in which the interviewer in dress and tie interrogates the White Anglo-Saxon Protestant head of the family in a proper detached house while having a cup of coffee.

This picture reflects a certain kind of shared, everyday ideology, a certain political style (Bornschiefer, 2008) in which the (white) middle class as an icon describes the whole society. This image may perhaps seem to be quite American or insufficient, but in a very self-evident way, it also fits for Western Germany in the late 1960s.

To understand the challenges of PR work supporting survey processes¹ in large-

1 Searching for literature about surveying, such as Dillman et al. (2008); Engel, Bartsch, Schnabel, and Vehre (2012); Groves and Couper (2012); Kuß (2012); Marsden and Wright (2010); Proner (2011);

scale assessments, it is necessary to add another analytic category to this simplifying starting point: complexity.

Luhmann (1998) describes at least twelve so-called “function-systems” (*Funktionssysteme*) of communication, a terminus we now explain, that evolved from the need to channel the overwhelming complexity of the modern world. The exclusive method of binary coding within these systems that processes their program to fulfill their special aims closes their horizon of sense and meaning: These function-systems of meaning build themselves from own elements and from their own material (autopoiesis), and in consequence, they only understand themselves, or as Luhmann would say: These function-systems are self-referential. The whole world, that is, all other function systems surround this network of exclusive communication as an environment. Communication between system and environment is more than unlikely because all these systems “live” on their own.

The neurobiologists Maturana and Varela (1980) proclaimed the discovery of the autopoiesis of consciousness: The psychic system of each and every living system is different. In simple experimental observations on the morphogenesis of reptile encephala, they were able to show that the structure of conjunction within the cerebral material depends on certain, different experiences with which the creatures have been treated. Luhmann (1998), who based his complete social theory on this assumption, in consequence, reduced mankind to those psychic systems that try to communicate but necessarily fail on the basis of self-reference. Thinking through this epistemological position, we must ultimately concede that the idea of communication for the sake of mutual comprehension cannot hold. Hence, we must recognize the problem of conceptualizing PR on the basis of this premise. A concept of communication without comprehension will become even more cumbersome and will require more adjustment the closer it comes to people’s everyday lives. Beyond this sphere, Luhmann’s idea of functional systems remains, however, a very fruitful approach.

Large-scale survey assessments regarding educational processes like the NEPS are primarily based in the scientific system, whereas they target the educational system. The code of the educational system is “good and bad grades”, and the code of the scientific system is “true or false”. If one tries to translate from one system to the other, the problem becomes obvious. The political system processes power following the code of cabinet or opposition, whereas the legal system processes the law in terms of legal and illegal ideas, and mass media play their role of providing information and entertainment coded in (non)information or gained attention. Money, the medium

Stoop, Billiet, Koch, and Fitzgerald (2010); and Vehre (2011), reveals that no one so much as mentioned PR as a part of the process!

Glaser (2012) is the only one who is aware of the existence of PR work in this context and lists some examples for high-profile public relation campaigns that were run in order to support surveys the way Dillman et al. conclude that participants care in Figure 2. However, even this contribution lacks a starting point for an explanation as to how and why public relation efforts could assist the surveying process.

Figure 2 Overview of the Tailored Design Method. Dillman et al. (2008, p. 38)

A. Tailored design is the development of survey procedures that work together to form the survey request and motivate various types of people to respond to the survey by establishing trust and increasing the perceived benefits of completing the survey while decreasing the expected costs of participation.

B. Successful tailored design attends to the multiple sources of survey error—coverage, sampling, measurement, and nonresponse—with a focus on minimizing overall survey error.

C. Tailored design involves customizing survey procedures for each particular survey situation based on knowledge about the topic and the sponsor of the survey, the types of respondents who will be asked to complete the survey, and the proposed budget and time frame for reporting the results.

D. Multiple aspects of the implementation process and the questionnaire can be combined in different ways to encourage respondents to participate by creating trust in the sponsor and influencing the perceived expectations of the benefits and costs of responding to the survey.

To establish trust	To increase benefits of participation	To decrease costs of participation
<ul style="list-style-type: none"> • Obtain sponsorship by legitimate authority • Provide a token of appreciation in advance • Make the task appear important • Ensure confidentiality and security of information 	<ul style="list-style-type: none"> • Provide information about the survey • Ask for help or advice • Show positive regard • Say thank you • Support group values • Give tangible rewards • Make the questionnaire interesting • Provide social validation • Inform people that opportunities to respond are limited 	<ul style="list-style-type: none"> • Make it convenient to respond • Avoid subordinating language • Make the questionnaire short and easy to complete • Minimize requests to obtain personal or sensitive information • Emphasize the similarity to other requests or tasks to which a person has responded

of the economic system, is binary-coded in “have” and “have not”. These codes of the function-systems in the Luhmann conceptualization and these media processing self-reference in its opaque meaning for everyday life pave the way to understanding the distance between what we know about the world and “expert systems”. Giddens would call these “functional circumstances”.

The discovery of self-reference joined the re-conceptualization of individualization (Beck, 1986), or in other words, the idea that beyond status and class, the individual is freed from inherited constraints of social origin, namely cultural institutions, at least in case of the German scientific community. Both ideas atomize the area of (mass) communication as well as the ideas of collective sense and inter-individual shared meaning. From the sociologically informed point of view, a challenge arises that Dillman et al. (2008) try to cover in three dimensions (following four programmatic headlines at the bottom of Figure 2).

Column 1. The “Giddens” point: The distancing of time and space, the increasing role of “expert systems”, and the weakening of the normative frame of agency by individualization mean that it is necessary for those responsible for surveys to establish trust. These modern gaps of “ontological security” may be partially bridged by the cooptation of legitimate authority. However, these efforts are not only necessary

when dealing with the confidentiality of data achieved or the legitimation and importance of the survey.

Column 2/3. The column headlines in Figure 2 ending on “participation” both focus on the two sides of a single coin: benefits and costs. There are real benefits, real tangible rewards, if those surveyed are offered incentives. All the other “costs” and “benefits” semantically pronounce respect and appreciation, group membership, and politeness. All these non-economic categories, freed from a specific ideological bias (the homo-oeconomicus idea) and (re)situated in what Schütz (1932) and later Habermas (1981) called “Life World” (*Lebenswelt*) with regard to Husserl, combine autopoiesis and homo oeconomicus, that is, the need to secure trust and self-reference on the challenges PR work is facing by supporting the survey process in large-scale assessments together. The Life World as the “given” world, that is, the world “as lived” prior to scientific analysis and reflective representation, provides valuable security. The Life World is host to sense and meaning in communication for everyday life, and meaning as sense grounds our praxis of respect and credit, of politeness and the rules of group membership, in other words: symbolic codes of appreciation and esteem.

At this point, PR work starts in the Life World with detecting the different dimensions of sense and meaning as well as the code(s) of everyday life. Here, trust is born; here, people live beyond strategies of optimizing economic benefits. From Life World to function systems, sense and meaning become the specific code of function systems if PR work addresses structures beyond everyday life. These codes must be translated if there is information provided from the science system that addresses other function systems. From this point of view, any scientific finding must be either transformed and translated to meaningful information targeting the Life World or must be coded to information that matches the programs of other function systems such as politics, economics, law, mass media, or the educational system. PR work in this abstract sense means to discover, tailor, and provide content and information for specific target groups. This simple conclusion sheds light on many challenges for PR work supporting the survey process in large-scale assessments.

2 First Challenge: Defining and Identifying Target Groups and Their Need for Information

Resulting from the study description and the aims of the project with its representativeness for Germany, it is not wrong to assume that the public relations work of the National Educational Panel Study (NEPS) should address more or less everybody in the Federal Republic of Germany. Resulting from the study conception, however, we focus on a number of target groups. The reason for this decision and the definition of these most important groups that have to be addressed specifically in the context of the PR work of the NEPS are outlined subsequently.

2.1 Participants in the Study

Both the persistent acceptance and the long-term willingness to participate for a total of approximately 100,000 persons are of the utmost importance for the successful realization of the panel study. Portioned to the six starting cohorts, the following groups participate:

- Starting Cohort 1—Early Childhood: 7-month-old babies (target), mothers (parents) as context persons
- Starting Cohort 2—Kindergarten: 4-year-olds in Kindergarten (target); parents, educators, heads of Kindergarten facilities as context persons
- Starting Cohort 3—Grade 5: Grade-5 students at regular schools and those attending special schools (target); parents, teachers, school principals as context persons
- Starting Cohort 4—Grade 9: Grade-9 students at regular schools and those attending special schools (target); parents, teachers, school principals as context persons
- Starting Cohort 5—First-Year Students: First-Year students (target, no context)
- Starting Cohort 6—Adults: Adults of birth cohorts 1944 to 1986 (target, no context)

As mentioned above, PR work in the case of the participants and their relatives means tailoring scientific information in the Life World sense of moral discourse, group-identity, and trust.

2.2 Gatekeepers in the Educational System

Participation is not compulsory for the randomly selected persons. It is therefore all the more important that we keep in close contact with gatekeepers of (or in) the educational system. Outside the educational system, it is more difficult to influence the willingness to participate.

Our gatekeepers in this context are (1) cooperation schools participating in pilot or main studies, (2) Ministries of Education and Cultural Affairs in all 16 Federal States, and (3) social interest groups such as teachers' and parents' associations and student councils.

At this point between the Life World and systems, we need to address the educational system in which the code operates binarily in good grades/bad grades and these grades are used to rate not only the students.

2.3 Stakeholders in Politics and Media (Public)

In addition to the communication with directly involved policy-makers, that is, the responsible persons at the Federal Ministry of Education and Research (*Bundesministerium für Bildung und Forschung*) and the German Federal State Ministries of Education and Science (especially the Bavarian State Ministry of Science, Research and the Arts (*Bayerisches Staatsministerium für Bildung und Kultus, Wissenschaft und Kunst*)), it is important for the LIfBi, on behalf of the NEPS, to get in contact with as many persons as possible who are active on the political and administrative level. In order to assure this, members of the NEPS consortium give presentations on the concept and design of the study to interested representatives on this level on site at the Central Coordination Unit in Bamberg as well as upon the invitation of third parties at random intervals. As addressees, special focus is placed on the organizational level of political parties, the Standing Conference of the Ministers of Education and Cultural Affairs (*Ständige Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland*), and the individual Ministries of Education and Cultural Affairs of the Federal States (*Kultus- und Wissenschaftsministerien der Länder*).

Information from the NEPS is now to be translated in the sphere and the codes of the political system: “True and false” must become “to have power or not to have power”.

In order to use the established media platforms, active contact with representatives of regional and national media is necessary. The aim is to increase the awareness of the NEPS by publishing articles and interviews in daily and weekly newspapers. Association magazines are also used.

2.4 The Scientific Community

The NEPS was set up as an infrastructure facility to collect data and to prepare Scientific Use Files for the national and international scientific community, that is, to build up a high-quality, extensive infrastructure in order to be able to find out more about how education is acquired, to understand how it impacts on individual biographies, and to describe and analyze the major educational processes and trajectories across the life span. Therefore, the scientific community is also one of our most important target groups for which, of course, the Research Data Center LIfBi providing NEPS data is the main point of contact. The public relations team, however, provides support in many aspects, for example, in the preparation of information material and organizing presences at national and international scientific conferences of the different disciplines.

Section 3 describes the methods and means of how the most important target groups are addressed to in more detail.

3 Second Challenge: From Face-to-Face to “Media”: Different Channels and Levels of Use in PR Work

The categorization by target groups is one possible approach to give an overview of target-specific public relations work. Another option would be using the specification and explication of the various channels and levels we use in doing public relations work. In the following section, we outline these methods and means.

3.1 Personal Contact

Personal contact is unsurprisingly extremely important and valuable in communicating with the different target groups. The face-to-face contact with our participants occurs mainly via the interviewers or test leaders of our two survey institutes, infas (Institut für angewandte Sozialwissenschaft GmbH) and IEA Data Processing and Research Center, depending, of course, on the mode of the survey. However, participants in all substudies mentioned above as starting cohorts have different possibilities to get in contact with the responsible research institute, and there is also a contact person for the NEPS at the Central Coordination Unit of the LIfBi (not face-to-face contact, but via e-mail or telephone hotline). The participants' feedback given to the contact person in Bamberg concerning, for example, changes of address or telephone number, is directly reported to infas under strict observation of the data-protection regulations in order to clarify the concern as soon as possible. Questions about the study in general and about the content of the survey are answered directly by the contact person in Bamberg with the support of the respective operational unit of the NEPS. These contact possibilities are very frequently used and represent a valuable contribution to panel care.

We are in regular and close contact with the gatekeepers of the educational system and give information sessions, for example, for schools participating in pilot or main studies, on whose acceptance and cooperation the study depends. We have to contact the Ministries of Education and Cultural Affairs in all 16 Federal States before each study in the school context. Furthermore, we are in contact with representatives of social interest groups such as teachers' and parents' associations or student councils. The information and exchanges given at these meetings have proven to be very valuable for both parties. On the one hand, the guests appreciate the proactive public relations work, and on the other hand, the level of awareness of the study is increasing and the acceptance in the population is growing constantly as the gatekeepers often act as disseminators and also afford the opportunity to place an article about the NEPS in their organization journal.

For stakeholders in politics and media, the case is similar: The LIfBi invites foremost local politicians more-or-less regularly and informs them about the current state of the NEPS and future plans. They pass on the information in their networks

and hence also work as disseminators. Of course, we are in very close contact with the mayor of Bamberg and the City Council. The economic importance of the LIfBi's execution in the NEPS for both the city and the region is obvious: With over 100 employees in Bamberg and an annual budget amounting to a notable eight-figure sum, the economic value of this research institution can be compared with that of a medium-sized company.

The face-to-face-contact with the scientific community takes place in various formats. As mentioned above, the public relations team organizes attendance at scientific conferences of different disciplines in Germany and abroad, for example, at the Deutsche Gesellschaft für Soziologie (DGS, biannual), the Deutsche Gesellschaft für Psychologie (DGPs, biannual), the Gesellschaft für Empirische Bildungsforschung (GEBF, annual), the Arbeitsgruppe für Empirische Pädagogische Forschung (AEPF, annual), the Society for Longitudinal and Life Course Studies (SLLS, annual), the European Survey Research Association (ESRA, biannual), the American Sociological Association (ASA, annual), the American Educational Research Association (AERA, annual), the European Sociological Association (ESA, biannual), the European Conference on Educational Research (ECER, annual), the European Association for Research on Learning, and the Instruction (EARLI, biannual). Furthermore, the LIfBi is present at public events such as the annual "Nuremberg Metropolitan Region Science Day". The conference attendance actively contributes to network-building and recruiting of data users. The Research Data Center LIfBi provides regular user trainings in Bamberg as well as in several other countries, such as Italy, Poland, and Korea. Furthermore, members of the Research Data Center give personal user support via e-mail and a telephone hotline.

3.2 Print ("Contact") Material

For participants, printed "contact" material—apart from the formal printed letters that target persons receive for courtesy reasons and legal restraints—is of special importance. If people are informed and interested, there are better chances for them to participate in the study (Budowski and Scherpeenzel, 2005). This material, which mostly consists of leaflets, brochures, and greeting cards with an attractive layout and short, easily comprehensible texts, aims to address the individual proactively, to show our high appreciation for and also the importance of their participation, for without them, the NEPS could not function. Furthermore, once a year, we provide printed feedback material to the participants containing information on the progress of the study, the first general results, and future steps. Therefore, this material also contributes to a complex panel-care system.

Persons in a "cared" panel are more likely to participate in the next wave than are persons in an "uncared" panel (Krebs, 1986). Generally, concerning printed materials, the flyer and brochure formats are possible depending on how much information we

would like to give. The content is developed by the corresponding responsible operational units of the NEPS. The public relations team transforms the scientific content into everyday language that the participants can understand and adjusts the findings and the outline of the proceedings to the necessary format and layout. Given the geographical distribution of the NEPS and the number of staff members involved, this process is quite complex. To structure it, the public relations team supplies a time line dating back from the deadline when the printed material has to be on site at one of our research institutes. Within this time line, several revision phases involving the related operational units, the commissioned research institute, and the Executive Director of Research have to be observed. To structure the complex process of the production and distribution of feedback materials for different target groups, a deliberated concept is needed to provide all relevant information and results for every participant from our six starting cohorts in every stage of their life courses.

The Operational Unit Public Relations and Respondent Communications also provides printed material for gatekeepers and stakeholders in politics and media in different layouts and contents. Here, the specific codes of the functional systems are used to gain access to their functional system operating power or public awareness and deepen their commitment to the study.

For the scientific community, a general-information brochure as well as an informational leaflet (both in German and English) is kept up to date. Furthermore, a research-data leaflet provides an overview on the NEPS's six starting cohorts and the two additional studies in Thuringia and Baden-Wuerttemberg, with information on samples, data structure, datasets, data access, and data releases.

3.3 Mass Media

In our context, we must distinguish between the “traditional” mass media we serve, that is, print media (newspapers, journals, and books) and radio, and the “new” mass media we serve, that is, digital media (such as internet and email newsletters). In addition to the approach via target groups or the various channels and levels we use in doing public relations work, we now would like to point out the different mass media we use to reach our target groups.

“Traditional” mass media

The difficulty in gaining the attention of “traditional” mass media is the news value—the audience would like to read or hear something new and specific, for example, from a specific person who reports concretely about his or her experience of participation in the NEPS. For data-protection reasons, this is, of course, not possible. It was hard to explain that within the first few years until the first data release of the NEPS, the average citizen did not see any specific output, whereas the amount of taxpayers' money financing the NEPS was a very specific sum. Even then, it was necessary to ex-

plain that the NEPS as a research infrastructure project has “only” the task of providing high-quality data to the scientific community; the analysis of these data and the reporting are the next steps, for which the NEPS is not financed. However, we have already often managed to gain the attention of newspapers and journals, both regional and national. This was done with the large opening ceremony in February 2009, when the former Federal Minister for Education and Research began with greetings, with various articles in newspapers and professional journals, with interviews with the Managing Project Director and scientific staff, and with the numerous occasions of visits of politicians and the process of institutionalization. The public relations team maintains a press review to accompany the reaction of the media over time. It is becoming obvious that the number of articles published on the basis of NEPS data will continue to grow with the number of datasets released. The results obtained with our data have a high potential to contribute to the public discussion. The presence of the NEPS in national and international media will be improved, and an even closer network of journalists than the current one will be built.

The scientific NEPS staff produces a constantly growing number of publications as well as articles in journals, books, and independent works, such as the special edition of the “*Zeitschrift für Erziehungswissenschaften*” (ZfE) and this book volume. An updated publication list can be found on our homepage.

“New” mass media

The main media here, of course, is our web presence. The operational unit Public Relations and Respondent Communications of the LIfBi maintains two web addresses on behalf of the NEPS, “www.neps-studie.de” intended for our participants and www.bildungspanel.de as well as www.neps-data.de, intended for the general public and the scientific community, respectively. The information on “www.neps-studie.de” is edited target-group specifically and updated regularly: The homepage, which contains more general information, guides the participants intuitively to their substudy via by photos with recognition value. For each substudy, detailed information can be found on clearly structured bottom pages, for example, on central questions, design, contact persons, and especially data-protection issues. In order to meet the dynamic requirements of modern societies and to increase the attractiveness for the users, this website is improved and expanded continuously. Parts of the new website are offered in Turkish, and Russian. Furthermore, as part of the panel care, an online form developed by the commissioned research institute infas is linked to this particular website to make it easier for the participants to update contact data—of course under the strict observation of all data-protection regulations. The homepage of “www.bildungspanel.de” and “www.neps-data.de” is one and the same because we promote “www.bildungspanel.de” for the general public as well as for stakeholders in politics and media, and we promote “www.neps-data.de” for the international scientific community. Of course, this site is available both in German and English. This website is structured in two parts: One part is for the general public that provides information

on the project in general, the staff, the NEPS boards, visiting scholars, and publications; the other part is the website of the Research Data Center LIfBi, which provides the NEPS with data and offers comprehensive information for data users, for example, all substudies and additional studies of the NEPS, data access, the data-release schedule, and user trainings and support. Both parts of this website have a news section in which relevant information about current events, guest researchers, and data releases are published on one or the other part, depending on the target group.

The operational unit Public Relations and Respondent Communications of LIfBi publishes two e-mail newsletters twice per year, both of which focus on two different target groups of the NEPS: (1) The scientific newsletter, LIfBi *data*, is sent via email to a constantly growing mailing list of interested researchers. The latest issue has been sent to more than 1,000 addressees. It is published in English and informs readers about the most interesting news for the scientific community, for example, upcoming data releases, dates for NEPS User Trainings, current NEPS Working Papers, and important developments and events. (2) We keep the political-administrative level up to date with a semi-annually newsletter, LIfBi *info*, in German, which contains short articles about the status of the NEPS and the other projects carried out by the LIfBi, latest developments, and recent events. The latest issue of LIfBi *info* has been sent to more than 400 addresses, including contact persons in the German Ministry for Education and Research and in the Ministries of Education in the 16 Federal States.

Within the social media sector, the NEPS decided not to become a member of Facebook. The main reasons for this are include the wish to not have any self-disclosure of our study participants, which would be inevitable. A second main reason is that the progress of messages and comments would be impossible to control, and one negative view could have severe consequences.

4 Third Challenge: Corporate Design

Discipline in appearance means visibility—a high external visibility in connection with an active and open information policy is indispensable for enhancing acceptance and trust in a broad public. A consistent corporate design is the basis for this. For the NEPS and its hosting institution LIfBi, visibility, acceptance, and trust are all the more important as the project is financed by public funds.

At the outset of the project, professional public relations agencies were engaged in the development of the official NEPS logos: the German “Nationales Bildungspanel,” English “National Educational Panel Study,” and our participants’ logo “Bildungsverläufe in Deutschland”. In a next step, a concept for a consistent visual appearance was developed over time, also taking into account the expertise of our two survey institutes. The concept includes standardized outlines for printing materials in various formats for the different target groups, for example, cover letters, flyers, brochures, and fixed templates for the various necessary official documents, such as busi-

ness letters, presentations, and posters. The corporate design concept also includes a key visual concept and a color concept for every substudy and starting cohort. The multi-locational network of the NEPS presents specific challenges as the central public relations team has to make sure that all members of the NEPS consortium use the different logos, which all members across Germany are obliged to use in the right way.

A common wording is another challenge we continually face. Participants receive several information documents, particularly when they are contacted for the first time, that include at least a cover letter, a privacy statement, and a declaration of consent. As there is some compulsory information, most cover letters are relatively long and difficult to read, which is, of course, not advantageous from the perspective of PR and not very motivating from the participants' perspective. An attractive wording and layout is thereby all the more important. As the very central operational unit of the LIfBi executing the NEPS involved in each study, the colleagues of the Survey Coordination check all materials that are used in the field with regard not only to content, but also to layout and wording, in order to ensure the NEPS standard and quality over time in all substudies.

5 Goals of PR Work Supporting a Longitudinal Survey Study—Conclusion

The overall aim of a target-specific public relations work against the backdrop of modern Western societies that support a longitudinal study is to ensure the target persons' willingness to participate over time and thus to keep panel attrition as low as possible. A necessary prerequisite for a good panel care system is knowing the target groups well (also apart from the study's main sample) and applying the measures of PR work accordingly. At the (function) system level, PR work aims to support the continuity of the survey concerning funding and acceptance. PR work in both fields means tailoring information to these target groups by translating the scientific content of large-scale survey assessments into something meaningful in Life World surroundings as well as transforming information to different codes for distinct function-systems. PR work creates codes of power or economy, of education or law. Content can also be offered in terms of group-identity in everyday life.

In a second step, this code and information have to be disseminated to increase the awareness and acceptance of the NEPS in the public. PR work here means being open and accessible to the public, being present on the World Wide Web, and publishing articles and interviews in daily and weekly newspapers, on the radio, and on TV. The uses of funding such a large-scale survey assessment for the political system are grounded in public awareness. The NEPS as a brand with a clearly defined corporate identity makes the survey participants part of something large and useful. The target persons' trust and commitment is deepened in all ranges of media-driven communication by representation on the World Wide Web and other (mass) media, by

tailoring and providing information leaflets and brochures, and by personal communication with participants via e-mail and telephone.

PR work not only helps to strengthen the acceptance of the general public and ensure the funding of research over years; it also paves the way to target persons for surveys, helps to avoid panel attrition in order to keep the sample representative over the years, and thus is a very important component of a longitudinal study.

References

- Beck, U. (1986). *Risikogesellschaft. Auf dem Weg in eine andere Moderne*. Frankfurt am Main: Suhrkamp.
- Bornschiefer, V. (2008). *Weltgesellschaft. Grundlegende soziale Wandlungen* (2nd ed.). Zürich: Loreto.
- Budowski, M., & Scherpenzeel, A. (2005). Encouraging and maintaining participation in household surveys: The case of the Swiss household panel. *ZUMA Nachrichten*, 29(56), 10–36.
- Dillman, D., Smyth, J., & Christian, L. (2008). *Internet, mail, and mixed-mode surveys: The tailored design method*. Hoboken, NJ: Wiley & Sons.
- Engel, U., Bartsch, S., Schnabel, C., & Vehre, H. (2012). *Wissenschaftliche Umfragen: Methoden und Fehlerquellen*. Frankfurt am Main, NJ: Campus.
- Giddens, A. (1990). *The consequence of modernity*. Cambridge: Polity.
- Glaser P. (2012). Respondents cooperation: Demographic profile of survey respondents and its implication. In Gideon L. (Ed.), *Handbook of Survey Methodology for the Social Sciences* (pp. 195–208). New York, Heidelberg, Dordrecht, London: Springer.
- Groves, R., & Couper, M. (2012). *Nonresponse in household interview surveys*. New York: Wiley & Sons.
- Habermas, J. (1981). *Theorie des kommunikativen Handelns* (Vol. 1-2). Frankfurt am Main: Suhrkamp.
- Krebs, D. (1986). Panelpflege—Eine Forschungsnotiz. *ZUMA Nachrichten*, 10(19), 76–80.
- Kuß, A. (2012). *Marktforschung: Grundlagen der Datenerhebung und Datenanalyse* (5th ed.) Wiesbaden: Gabler.
- Luhmann, N. (1998). *Die Gesellschaft der Gesellschaft* (Vol. 2). Frankfurt am Main: Suhrkamp.
- Marsden, P., & Wright, J. (Eds.) (2010). *Handbook of survey research* (2nd ed.) San Diego: Emerald.
- Maturana, U., & Varela, F. (1980). *Autopoiesis and cognition: The realization of the living*. Dordrecht: D. Reidel.
- Proner, H. (2011). *Ist keine Antwort auch eine Antwort? Die Teilnahme an politischen Umfragen*. Wiesbaden: VS Verlag für Sozialwissenschaften.
- Schütz, A. (1932). *Der sinnhafte Aufbau der sozialen Welt: Eine Einleitung in die verstehende Soziologie*. Wien: Springer.

- Stoop, I., Billiet, J., Koch, A., & Fitzgerald, R. (2010). *Improving survey response: Lessons learned from the European Social Survey*. Chichester: Wiley.
- Vehre, H. (2011). "Sie wollen mir doch was verkaufen!" *Analyse der Umfrageteilnahme in einem offline rekrutierten Access Panel*. Wiesbaden: VS Verlag für Sozialwissenschaften.

About the authors

J. Göpel
Leibniz Institute for Educational Trajectories (LifBi), Bamberg.

G. Lechner
Leibniz Institute for Educational Trajectories (LifBi), Bamberg.
e-mail: goetz.lechner@lifbi.de

A. Passmann
Leibniz Institute for Educational Trajectories (LifBi), Bamberg.

III. Longitudinal Measurement of Educational Processes: Surveys and Constructs

Video-Based Assessment and Rating of Parent-Child Interaction Within the National Educational Panel Study

Anja Sommer, Claudia Hachul and Hans-Günther Roßbach

Abstract

There is strong evidence that the learning opportunities offered in familial learning environments have a long-lasting impact on children's development and educational career. As one of only a few large-scale longitudinal studies, the National Educational Panel Study (NEPS) Starting Cohort 1—Newborns is taking up the challenge of direct assessment of parent-child interaction in familial learning environments. This article describes how this assessment was developed, comparing existing observational designs and instruments with regard to their large-scale practicability and utility for the NEPS. To gain reliable data on parent-child interaction, we apply the following procedure: (1) an overt, non-participant field observation of parent-child interaction embedded in a semi-standardized play situation, which is videotaped, and (2) an analysis of the videotaped parent-child interaction using a macroanalytic rating instrument adapted from the NICHD Study of Early Child Care and Youth Development (NICHD SECCYD). We illustrate the general practicability and reliability of this assessment with results from the first pilot study (N = 466). We point out potential pitfalls in implementing this approach by discussing the results of different in-depth analyses. Finally, we detail the resulting adaptations in the assessment and rating of parent-child interaction for the first main study (N = 3,481).

1 Parent-Child Interaction and its Importance for Child Development

The familial learning environment is of profound significance, especially in early life. Familial learning opportunities are most important at a very young age, and there is strong evidence that these opportunities have a long-lasting impact on child devel-

opment (Belsky et al., 2007; Blomeyer, Laucht, Pfeiffer, & Reuß, 2010; NICHD Early Child Care Research Network, 2002).

Structural characteristics of the familial learning environment, such as income and parental education, are often considered when associations between family characteristics and cognitive or social development in early life are studied (Halle et al., 2009; Hillemeier, Farkas, Morgan, Martin, & Maczuga, 2009). However, not only structural characteristics, but also educational processes such as parent-child interactions seem to play a key role in children's cognitive, linguistic, and socio-emotional development. Even if structural factors are controlled, associations between a child's development and educational processes remain significant (Belsky et al., 2007; Bornstein & Tamis-LeMonda, 1989; Bromley, 2009; Leerkes, Blankson, & O'Brien, 2009; NICHD Early Child Care Research Network, 2002; Page, Wilhelm, Gamble, & Card, 2010). Accordingly, an assessment of learning environments should not only consider structural characteristics, but also educational processes (Bäumer, Preis, Roßbach, Stecher, & Klieme, 2011).

A detailed look at these educational processes is offered through the observation of parent-child interactions. In these interactions, different factors, such as activating behavior, sensitivity, and responsiveness, have been found to be related to different aspects of later child development (Blomeyer et al., 2010; Leerkes et al., 2009; NICHD Early Child Care Research Network, 1998; Page et al., 2010). Therefore, the National Educational Panel Study (NEPS) assesses processes in familial learning environments beyond parent self-reports and observes parent-child interaction in the very first years of a child's life. In order to assess these aspects (in addition to others), the Newborn Cohort of the NEPS used a nationally representative sample of 3,481 children¹ born in Germany from March to August 2012 and follows these children longitudinally (Aßmann et al., 2011). In the first three years of the child's life, three measurement points at the age of 7, 16, and 26 months are given in the longitudinal study design of the NEPS.

2 Assessment of Parent-Child Interaction in Large-Scale Studies

Diverse methodological approaches can be applied for the assessment of parent-child interaction. To justify the choice of the methodological approach used in the NEPS, we discuss different observational designs and instruments regarding these approaches' large-scale practicability and utility for the NEPS.

¹ National Educational Panel Study (NEPS): Starting Cohort 1 – Newborns (SC1), doi:10.5157/NEPS:SC1:1.0.0

2.1 Observational Designs

For the classification of structured observation, Greve and Wentura (1997) distinguish different observational designs along several bipolar classifications: (1) overt vs. covert observation, (2) participant vs. non-participant observation, (3) laboratory vs. field observation, and (4) technically mediated vs. non-mediated observation. These designs differ with regard to their capability of assessing the targeted observational subject and regarding their large-scale practicability.

(1) Considering ethical correctness, an assessment of parent-child interaction has to be overt (Greve & Wentura, 1997). (2) Aiming at a standardized assessment of the interaction between parent and child, active participation of the observer is not constructive for the assessment of this dyadic situation. (3) Laboratory observation offers the opportunity to control framework better than field observation. However, with respect to the assessment of parent-child interaction, field observation in the natural home setting of the family may decrease reactive effects (Rentzsch & Schütz, 2009). (4) These days, most studies use video-mediated observation for the assessment of interactions. Because the assessment of an interaction sequence and rating parent-child interaction is separated, interviewers as well as raters are prevented from managing too many tasks simultaneously, which improves the quality of the assessment. Additionally, storage and repeatability of the data allow for consistent field monitoring and checking for quality via the possible application of several raters. Therefore, video-based observation is highly practicable for large-scale studies.

For these reasons, the Newborn Cohort of the NEPS applies an overt, non-participant field observation of parent-child interaction, which is videotaped. Therefore, the assessment is subdivided into the assessment of the interaction sequence and the subsequent rating of the parent-child interaction.

2.2 Observational Instruments

Level of observation

Different types of observational instruments can be applied based on the specifications of the observational design. Instruments for observational assessment can be classified as micro- and macroanalytic, differing in their level of observation.

Microanalytic instruments aim at specific aspects of interaction and focus mainly on the categorization or coding of frequency and the duration of behavior (Faßnacht, 1995; Greve & Wentura, 1997). Faßnacht (1995) distinguishes two microanalytic approaches: Event-sampling methods record every occurrence of a preselected behavioral pattern over a specific observational period. Time-sampling methods separate the stream of time into short, continuous, consecutive time sequences, often lasting for 5 to 10 seconds. Observers decide on the occurrence of predefined behavior with regard to each sequence following an all-or-nothing principle. Both microanalytic

approaches are rather time-consuming and are commonly used in small-scale studies (e. g., Bornstein & Tamis-LeMonda, 1989; Hirschmann, Kastner-Koller, Deimann, Aigner, & Svecz, 2011). In contrast, macroanalytic rating procedures have a high level of aggregation, downplaying minute contextual variability (Bornstein, Hahn, Suwalsky, & Haynes, 2011). They offer a rather global impression and capture characteristics and enduring traitlike features of individuals and are therefore commonly used for assessing intensity or behavior as a whole (Faßnacht, 1995). Time effectiveness and a broad global assessment of the targeted construct are highly important for large-scale studies. Therefore, like the majority of large-scale longitudinal studies, NEPS implements a macroanalytic instrument for the rating of the videotaped parent-child interactions. However, due to the videotaping, microanalytic approaches focusing on details of the mother-child interaction are applied later on.

Instruments

There seems to be no standard macroanalytic instrument for rating the parent-child interaction that fits different study designs and requirements. To detail the decision for the instrument used in the NEPS, we list existing instruments regarding the included constructs and aspects indicating large-scale practicability in Table 1. For this purpose, we used an overview of Wiefel et al. (2007), but for our purpose, we excluded instruments that do not aim at the age group under consideration (FIT-K98, a family- and kindergarten-interaction test, and Mahoney's Maternal Behavior Rating), or these instruments were used for psychiatric mother and baby units (BMIS, Bethlem Mother-Infant Interaction Scale). Additionally, we considered instruments used in foregoing birth-cohort studies (see Schlesiger, Lorenz, Weinert, Schneider, & Roßbach, 2011 for an overview).

Any instrument to be used in the NEPS has to meet the discussed methodological requirements concerning the observational design and the level of observation. Regarding observational design, all listed instruments are based on an overt, non-participant video-mediated observation. Additionally, all instruments offer the opportunity to rate interaction sequences that are videotaped in home settings. Regarding the level of observation, all listed instruments are classified as macroanalytic. Although macroanalytic instruments are usually time-efficient, some approaches are more time-consuming than others. Aiming towards a short duration of rating with a high-quality analysis and reliable data at the same time, time-consuming macroanalytic instruments, such as the Mannheim Rating System for Mother-Infant Face-to-Face Interaction (MBS-MKI-S) and the Nursing Child Assessment Teaching Scale (NCATS) (see Table 1), were excluded.

In addition to these aspects, large-scale practicability for the NEPS can also be discussed along two points: First, the burden of every assessment of interaction sequence should be kept as low as possible to avoid high rates of panel attrition. Because of time constraints, time spent in the home setting of the family should be kept as low as possible (Schlesiger et al., 2011). Therefore, instruments whose rating de-

depends on interactional sequences that exceed 10 minutes cannot be used (this applies to the Emotional Availability Scales, EA-III; see Table 1). Second, accessibility of the instrument has to be considered. Coding Interactive Behavior (CIB) and CARE-Index (CARE) (see Table 1) have not been published yet. EA-III and NCATS (see Table 1) are only accessible after an intensive training by the author or other licensed trainers, who are partly not located in Germany. For a large-scale study like the NEPS, rater training should instead be flexible in time and persons.

Therefore, we decided to adapt the instrument from the NICHD-SECCYD study (see Table 1) (NICHD Early Child Care Research Network, 1991). Large-scale practicability is fulfilled regarding the discussed points: The NICHD-SECCYD study uses technically mediated observation through video, and analyses are conducted using a macroanalytic rating instrument, which can easily be taught and applied. Furthermore, this method is time-effective because the instrument is designed for rating short video-sequences not exceeding 10 minutes. Additionally, the NICHD-SECCYD study has reported good-quality indicators regarding internal consistency, reliability, concurrent validity, and predictive validity, which are illustrated in the examples below.

The NICHD Child Care Research Network (2005) reports an internal consistency and inter-rater-reliability for the sensitivity composite (subsuming three items; see also Section 4.2) indicated by Cronbach's alpha ($\alpha = .75$.) and Pearson's correlation coefficient ($r = .78$, $p = \text{n. a.}$) (Bland, Batten, Appelbaum, Wendell, & NICHD Early Child Care Research Network, 1995). Additionally, the correlation of the sensitivity composite with a positive parenting subscale of the Home Observation for Measurement of the Environment (HOME) Inventory ($r = .34$, $p < .0001$) indicates concurrent validity (Bland, Appelbaum, Batten, Wendell, & NICHD Early Child Care Research Network, 1994). In addition, the correlation of the sensitivity composite (averaged repeated measures for 6, 15, and 24 months) with different child outcomes at 36 months signals predictive validity (school readiness: $r = .37$, $p < .001$; receptive vocabulary: $r = .52$, $p < .001$; social competence: $r = .27$, $p < .01$) (NICHD Early Child Care Research Network, 1998). For further impacts on child development, see also NICHD Early Child Care Research Network (1999; 2005).

Table 1 Overview of Observational Instruments for Rating of Parent-Child Interaction

Instrument title	Scales/domains	Estimated duration in min. (interaction/rating)	Published	Training applicable for large-scale studies	Video-taping intended	References
CARE <i>CARE-Index, infant version</i>	7 scales <i>Parental and child behavior:</i> Facial expression, vocal expression, position and body contact, expression of affect, pacing of turns, control of the activity, developmental appropriateness of the activity	3/10	–	–	✓	Crittenden (2004, 2005); Letourneau & Tryphonopoulos (2012)
CIB <i>Coding Interactive Behavior</i>	43 items, divided into 6 domains <i>Parental behavior:</i> Sensitivity and responsiveness; intrusiveness <i>Child behavior:</i> Positive affect, negative emotionality, initiation, and involvement <i>Dyad:</i> Dyadic reciprocity	5/n.a.	–	✓	✓	Feldman (1998); Feldman, Weller, Sirota, and Eidelman (2003)
EA-III <i>Emotional Availability Scales (3rd edition)</i>	6 scales <i>Parental behavior:</i> Sensitivity, structuring, nonintrusiveness, nonhostility <i>Child behavior:</i> Responsiveness to adult, involvement of adult	20/n.a.	–	–	✓	Biringen (1998); Strauß & Schuhmacher (2005)
NCATS <i>Nursing Child Assessment Teaching Scale</i>	73 items divided into 6 domains <i>Parental behavior:</i> Sensitivity to child's cues, response to child's distress, social-emotional growth-fostering behavior, cognitive growth-fostering behavior <i>Child behavior:</i> Child's clarity of cues, responsiveness to the mother	6/n.a.	✓	–	✓	Gross, Conrad, Fogg, Willis, & Garvey (1993); Harrison, Magill-Evans, & Sadoway (2001); Sumner & Spletz (1995)

Instrument title	Scales/domains	Estimated duration in min. (interaction/rating)	Published	Training applicable for large-scale studies	Video-taping intended	References
NICHD-SECCYD <i>Rating instrument of mother-child interaction—semi-structured procedure six months home visit</i>	13 scales <i>Parental behavior:</i> Sensitivity to distress, sensitivity to nondistress, intrusiveness, detachment, stimulation of development, positive regard for the child, negative regard for the child, flatness of affect <i>Child behavior:</i> Positive mood, negative mood, activity level, sociability, sustained attention	10/10	✓	✓	✓	NICHD Early Child Care Research Network (1991)
MBS-MKI-S <i>Mannheim Rating system for mother-infant face-to-face interaction</i>	13 scales <i>Parental behavior:</i> Emotion, affectionateness, vocalization, verbal restriction, congruence, variability, reactivity/sensitivity, stimulation <i>Child behavior:</i> Emotion/facial expression, vocalization, viewing direction, reactivity, willingness to interact	10/(rating interval 1 min)	✓	✓	✓	Blomeyer et al. (2010); Esser, Scheven, Petrova, Laucht, & Schmidt (1989)

3 Assessing Parent-Child Interaction in the Early Childhood Cohort of the NEPS

After discussing the reasons for the selected design of the assessment and rating of the parent-child interaction, we now specify the form and organization of videotaping and rating the parent-child interaction in the NEPS's Newborn Cohort.

3.1 Assessment of Interaction Sequences

Videotaping of the interaction between the parent (primarily the mother) and his or her child is embedded in a personal interview in the home setting of the family. Parent-child interaction takes place in a semi-standardized play situation. Standardization covers (1) place, (2) play material, and (3) frame of the play situation but does not include strict instruction for interaction. Therefore, the parent is asked to interact with the child as usual. (1) The play situation is carried out on a blanket on the floor, which only serves as a visual localization of the play situation to support the interviewer (for the focus of the camera). (2) Play material included in the NEPS toy set had to meet different criteria regarding type and quality. Considering their type, toys were selected that aimed towards a specific goal of action outcome at different levels of difficulty. As Heckhausen and Heckhausen (2010) point out, some goals of action are more difficult because they demand higher levels of the child's activity regulation than others: Sudden-discrete effects get attention easily and are therefore attractive goals of action for very young children (e.g., squeezing a toy). Continuous effects, which are in conjunction with the action (e.g., the rattle of a car moved back and forth), demand a higher level of self-regulation, whereas stateful goals of a chain of activities are highly demanding because they appear only at the end of an activity (e.g., a finished tower of stacking cups). We selected the number and type of play materials aligned to children's age along this classification of effects (see Table 2). Additionally, we completed this compilation with toys evoking symbolic play and joint-attention episodes. Moreover, the quality of the toys has to be given: First, they had to be age-appropriate (resistant to saliva, not have small parts that can be swallowed); second, they had to offer a seal of quality; and third, they had to be easy to clean with disinfectant wipes because interviewers used the same toys for different households. As in the NICHD-SECCYD study, the framing of the play situation is adapted to the changing requirements of young children throughout their development. The frame of the play situation differs slightly from Wave 1 to Waves 2 and 3. In Wave 1, mothers were asked to play with their infants with five toys of their own for 3 minutes, then for another 5 minutes with toys from the NEPS toy set (see Table 2). In Waves 2 and 3, the observation procedure followed a three-bag procedure in which mothers were asked to play with their children for 10 minutes with toys divided into three bags in a set order (NICHD Early Child Care Research Network, 2005).

Table 2 Play Material for Parent-Child Interaction

	Age of child in months	Sudden-discrete effects	Continuous effects in conjunction with the action	Stateful goal of chain of activities	Symbolic play	Joint attention
<i>Wave 1</i>	7	Rattle, squeaking book	Duckling Ball	Stacking cups		
<i>Wave 2</i>	16	Squeaking animal	Rattling car	Stacking cups, sorting box	Plates, spoons	Book
<i>Wave 3</i>	26	Xylophone	Rattling car	Puzzle	Plates, spoons, animals	Book

The administration of the assessment is conducted by female interviewers in order to provide easier access to the homes of mothers and their 7-month-olds. Interviewer training was provided over several days, focusing on the requirements of the target group and correct assessment.

3.2 Rating of Parent-Child Interaction

Based on the videotaped interaction sequences, the rating of parent-child interaction is conducted by trained coders. Videos of parent-child interaction are delivered to the NEPS and stored in a special room in which access is strictly regulated according to NEPS data-protection standards.

As already described in Section 2, a macroanalytic rating instrument of the NICHD-SECCYD study was chosen for rating the parent-child interaction, which is shown in Table 1. The instrument covers parental and filial interaction style, which can be rated on a 4-point scale ranging from not-at-all characteristic to highly characteristic, supplemented by one missing category. We translated the English version into German, added additional examples for different scale points, and tested the instrument in a feasibility study ($n = 20$). As in the NICHD-SECCYD study, the raters in the NEPS rate all items after viewing five minutes (or 10 minutes for Waves 2 and 3) of videotaped parent-child interaction. Because of great demands of a highly inferent rating instrument, raters in the NEPS were trained extensively during a 50-hour rater training.

4 Results of Pilot Study Wave 1

In order to provide high-quality data, assessment of videos of interaction sequences in the field and the rating of parent-child interactions are tested extensively at the outset of every main study with the help of diverse pre-studies. For illustration, we concentrate on results of the first pilot study, which took place from October 2011 to January 2012. This pilot study led to different adjustments for the assessment and rating of parent-child interaction in the first main study, with a field time lasting from August 2012 to February 2013.

4.1 Results of Assessment of Interaction Sequences

466 interviews could be realized. The acceptance for participation in videotaped parent-child interaction was very high: 422 participants gave their written consent to be videotaped (90 %); after completion of a videotaped test for competencies, videotaping for parent-child interaction began in 376 parent-infant dyads (80 %) and was completed in 360 cases (77 %). Finally, 170 cases could be analyzed regarding aspects of parent-child interaction (190 interaction sequences were discarded due to different assessment faults, which partly occurred in the same cases). Misframed videos (141 cases) and/or an unfavorable camera setup and location of the play situation (75 cases) constituted the main assessment faults. In most cases, this resulted in videos in which the head or face of the mother or child was not visible for a significant amount of time. Thus, a valid analysis of interactional behavior that also covers facial expressions could no longer be given. Other types of faults in the assessment included an incorrect execution of the play situation, for example, when the mother and child played on a table or a couch instead of on a blanket on the floor (17 cases); when the relevant interactional sequence lay significantly below time limit needed for valid analysis (13 cases); and when technical faults occurred, such as the failure to record sound (7 cases). The reasons for the types of faults were twofold: First, interviewers had to adapt to differing framing conditions. In some cases, home settings were too small for administering the standard setup of the play situation and camera. Second, differences in the performance of the interviewers were noticeable. Presumably based on their previous technical knowledge, some interviewers administered the assessment of their cases nearly free of fault, whereas single interviewers failed at the assessment in a majority of their cases.

4.2 Results of Rating the Parent-Child Interaction

Rating the parent-child interaction was based on the 170 analyzable videotaped play situations. Rating was conducted by two raters (47 % of the videos were coded by

Table 3 Rater Agreement on Composites Level

	Sensitivity composite ^a	Detachment composite ^b
Joint probability in %	52	71
κ	.39***	.41***
ICC ^c	.76***	.53***
r	.76***	.55**

a Includes items: sensitivity to nondistress, positive regard for the child, intrusiveness (reversed score).

b Includes items: detachment, flatness of affect (recoded from 4-point to binary scale).

c Two-way random, nonadjusted; $n = 31$.

+: $p < .1$; *: $p < .05$; **: $p < .01$; ***: $p < .001$

Rater 1; 53 % by Rater 2). To check for inter-rater agreement, a double rating of 18 % of the cases was established ($n = 31$). To provide comparability with the NICHD-SECCYD study, we calculated the same composites of items: a sensitivity composite and a detachment composite (see Table 3). Although we constructed the scale for rating parent-child interactions so as it possesses equal intervals, the scale is in a conservative sense on an ordinal level of measurement. Therefore, we report parametric as well as non-parametric statistics in Table 3. Joint probability and Cohens Kappa indicate a rather poor agreement, with values of 52 % and 71 % as well as $\kappa = .39$ and $\kappa = .41$ (both $p < .001$), respectively² (for details, see Table 3). Intra-Class-Correlation (ICC) and Pearson's r , show moderate to good agreement, with values of ICC = .76 and ICC = .53 (both $p < .001$) and $r = .76$, $p < .001$ and $r = .55$, $p < .01$, respectively. Results of Pearson's r are comparable with the findings of NICHD-SECCYD, which reports $r = .78$ for the sensitivity composite and $r = .69$ for the detachment composite (Bland et al., 1995).

To enhance the quality of the rating instrument beyond the given results, the identification of the possible clarification of items was necessary. Therefore, a more precise look at rater agreement was required. As Uebersax (2010) points out, disagreement should be treated as a construct that can be subdivided into different components. Accordingly, an index reporting the different components simultaneously in one numerical value, such as the ICC, is not useful for identifying steps to improve agreement. Components of disagreement are mainly based on two different sources: differences between the raters in their trait definition or their definition of specific rating levels (Uebersax, 2010). As a consequence, item-level analysis is conducted for different components of disagreement regarding rater association, rating distribution, and rater bias.

2 All calculations were conducted with SPSS IBM Statistics 19.

For an indication of rater association, we conducted a simple Pearson correlation at the item level. While a majority of items show values between $r = .68$ and $r = 1$ (for p-values, see Table 4), indicating a good agreement, rater agreement for two items is rather poor, with values below $r = .20$. In addition to items with clear trait definition, evidence for a different interpretation of basic constructs or differences in the weight of trait factors is given for single items.

Disagreement can also be based on raters' differences in the definition of rating categories. A test for marginal homogeneity is used for the examination. Marginal homogeneity reflects the similarity of frequencies with which two raters use various rating categories (Uebersax, 2010). Therefore, we included all rated cases and compared marginal frequencies using a Pearson chi-square test. The significance of a single Pearson chi-square test indicates that the rater and distribution are significantly related, which implies differences in frequencies in the use of each rating category. The significance of Pearson's chi-square test is evident for the majority of items, although it only indicates a moderate significance (see Table 5). Therefore, the definition of rating levels should be clarified for the majority of items.

Finally, differences in the interpretation of the calibration of the rating scale could result in disagreement. In addition to other methods, we also tested the tendency to make generally higher or lower ratings with a t-test. Rater bias is displayed by results indicating significant differences between the means of the raters. The majority of items are not biased; only two items display a significant t-test at a 5 %-level (see Table 6).

5 Adaptations and Consequences for Upcoming Waves

Having pointed out potential pitfalls in the implementation of the assessment and rating of parent-child interaction in the first pilot study of NEPS Starting Cohort 1—Early Childhood, we now detail the resulting adaptations regarding the assessment of interaction sequences and the rating of parent-child interaction for the first main study.

First of all, the quality of the videos was enhanced. Adaptions covered interviewer training, the selection of interviewers, and supporting material for interviewers. Interviewer training for video-based assessment was expanded extensively, and an additional hands-on training was established. Furthermore, as interviewers differed in their number of faults in data collection, the selection of interviewers is now conducted based on a test-assessment. Additionally, the interviewer manual has been adapted, and a short pictorial instruction is now also handed out to support the assessment process. The implementation of these adjustments was also conducted for the pilot study assessment in Wave 2: Here, only two interaction sequences were distorted due to faults in assessment ($n = 64$).

Table 4 Rater Association on Item-Level; Pearson's *r*

	M1	M2	M3	M4	M5	M6	M7	M8	C1	C2	C3	C4	C5
<i>r</i>	1.00	.47**	.74**	.70***	.48**	.68***	-	.48**	.20	.74***	.74***	.43*	.08

Parental behavior: sensitivity to distress (M1), sensitivity to nondistress (M2), intrusiveness (M3), detachment (M4), stimulation of development (M5), positive regard for the child (M6), negative regard for the child (M7), flatness of affect (M8); child behavior: positive mood (C1), negative mood (C2), activity level (C3), sociability (C4), sustained attention (C5); *n* = 31; for M1, *n* = 2 (M1 can only be rated if child displays distress); +: *p* < .1; *: *p* < .05; **: *p* < .01; ***: *p* < .001

Table 5 Rating Distribution on Item-Level, Chi-Square Test

	M1	M2	M3	M4	M5	M6	M7	M8	C1	C2	C3	C4	C5
χ^2	3.75	6.01 ⁺	8.66*	6.17*	7.08 ⁺	10.46**	2.33	8.06 ⁺	8.35 ⁺	9.97*	2.39	11.00**	3.84
<i>df</i>	3	2	2	2	3	3	1	3	3	3	2	3	2

Parental behavior: sensitivity to distress (M1), sensitivity to nondistress (M2), intrusiveness (M3), detachment (M4), stimulation of development (M5), positive regard for the child (M6), negative regard for the child (M7), flatness of affect (M8); child behavior: positive mood (C1), negative mood (C2), activity level (C3), sociability (C4), sustained attention (C5); differences in degrees of freedom result from single scale levels not being used. *n* = 170; for M1 *n* = 10. (M1 can only be rated if child displays distress); +: *p* < .1; *: *p* < .05; **: *p* < .01

Table 6 Rater Bias on Item Level; t-Test

		M1	M2	M3	M4	M5	M6	M7	M8	C1	C2	C3	C4	C5
Rater 1	<i>M</i>	2.50	3.16	1.42	1.48	2.58	3.23	1.00	1.84	2.10	1.35	2.29	2.71	2.52
	<i>SD</i>	2.12	0.69	0.67	0.68	0.72	0.67	0.00	0.74	0.54	0.80	0.53	0.69	0.57
Rater 2	<i>M</i>	3.00	3.16	1.26	1.26	2.45	3.10	1.00	2.23	2.06	1.16	2.23	2.77	2.58
	<i>SD</i>	1.41	0.64	0.51	0.58	0.57	0.54	0.00	0.76	0.25	0.52	0.43	0.56	0.50
	<i>t</i>	-1.00 ⁺	0	1.98 ⁺	2.53*	1.10	1.44	-	-2.83**	0.33	1.99	1.00	-0.53	-0.49
	<i>df</i>	1	30	30	30	30	30	30	30	30	30	30	30	30

Parental behavior: sensitivity to distress (M1), sensitivity to nondistress (M2), intrusiveness (M3), detachment (M4), stimulation of development (M5), positive regard for the child (M6), negative regard for the child (M7), flatness of affect (M8); child behavior: positive mood (C1), negative mood (C2), activity level (C3), sociability (C4), sustained attention (C5); *n* = 31; for M1 *n* = 2 (M1 can only be rated if child displays distress); +: *p* < .1; *: *p* < .05; **: *p* < .01

The rating of parent-child interaction was modified regarding both the instrument itself and rating processes. The rating manual of the instrument was restructured. While the definition of trait and construct remained unmodified, the structure of each item and item-level description were unified. Additionally, we accommodated the fact that the trait is more continuous than discrete and expanded the rating scale from four to five levels, thereby providing a detailed description and example for each item level. These adaptations resulted in an adapted version of the rating instrument from the NICHD-SECCYD study, which is used for rating videotaped interactions in the first main study (Sommer & Mann, 2015).

Second, the rating process itself was adjusted: To avoid observer drift, the duration of the period of rating is kept as low as possible, and regular refreshment-trainings during the rating are conducted in addition to the rater training.

For the implementation of a video-based assessment and the rating of parent-child interaction in large-scale studies, different challenges had to be faced. After testing the assessment and rating in first pilot study in Wave 1 and identifying potential pitfalls, we made different adjustments for the main study in Wave 1 and subsequent waves. By assessing and rating parent-child interaction in the first main study ($N = 3,481$), the NEPS will gather substantiate information about educationally relevant processes in familial learning environments. The data were released in 2015 in a Scientific Use File.

References

- Aßmann, C., Steinhauer, H. W., Kiesl, H., Koch, S., Schönberger, B., Müller-Kuller, A., ... Blossfeld, H.-P. (2011). Sampling designs of the National Educational Panel Study: Challenges and solutions. In H.-P. Blossfeld, H.-G. Roßbach, & J. von Maurice (Eds.), *Zeitschrift für Erziehungswissenschaft, 14. Education as a lifelong process: The German National Educational Panel Study (NEPS)* (pp. 51–65). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Bäumer, T., Preis, N., Roßbach, H.-G., Stecher, L., & Klieme, E. (2011). Education processes in life-course-specific learning environments. In H.-P. Blossfeld, H.-G. Roßbach, & J. von Maurice (Eds.), *Zeitschrift für Erziehungswissenschaft, 14. Education as a lifelong process: The German National Educational Panel Study (NEPS)* (pp. 87–101). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Belsky, J., Vandell, D., Burchinal, M., Clarke-Stewart, K. A., McCartney, K., Owen, M., & The NICHD Early Child Care Research Network. (2007). Are there long-term effects of early child care? *Child Development, 78*(2), 681–701.
- Biringen, Z. (1998). *Emotional Availability Scales* (3rd ed.). Unpublished manual, Center for family studies, Colorado State University, Fort Collins.
- Bland, S., Appelbaum, M., Batten, D. A., Wendell, C., & NICHD Early Child Care Research Network. (1994). *Correlations between HOME-HOME and structured interac-*

- tion variables* (Child Care Data Report—8b). Nashville: Peabody College, Vanderbilt University.
- Bland, S., Batten, D. A., Appelbaum, M., Wendell, C., & NICHD Early Child Care Research Network. (1995). *Mother-child-interaction—Mother variables. Six months reliabilities* (Child care data addendum—8c). Nashville: Peabody College, Vanderbilt University.
- Blomeyer, D., Laucht, M., Pfeiffer, F., & Reuß, K. (2010). *Mutter-Kind-Interaktion im Säuglingsalter, Familienumgebung und Entwicklung früher kognitiver und nicht-kognitiver Fähigkeiten: Eine prospektive Studie*. Mannheim: Zentrum für Europäische Wirtschaftsforschung.
- Bornstein M. H., Hahn C.-S., Suwalsky J. T., & Haynes, O. M. (2011). Maternal and infant behavior and context associations with mutual emotion availability. *Infant Mental Health Journal*, 32(1), 70–94.
- Bornstein, M. H., & Tamis-LeMonda, C. S. (1989). Maternal responsiveness and cognitive development in children. In M. H. Bornstein (Ed.), *Maternal responsiveness: Characteristics and consequences. New directions for child development* (pp. 49–61). San Francisco, CA: Jossey-Bass.
- Bromley, C. (2009). *Growing up in Scotland: The impact of children's early activities on cognitive development*. Edinburgh: Scottish Government.
- Crittenden, P. M. (2004). *CARE-Index: Coding manual*. Unpublished manual.
- Crittenden, P. M. (2005). Der CARE-Index als Hilfsmittel für Früherkennung, Intervention und Forschung. *Frühförderung Interdisziplinär*, 24(3), 99–106.
- Esser, G., Scheven, A., Petrova, A., Laucht, M., & Schmidt, M. H. (1989). Mannheimer Beurteilungsskala zur Erfassung der Mutter-Kind-Interaktion im Säuglingsalter (MBS-MKI-S). *Zeitschrift für Kinder- und Jugendpsychiatrie*, 17, 185–193.
- Faßnacht, G. (1995). *Systematische Verhaltensbeobachtung. Eine Einführung in die Methodologie und Praxis*. München: Ernst Reinhardt Verlag.
- Feldman, R. (1998). *Coding interactive behavior manual*. Unpublished manual, Bar-Ilan University, Ramat-Gan, Israel.
- Feldman, R., Weller, A., Sirota, L., & Eidelman, A. I. (2003). Testing a family intervention hypothesis: The contribution of mother—infant skin-to-skin contact (kangaroo care) to family interaction, proximity, and touch. *Journal of Family Psychology*, 17(1), 94–107.
- Greve, W., & Wentura, D. (1997). *Wissenschaftliche Beobachtung. Eine Einführung*. Weinheim: Beltz Psychologie Verlags Union.
- Gross, D., Conrad, B., Fogg, L., Willis, L., & Garvey, C. (1993). What does the NCATS measure. *Nursing Research*, 42(5), 260–265.
- Halle, T., Forry, N., Hair, E., Perper, K., Wnadner, L., Wessel, J., & Vick, J. (2009). *Disparities in early language development: Lessons from the Early Childhood Longitudinal Study—Birth Cohort (ECLS-B)*. Washington, DC: Child Trends.
- Harrison, M., Magill-Evans, J., & Sadoway, D. (2001). Scores on the Nursing Child Assessment Teaching Scale for father-toddler dyads. *Public Health Nursing*, 18(2), 94–100.

- Heckhausen, J., & Heckhausen, H. (2010). Motivation und Entwicklung. In J. Heckhausen, & H. Heckhausen (Eds.), *Motivation und Handeln* (pp. 427–488). Berlin: Springer.
- Hillemeier, M. M., Farkas, G., Morgan, P. L., Martin, M. A., & Maczuga, S. A. (2009). Disparities in the prevalence of cognitive delay: How early do they appear? *Paediatric and Perinatal Epidemiology*, *23*(3), 186–198.
- Hirschmann, N., Kastner-Koller, U., Deimann, P., Aigner, N., & Svecz, T. (2011). INTAKT: A new instrument for assessing the quality of mother-child interactions. *Psychological Test and Assessment Modeling*, *53*(3), 295–311.
- Leerkes, E. M., Blankson, A. N., & O'Brien, M. (2009). Differential effects of maternal sensitivity to infant distress and nondistress on social-emotional functioning. *Child Development*, *80*(3), 762–775.
- Letourneau, N., & Tryphonopoulos, P. (2012). Der CARE-Index: Ein Instrument zur Erfassung der Beziehungsqualität zwischen Bezugsperson und Kind ab der Geburt. In M. Stokowy, & N. Sahhar (Eds.), *Bindung und Gefahr. Das Dynamische Reifungsmodell der Bindung und Anpassung* (pp. 19–32). Gießen: Psychosozial-Verlag.
- NICHD Early Child Care Research Network. (1991). *NICHD study of early child care: Volume II: 5 month manual, 6 month manuals, time use manuals*. Unpublished manuscript, NICHD Study of Early Child Care.
- NICHD Early Child Care Research Network. (1998). Relations between family predictors and child outcomes: Are they weaker for children in child care. *Developmental Psychology*, *34*(5), 1119–1128.
- NICHD Early Child Care Research Network. (1999). Child care and mother—child interaction in the first 3 years of life. *Developmental Psychology*, *35*(6), 1399–1413.
- NICHD Early Child Care Research Network. (2002). Child care structure—Process—Outcome: Direct and indirect effects of child care quality on young children's development. *Psychological Science*, *13*(3), 199–206.
- NICHD Early Child Care Research Network. (2005). *Child care and child development. Results from the NICHD Study of Early Child Care and Youth Development*. New York: Guilford.
- Page, M., Wilhelm, M. S., Gamble, W. C., & Card, N. A. (2010). A comparison of maternal sensitivity and verbal stimulation as unique predictors of infant social-emotional and cognitive development. *Infant Behavior and Development*, *33*(1), 101–110.
- Rentsch, K., & Schütz, A. (2009). *Psychologische Diagnostik. Grundlagen und Anwendungsperspektiven*. Stuttgart: Kohlhammer.
- Schlesiger C., Lorenz, J., Weinert, S., Schneider, T., & Roßbach, H.-G. (2011). From birth to early child care. In H.-P. Blossfeld, H.-G. Roßbach, & J. von Maurice (Eds.), *Zeitschrift für Erziehungswissenschaft*, *14. Education as a lifelong process: The German National Educational Panel Study (NEPS)* (pp. 187–202). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Sommer, A., & Mann, D. (2015). *Qualität elterlichen Interaktionsverhaltens. Erfassung von Eltern-Kind-Interaktionen mithilfe eines makroanalytischen Ratinginstruments im Na-*

- tionalen Bildungspanel*. (NEPS Working Paper No. 56). Bamberg: University of Bamberg, National Educational Panel Study.
- Strauß, B., & Schuhmacher, J. (Eds.). (2005). *Klinische Interviews und Ratingskalen*. Göttingen: Hogrefe.
- Sumner, G., & Spietz, A.L. (1995). *NCAST caregiver/parent-child interaction teaching manual* (2nd ed.). Seattle, WA: NCAST Publications, University of Washington.
- Uebersax, J. (2010). *Statistical methods for rater and diagnostic agreement*. Retrieved from <http://john-uebersax.com/stat/agree.htm>
- Wiefel, A., Titze, K., Kuntze, L., Winter, M., Seither, C., Witte, B., ... Lehmkuhl, U. (2007). Diagnostik und Klassifikation von Verhaltensauffälligkeiten bei Säuglingen und Kleinkindern von 0–5 Jahren. *Praxis der Kinderpsychologie und Kinderpsychiatrie*, 56(1), 59–81.

About the authors

C. Hachul
University of Bamberg, Bamberg.

H.-G. Roßbach
Leibniz Institute for Educational Trajectories (LifBi), Bamberg.
Chair of Early Childhood Education, University of Bamberg, Bamberg.

A. Sommer
Chair of Early Childhood Education, University of Bamberg, Bamberg.
e-mail: anja.sommer@uni-bamberg.de

Measuring Personality Traits of Young Children—Results From a NEPS Pilot Study

Doreen Müller, Tobias Linberg, Michael Bayer, Thorsten Schneider
and Florian Wohlkinger

Abstract

Measuring the Big Five personality traits is part of the research program in different starting cohorts of the National Educational Panel Study (NEPS). The Big Five are usually measured through self-ratings via self-administered questionnaires. However, children of preschool age cannot easily report on their self-concept in a sufficient way, even when more extensive research settings are applied. Studies using parental and teacher ratings show that the Big Five can capture individual differences in the behavioral tendencies of children (Digman, 1990; Mervielde, 2005; Weinert, Asendorpf, Beelmann, Doil, & Frevert, 2007), but there are no short survey versions of the Big Five for parental ratings that are done via telephone interviewing. In order to obtain data on the Big Five of five-year-old children in the NEPS, we used a bipolar 10-item scale and asked parents and Kindergarten teachers to rate the children. Since the parents answered questions in computer-assisted telephone interviews (CATI), we adapted the items to this mode of surveying. In order to gather information on measuring the Big Five in this bipolar rating scale via telephone interviews, we conducted cognitive interviews with $n = 15$ parents and then tested two different versions within a split-half design in a pilot study (with total $n = 89$ parents). This paper presents results from cognitive interviews on parents' abilities to rate their children's behavior in this way. We compare the results of the two versions applied in the pilot study as well as the ratings of Kindergarten teachers and parents. Finally, our paper draws conclusions on the measurement of personality traits of young children within the NEPS.

1 Introduction

Measuring personality at a younger age presents some challenges that survey studies have to deal with in a systematic manner. First of all, in multi-thematic, large-scale studies, there are some limitations in measuring the personality of young children. This is not because children are principally unable to report on their personalities, as Measelle, John, Ablow, Cowan, and Cowan (2005) showed in a profound manner, but rather because there are age limitations when using very short instruments required by large-scale studies because of limited time resources. The personality of young children develops alongside their context-specific experiences, which they gain both in the family and in the institutional context.

Consequently, we decided to use external assessments from two different sources. Since all of the target children attended Kindergartens, we asked both the educators as well as the parents. Parents are able to rate the personality of their young children in a distinct and replicable way to reproduce the five factors of personality (Kohnstamm, Geldolph, Mervielde, Besevegis, & Halverson, 2005), which together form one of the most influential models in personality research (McCrae & Costa, 1987). To measure the manifestations of the specific factors, we used the “Fünf Faktoren Fragebogen für Kinder—Kurzform” (FFFK-K, Five Factor Questionnaire for Children—Short Form) instrument developed by Asendorpf (Weinert et al., 2007), which represents a short and age-adapted version of the “Big Five bipolar adjective scales” that Asendorpf and van Aken (2003) developed and used in a longitudinal study on the validity of personality judgments in childhood.

1.1 A Short Version of the Big Five Bipolar Adjective Scales

Bipolar adjective scales are, in accordance with the lexical tradition, one of the most prominent approaches for measuring personality traits and their development (Digman, 1990). As Hofstee (2003) states, “The lexical approach reflects and fosters a lay definition of personality” (p. 235). Mervielde, Buyst, and de Fruyt (1995) argue that the five factors “are also a major component of teachers’ and parents’ natural language discourse on children” (p. 532). This means that teachers’ and parents’ everyday understanding of children’s personality traits corresponds to the adjectives used in research. Based on this long-lasting tradition of analyzing lay personality descriptions and developing practicable and reliable measures, Asendorpf’s short version of the Five-Factor Model—the FFFK-K—was of high interest for us. Table 1 presents an overview of the bipolar adjective pairs and their respective assignment to the personality dimensions.

Ratings of personality traits are usually measured using paper-and-pencil questionnaires so that the respondents have a good impression of the bipolarity of the scale and the range of the possible assessments. In our study, we used two different

Table 1 Items and Scales of Parent/Educator Bipolar Adjective Big Five Instrument

	Items	Scale
A	Talkative/quiet	Extraversion
B	Disorderly/orderly	Conscientiousness
C	Good-Natured/touchy	Agreeableness
D	Uninterested/interested	Openness/intellect
E	Self-Assertive/insecure	Neuroticism
F	Withdrawn/sociable	Extraversion
G	Focused/distractable	Conscientiousness
H	Stubborn/gentle	Agreeableness
I	Quick/slow	Openness/intellect
J	Worried/calm	Neuroticism

Note. Source: Weinert et al. (2007).

modes. The educator ratings were collected via paper-and-pencil questionnaire in which we used the original instrument of the FFFK-K, and the two items of each dimension are presented in a reverse manner (e.g., ‘talkative/quiet’ and ‘withdrawn/sociable’ for the dimension of extraversion). The parent-ratings, on the other hand, were collected via telephone interview, which is very challenging, especially with respect to the problem of the desirability of the traits.

1.2 Bipolar-Adjective Scales in a Telephone Interview

Bipolar adjective scales confront respondents with high requirements. These respondents have to decide (1) which of the two adjectives describes the person appropriately and (2) to what degree. While the educator questionnaire is a paper-and-pencil version, the parent-ratings are collected via telephone interviews. Since there is very little precedent for measuring children’s personality traits via telephone interview, our first research question is:

- 1) Can bipolar rating scales be used in a telephone interview?

This question implies the analyses and descriptions of necessary strategies of adaptation (see Sections 2 and 3). Despite this fundamental clarification of the possibilities and adaptations when using bipolar rating scales within telephone interviews, we also

look at the consequences of the simultaneous measurement of personality characteristics when two different modes of data collection (paper-and-pencil and telephone interviewing) are applied. Our second research question is therefore:

- 2) Are there any consequences of adapting the parent-ratings with regard to the differences between the ratings of different evaluators?

With this question, we focus on the comparison between parent and educator ratings with two different versions of the FFFK-K for two different populations (see Section 4).

2 Evaluation of Two-Scale Versions With Cognitive Interviews

2.1 Methodology and Data

To answer the question of whether bipolar rating scales can be used in a telephone interview, we decided to conduct cognitive interviews, which offer possibilities to identify problems in the respondents' comprehension of the questions and to assess their cognitive effort in answering the questions (for more detailed information, see Prüfer & Rexroth, 2005; Schlechter, Blair, & Vande Hey, 1996; Wallis, 2005).

For our cognitive interviews, we used two different techniques: *paraphrasing* (reflecting questions in own words) and *general probing* (naming potential problems). The interviews were conducted via phone by trained interviewers. Each interview was taped for later analyses, and the interviewers took notes on the respondents' answers according to our protocol during the interviews. The cognitive interviews were conducted between October and December 2011.

The sample of our interviews was recruited in Kindergartens and comprises 15 parents (with different social backgrounds) within three German federal states.

2.2 Results and Consequences

All parents received the same standardized questions to measure children's personality traits. The introduction to the question is: "The following oppositional characteristics intend to assess the fit with [target's name]'s characteristics. To grade the characteristic's fit, you can use numbers from 0 to 10. Lower numbers indicate a fit with the first characteristic, higher numbers a fit with the second"¹ (see Table 1).

1 Original: "Bei den folgenden gegensätzlichen Eigenschaften sollen Sie angeben, welche Eigenschaften eher auf [Name des Zielkinds] zutreffen. Wie stark die Eigenschaften zutreffen, können Sie

The respondents were asked to replicate this introduction in their own words via the cognitive technique of paraphrasing. They were also asked to explain their rating for the first pair of adjectives. Only two respondents could not reproduce the introduction properly. They did not understand the bipolar scale at all and constantly answered in a unipolar format (true—not true). However, all other respondents were able to recognize the bipolar scale and reproduce the introduction correctly. For example, a respondent answered, “[...] the first characteristic is up to 5 and then continues to 10 [...]. So you start with ‘stubborn’ from a scale between 0 and 5, and the other one starts at 5.”² In summary, it can be noted that problems regarding the understanding of the instruction were rare.

Moreover, the second cognitive technique, namely general probing, did not indicate problems in handling a bipolar scale in a telephone interview in general. However, the respondents’ answers indicated another problem: the assignment of negatively connoted characteristics to high numbers and the assignment of positively connoted ones to low numbers, as well as the changing of these assignments. Eight out of 15 respondents showed difficulties in doing this. For example, the scale was often recalled by saying “10 is quiet?” “Ahh, 0 is talkative, that’s what that means,” or “0 is talkative?”³ Additionally, the open questions about problems with this scale presented difficulties. Subsequent responses clearly indicated the problem.

“It is very confusing what 0 and 10 are. You have to pay close attention! It would be different if you could read.” “[...] and if you don’t pay attention, it is possible to give a wrong answer [...] because low numbers are usually associated with something negative. You have to concentrate hard.” “What I find difficult with 10 and 0 is that sometimes 10 is the desirable one, and sometimes 0.” “You always have to recall which was 0 and which was 10 because you always have the idea in your mind that one of the two is positive and the other is negative, and the negative one is always 0 and the positive one is 10.”⁴

While the results show that using bipolar scales in telephone interviews is not problematic in general, respondents had difficulties assigning suitable values when the connoted poles of the items changed regularly. This means that the FFFK-K is

mit Zahlen von 0 bis 10 abstufen. Bei einer kleinen Zahl trifft eher die erste Eigenschaft zu, bei einer großen Zahl eher die zweite.”

2 Original: “Die erste Eigenschaft bis 5 geht und dann weiter bis 10. [...]. Also man fängt bei trotzig an bei einer Skala von 0 bis 5 und dann das andere ab 5.”

3 Original: “10 ist still?” “0 ist gesprächig. Aha so rum.” “0 ist gesprächig?”

4 Original: “Ziemlich verwirrend ist, dass das Positive oft getauscht wird, was jetzt 0 und 10 ist; man muss ziemlich aufpassen! Anders wenn man lesen könnte.”/“[...] Und wenn man dann nicht aufpasst, kann’s passieren, dass jemand eine falsche Angabe macht. [...] weil man gewohnheitsmäßig die kleinen Zahlen mit Negativem verbindet. Da muss man sich wahnsinnig konzentrieren.”/“Was ich jetzt schwierig finde bei 10 und 0, manchmal ist 10 das, was ja eigentlich wünschenswert ist und manchmal ist das, was wünschenswert ist, 0.”/“Man musste immer nachdenken, was war 0, was war 10. Weil man ja im Kopf immer hat, das eine ist immer das Positive und das andere eher negativ, und so hat das Negative immer als 0 und das Positive als 10.”

not comprehensible in its original presentation of the items because the cognitive demand is—in accordance with our results—too high for the respondents on the phone. To gain more insight into these problems, further analyses are presented. We used *live recording* of parental telephone interviews and *split-half design* in the pilot study.

3 Live Recording and Split-Half Design in the Pilot Study for Testing and Evaluating the Big Five Rating Scale in a Telephone Interview

3.1 Methodology and Data

We used a split-half design to analyze the described problems. We generated two test groups. The first test group used the original scale, and the second used an adapted version in which the poles of items A, E, G, and I were rotated (see first column in Table 2).

The pilot study was conducted with 89 parents from four federal states in Germany. Furthermore, some interviews were recorded live ($n = 20$).

3.2 Results and Consequences

The live records confirm the results from the cognitive interviews. Every second respondent from the first test group had difficulties with the polarity. This difficulty again shows that the scale is highly demanding and that respondents are assigned items incorrectly. The second test group showed no such problems.

The descriptive analysis of the interview data reveals no marginal differences between the two test groups. For an overview, see Table 2.

In a second step, we examined whether a polarity reversal shows effects on the level of the scale (see Table 3 below). The reliability of the scales for extraversion, conscientiousness, and openness/intellect increased, whereas the reliability of the neuroticism dimension decreased. The scale agreeableness, which is based on identical item polarity in both test groups, showed little change.

In summary, the data shows evidence that the second version is preferable all in all. The option from the second test group is generally more comprehensible via telephone, and false assignment of answers can thus be avoided. Consequently, higher reliabilities can generally be reached. For measuring neuroticism, the findings of the first test group showed clearly better results with respect to scale reliability even though this group contained one rotated item. As a result, the final version is slightly different than the one tested for in the second group and contains the original version for neuroticism and agreeableness, whereas the second version was already used in the pilot study for the second test group. In the end, three out of five dimensions

Table 2 Means and Standard Deviations of the Big Five Items

	Item	Test Group 1 (<i>n</i> = 46)		Test Group 2 (<i>n</i> = 43)	
		<i>M</i>	<i>SD</i> ^a	<i>M</i>	<i>SD</i> ^a
A	Talkative/quiet ^b	2.74	2.82	1.81	1.88
B	Disorderly/orderly	5.85	1.98	6.19	1.89
C	Good-Natured/touchy	3.89	2.50	4.53	2.44
D	Uninterested/interested	8.43	1.42	8.48	1.24
E	Self-Assertive/insecure ^b	3.67	2.67	2.74	1.79
F	Withdrawn/sociable	7.98	1.87	8.14	1.85
G	Focused/distractible ^b	4.28	2.53	3.79	1.83
H	Stubborn/gentle	5.12	2.16	5.44	2.20
I	Quick/slow ^b	2.26	1.91	2.40	1.56
J	Worried/calm	6.13	2.53	6.51	1.98

Note. ^a*SD* standard deviation; ^bthese items are rotated for test group 2. T-tests to compare the mean of the items by test groups showed no significant differences at 5% level. Source: Data from the pilot study, own calculations.

Table 3 Reliabilities of the Big Five Scales

Scale	Alpha Test Group 1	Alpha Test Group 2
Extraversion	.553	.818
Conscientiousness	.509	.652
Agreeableness	.597	.514
Openness/intellect	.469	.618
Neuroticism	.693	.497

Note. Source: Data from the pilot study, own calculation.

of personality traits were measured with a rotated version of the original instrument used in paper-and-pencil questionnaires.

4 Does Our Adaption of the Parent-Rating Affect the Difference Between Parent and Educator Ratings?

Previously, we learned about strengths and difficulties when parents were asked to rate their children's personality traits by employing a bipolar scale via computer-assisted telephone interviewing. Because we used different versions of the instrument in the parent interviews, we now analyze if our changes in the polarity of some items within the parent interview affected the differences of the ratings from the different evaluators. That is, these analyses can give information on the differences of rating children's personality traits in different contexts and from different evaluators in general, and they can also provide information as to whether the changes we made affect this indicator of validity. If changing the polarity of some items does not affect the similarity of the ratings, the differences between parent and educator ratings should not change substantially between the two versions of the instruments that we employed.

4.1 Data and Methods

The analytical samples contain only complete datasets within the variables of interest. That is, we considered only those children for analysis who were rated by their parents and their Kindergarten educators without missing values. The analytical sample comprises parent- and educator ratings for $n = 31$ children within the pilot study test group 1, and for $n = 1,771$ children within the main study (which is available for the scientific community).

We focus on reporting the differences in the mean values of parent and educator ratings at the scale level in order to gain information on the question of whether our adaption of the parent instrument affected the similarity of ratings based on the result of the cognitive interviews. test group 1 of the pilot study data contains the "original" version of the Big Five instrument, whereas the main study data contains the instrument that was changed according to the results of the cognitive interviews.

Because educators were asked to rate all participating children within their Kindergarten group, the data has a multilevel structure. Therefore, we report and use robust standard errors (clustered at the level of Kindergarten educators) for the mean scores of educator ratings.

4.2 Results

Figure 1 shows the mean scores of the parent and educator ratings from the pilot study using the original version within the parent interviews (test group 1). Rating the same children, parents and educators do not differ systematically in their mean values concerning neuroticism and conscientiousness, but the ratings differ concerning openness, agreeableness, and extraversion (albeit only slightly).

Figure 2 now shows the mean scores of parents and educators based on the data of the main study using the changed version within the parent interviews. Comparing the overall tendencies of parents' and educators' ratings of both figures, one can see that the patterns are similar: While neuroticism has the greatest similarity, the ratings differ the most for openness and agreeableness.

Table 4 shows an overview of the mean values and the corresponding differences for each scale. Since the scale ranges from 0 to 10, even the mean differences of about 1 point can be considered small considering that the observed behaviors can be assumed to vary systematically by context. However, according to our research question, it is more important to state that the numbers do not indicate that our changes affected the tendencies and relations of the overall differences in the ratings.

Figure 1 Comparing Parent and Educator Ratings: Results from the Pilot Study (Test Group 1)

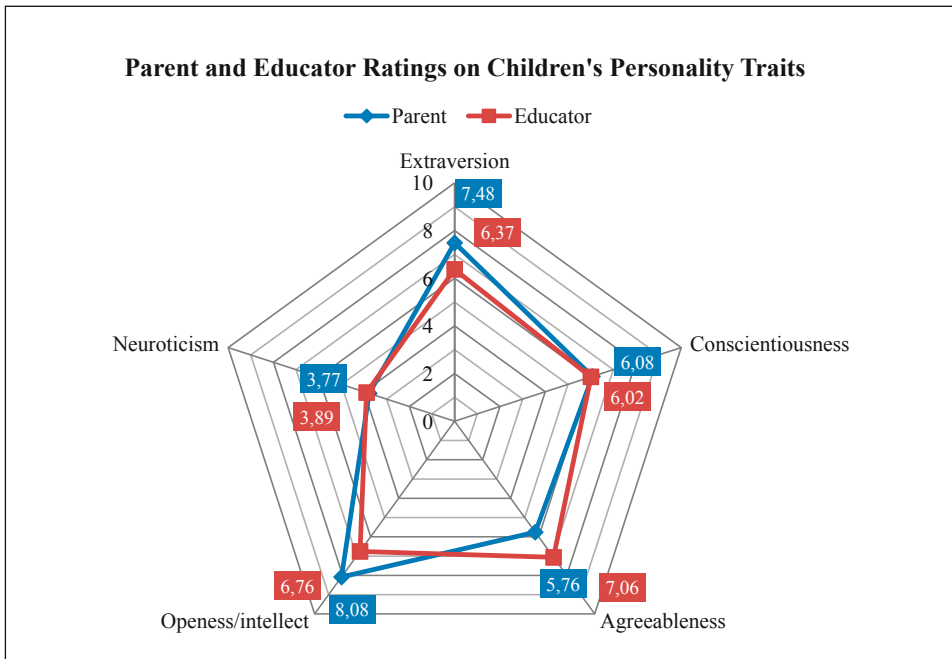
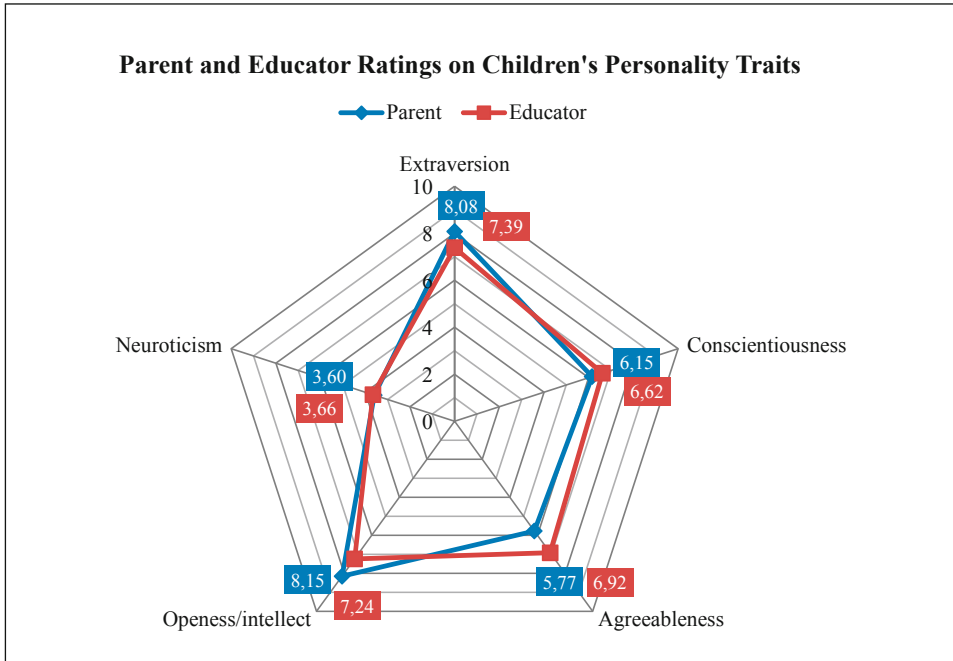


Figure 2 Comparing Parent and Educator Ratings: Results from the Main Study*



* Source: DOI: 10.5157/NEPS:SC2:1.0.0

Table 4 Mean Values and Mean Differences for the Ratings in the Pilot and Main Study

	Pilot Study ^a (Test Group 1)			Main Study ^b		
	Parent <i>M</i> (<i>SE</i>)	Educator <i>M</i> (<i>SE</i>)	Diff. Δ_{P-E}	Parent <i>M</i> (<i>SE</i>)	Educator <i>M</i> (<i>SE</i>)	Diff. Δ_{P-E}
Extraversion	7.48 (.33)	6.37 (.37)	-1.11	8.08 (.04)	7.39 (.06)	-.69
Conscientiousness	6.08 (.35)	6.02 (.42)	-.06	6.15 (.04)	6.62 (.06)	.46
Agreeableness	5.76 (.34)	7.06 (.36)	1.31	5.77 (.04)	6.92 (.07)	1.15
Openness/intellect	8.08 (.25)	6.76 (.45)	-1.32	8.15 (.03)	7.24 (.06)	-.91
Neuroticism	3.77 (.40)	3.89 (.21)	.11	3.60 (.04)	3.66 (.07)	.06

Note. ^a Standard error adjusted for 12 clusters; *n* = 31; total difference: 3.91; ^b standard error adjusted for 570 clusters; *n* = 1,771; total difference: 3.27.

In summary, the results indicate the possibility of measuring young children's personality traits indirectly and that changing the polarity of some items in the parent instrument did not affect this measure of validity substantially since the pattern of the differences remained the same over both versions of the instruments that we employed in the studies.

5 Summary

Based on the research questions formulated at the beginning of this chapter, we can now summarize the results of our analyses. The first question asked about the possibility of using bipolar rating scales in telephone interviews, and this can be affirmatively answered. Bipolarity doesn't produce unsolvable problems for interviewees. However, the cognitive interviews and the pilot study provide clear evidence that a combination of bipolar adjective scales and switches in the direction of the scale can overstrain the interviewees and increases the possibility of unintended answers. The final version of the bipolar ratings scales for measuring personality traits via telephone therefore contains both of these results.

Our second question concerned the consequences of adapting the bipolar rating scales for the telephone interview mode. Comparing the original and the adapted versions of the parent instrument with the educators' ratings for the same children, we did not find evidence that our adaption affected this measure of validity systematically. Educators and parents differ in their ratings, but these differences were not affected systematically by changing the parent instrument.

References

- Asendorpf, J. B., & van Aken, M. A. G. (2003). Validity of big five personality judgments in childhood: A 9 year longitudinal study. *European Journal of Personality, 17*(1), 1–17.
- Digman, J. M. (1990). Personality structure: Emergence of the five-factor model. *Annual Review of Psychology, 41*(1), 417–440.
- Hofstee, W. K. B. (2003). Structures of personality traits. In T. Millon, & M. J. Lerner (Eds.), *Handbook of psychology* (pp. 231–254). Hoboken, NJ: John Wiley & Sons.
- Kohnstamm, G. A., Mervielde, I., Besevegis, E., & Halverson, C. F. (1995). Tracing the big five in parents' free descriptions of their children. *European Journal of Personality, 9*(4), 283–304.
- McCrae, R. R., & Costa, P. T. (1987). Validation of the five-factor model of personality across instruments and observers. *Journal of Personality and Social Psychology, 52*(1), 81–90.
- Measelle, J. R., John, O. P., Ablow, J. C., Cowan, P. A., & Cowan, C. P. (2005). Can children provide coherent, stable, and valid self-reports on the big five dimensions? A longi-

- tudinal study from ages 5 to 7. *Journal of Personality and Social Psychology*, 89(1), 90–106.
- Mervielde, I. (2005). Persönlichkeitsbeurteilung aus entwicklungspsychologischer Perspektive. In J. B. Asendorpf (Ed.), *Enzyklopädie der Psychologie: Themenbereich C: Theorie und Forschung. Serie V: Entwicklungspsychologie* (pp. 563–616). Göttingen: Hogrefe.
- Mervielde, I., Buyst, V., & de Fruyt, F. (1995). The validity of the big-five as a model for teachers' ratings of individual differences among children aged 4–12 years. *Personality and Individual Differences*, 18(4), 525–534.
- Prüfer, P., & Rexroth, M. (2005). *Kognitive Interviews* (ZUMA How-to-Reihe No. 15). Mannheim: Zentrum für Umfragen, Methoden und Analysen.
- Schlechter, S., Blair, J., & Hey, J. V. (1996). Conducting cognitive interviews to test self-administered and telephone surveys: Which methods should we use? In American Statistical Association (Ed.), *Proceedings of the survey research methods section* (pp. 10–17). Alexandria, V. A.: American Statistical Association.
- Weinert, S., Asendorpf, J. B., Beelmann, A., Doil, H., & Frevert, S. (Eds.). (2007). *Expertise zur Erfassung von psychologischen Personmerkmalen bei Kindern im Alter von fünf Jahren im Rahmen des SOEP* (DIW Data Documentation No. 20). Berlin: Deutsches Institut für Wirtschaftsforschung.
- Willis, G. B. (2005). *Cognitive interviewing: A tool for improving questionnaire design*. Thousand Oaks: Sage Publications.

About the authors

M. Bayer
Leibniz Institute for Educational Trajectories (LifBi), Bamberg.
Lutheran University of Applied Sciences, Nuremberg.

T. Linberg
Leibniz Institute for Educational Trajectories (LifBi), Bamberg.

D. Müller
Leibniz Institute for Educational Trajectories (LifBi), Bamberg.
e-mail: doreen.mueller@lifbi.de

T. Schneider
Institute of Sociology, University of Leipzig, Leipzig.

F. Wohlkinger
Department of General Pedagogy, Education and Socialization Research,
University of Munich, Munich.

Measuring Self-Concept in the NEPS

Florian Wohlking, Michael Bayer and Hartmut Ditton

Abstract

In educational science, the idea of self-concept is well-known to be substantially correlated with learning behavior, decision making, and academic performance (cf. Shavelson and Bolus 1982; Helmke and van Aken 1995; Bong and Clark 1999; Kaufmann 2008). Therefore, it is a crucial concept in educational research, with importance for different purposes. In the National Educational Panel Study (NEPS), the measurement of self-concept needs to meet the requirements of several stages over the life course: academic self-concept during elementary school and high school, as well as a more general dimension of self-concept after leaving the highly structured context of educational institutions and entering the labor market. This task can be performed due to the hierarchical structuring of self-concept (cf., e.g., Shavelson et al. 1976; Marsh and Shavelson 1985; Marsh 1987; Lichtlein 2000). By distinguishing between two major levels, general self-concept on the one hand and domain-specific self-concept on the other, it is possible to monitor the individual's perception of him- or herself across the complete life course. This article outlines the insertion of self-concept measures used in the NEPS. Information on the theoretical concepts is given, and the chosen measures of investigation are introduced. Subsequently, selected results of students in the 5th and 9th Grade are presented.

1 Introduction

Self-related perceptions play an important role in educational research as well as in research on personality and social psychology (cf. Gecas 1982). The way people view themselves affects their behavior and thus substantially influences their lives (cf. Epstein 1973). As a result of this far-reaching impact, self-perceptions have be-

come an inherent part of research. Educational scientists, in particular, address a great deal of interest in self-related beliefs, such as self-efficacy and self-concept. For educational research, self-concept is especially interesting in its hierarchical structure (Marsh and Shavelson 1985; Marsh 1987) and its implications regarding development issues. There is a consensual understanding that the self-concept of a person should be described on different levels. On a more abstract level, constructs like general self-esteem or general self-efficacy can be found, while aspects like “academic self-concept” or “school-related self-concept” are seen as being more context-specific. Academic self-concept is well-known to correlate with academic achievement (cf., e.g., Eckert et al. 2006; Köller et al. 2006) even though the nature of this correlation is discussed controversially (cf. Kammermeyer and Martschinke 2006). Beyond this, questions about causality are even harder to answer (Helmke and van Aken 1995). It is not easy to give a precise definition of the term self-concept, especially because of the widespread usage beyond disciplinary boundaries. Rosenberg’s definition of self-concept as “the totality of the individual’s thoughts and feelings having reference to himself as an object” (Rosenberg 1979, p. 7) is well-known but also very broad.

2 Theoretical and methodological background

In educational research, self-concept is often defined as a person’s perception of him- or herself and his or her abilities (cf. Shavelson et al. 1976; Marsh and Shavelson 1985; Watermann et al. 2010). Its main characteristics are multidimensionality on the one hand and a hierarchical structure on the other (cf. Shavelson et al. 1976). At the top level of the hierarchy, there is a general dimension of self-concept, which then unfolds into several distinctive subdimensions, such as social self-concept, emotional self-concept, physical self-concept, and academic self-concept (cf. Shavelson et al. 1976; Shavelson and Bolus 1982). Each of these subdimensions can be further disaggregated into more specific subareas. For example, the academic self-concept can be disassembled into subject-specific components.

The different aspects of self-related perceptions can be used to address a great variety of questions. In his classical approach, Rosenberg (1979) used the general aspect of self-esteem to analyze differences between blacks and whites in the U.S. In addition, Kohn (1981) focused on the connections between more general dimensions of the self-concept and vocational and occupational developments. In educational research, the development of academic self-concept (or subject-specific subdimensions) is typically monitored together with the development of academic performance. Though there is substantial proof for the positive correlation between these two factors, the concrete (causal) mechanism underlying this interdependency is still unclear (cf. Dickhäuser 2006). The causal relation can be formulated in two oppositional approaches. Skill development theorists argue that social and dimensional comparisons of achievement lead to a person’s perception of his/her ability, while

self-enhancement theorists consider self-concept to be a cause of performance (cf. e.g., Calsyn and Kenny 1977; Marsh 1990a; van Aken et al. 1997; Dickhäuser 2006; Kammermeyer and Martschinke 2006). Both traditions find support in empirical analyses, and neither appears to be superior.

In addition to the ambiguousness of findings on achievement and academic self-concept, the nature of the mechanism is strongly shaped by the characteristics of the investigated school system. As Watermann et al. (2010) pointed out, the findings of American research cannot be transferred to the German situation without restrictions. Kammermeyer and Martschinke (2006) found a shift from skill-development to self-enhancement after the first Grade for the German school system.

Research on the transition to different school types after elementary school often focusses on the transition's impact on academic self-concept (cf. Köller and Baumert 2001). The changing frame of reference (the composition of students changes from heterogeneous achievement groups in elementary school to homogeneous groups after school-type selection) leads to a reevaluation of self-concept. Low-performing students' self-concepts benefit from the new reference group in which their own achievement lies closer to or even above the class average, while students demonstrating high performance find themselves in a composition in which their own achievement might not be as outstanding as it was before and they therefore have to deal with losses in self-concept. The described phenomena of the reference group is known as Big-Fish-Little-Pond effect (cp., e.g., Marsh 1990b; Marsh and Hau 2003; Marsh 2005) and can also be found in the German school system (cp. Köller 2004; Köller et al. 2006).

3 Self-Concept Measures in the German National Educational Panel Study

The National Educational Panel Study (NEPS; cf. Blossfeld et al. 2011) provides a great framework for answering questions like the ones outlined above. Its longitudinal design from early childhood to late adulthood provides a unique chance to monitor the development of constructs, such as the self-concept of abilities across a long time period containing important educational transitions, and to embed it in the context of the whole life course.

The above-mentioned hierarchy of self-concept offers the possibility to link academic research with questions on general educational processes. The distinction of a general dimension of self-concept and domain-specific subdimensions can be used to form a coherent measurement fulfilling all the needs of different life stages (cf. Wohlkinger et al. 2011).

3.1 General Self-Concept

General self-concept represents the top level of the self-concept hierarchy. Conceptually, it is not linked to any domain such as school, university, work, or family. Therefore, this measure can be used in an identical manner across all age cohorts. This allows for age-group comparisons and for testing measure stability assumptions across the whole life span.

Among potential instruments appropriate for this purpose, the Rosenberg self-esteem scale (Rosenberg 1965) was selected since self-esteem is assumed to be the base of domain-specific and situational self-evaluations and thus generally forms the key element of self-concept (cf. Ferring and Filipp 1996). Self-esteem has a strong theoretical grounding in social psychology and contains the two dimensions of “self-worth” and a kind of “competence.” Self-esteem can be seen as “outcome, motive, and buffer” and is, in this sense, an important aspect for developing processes over the whole life course (Cast and Burke 2002). Robins and Trzesniewski (2005) showed that there is a kind of normal trajectory of self-esteem across the life-span and that the existing discontinuities are connected with important life experiences at different ages. Von Collani and Herzberg (2003a; 2003b) presented a short 10-item German version of Rosenberg’s self-esteem scale that combines good psychometric characteristic (reliability, validity) and includes positive as well as negative item wording. The instrument is used with students starting from Grade 5 up to the adult stage (Roth et al. 2008).

3.2 Domain-Specific Self-Concept

Following the hierarchy of the self-concept, domain-specific measures are necessary to obtain a better-defined look on the different aspects of person’s view of him- or herself. The stage structure of the NEPS provides a quite convenient way to implement domain-specific instruments. At the school and higher-education stages, there is a focus on academic self-concept, whereas the adult stage concentrates on the spheres of work life and family.

To contribute to the needs of the academic self-concept research tradition, the domain-specific self-concept at the school stages is further disaggregated. A general dimension of academic self-concept was implemented to provide a measure for overall self-rating of school performance. Additionally, along with the NEPS emphasis on the subjects of German and mathematics, both these subjects were incorporated separately. In PISA 2000, a similar conception led to the development of a very economical instrument consisting of three short scales on verbal, mathematical, and overall academic self-concept (Kunter et al. 2002). These scales were applied for students of Grade 5 and Grade 9, enabling comparisons with the cross-sectional data acquisition of PISA within the framework of a longitudinal study.

In addition to the positive facet of self-rating, we measure learned helplessness. The conception of learned helplessness was introduced by Abramson and colleagues (Abramson et al. 1978) and is understood as counterpart of the positive self-concept. The instrument used in the NEPS was originally utilized in KOALA-S (cf. Ditton, 2007), a longitudinal study in elementary schools that thereby also provided experiences with young students. For the NEPS, we adjusted this instrument to the domain-specific level and now use it to gather helplessness in the subjects of German and mathematics separately.

Altogether, five different self-concept measures are being used at the school stages that cover different levels of the self-concept hierarchy and ensure that a great variety of questions are answered with the NEPS data.

At the stage of higher education, the school-related dimensions of German and mathematics don't play a major role for students of most subjects. For this reason, the distinction of these domains within the school context is not very applicable for other domains and was thus removed for non-school stages. Still, we differentiate between positive and negative aspects. The positive facet is covered by taking the absolute academic self-concept from Dickhäuser et al. (2002), while the student helplessness instrument is based on Jerusalem and Schwarzer (2006).

At the adult stage, not only is the differentiation between the school-typical dimensions of German and mathematics no longer appropriate, but the higher level dimension of academic self-concept also doesn't apply to the respondent's reality anymore. Therefore, only the universal dimension of self-concept, namely the Rosenberg self-esteem scale, is surveyed.

Altogether, the self-concept framework provided in the NEPS takes advantage of the structural characteristics of self-concept. The hierarchical formation, in particular, as well as the separation into positive and negative facets, is used to fulfill the peculiar needs of each life stage. With this framework, different disciplines are able to address a great variety of questions connected with the self-concept to the NEPS data.

4 First Results

To get an impression of the self-concept measures used in the NEPS, we hereby present an overview of the positive domain-specific self-concept measures and their correlation with grades for students in both Grade 5 and Grade 9.¹

1 This paper uses data from the National Educational Panel Study (NEPS): Starting Cohort Grade 5, doi: 10.5157/NEPS:SC3:1.0.0 and Starting Cohort Grade 9, doi:10.5157/NEPS:SC4:1.0.0. From 2008 to 2013, NEPS data were collected as part of the Framework Programme for the Promotion of Empirical Educational Research funded by the German Federal Ministry of Education and Research (BMBF). As of 2014, the NEPS survey is carried out by the Leibniz Institute for Educational Trajectories (LifBi) at the University of Bamberg in cooperation with a nationwide network.

Altogether, the Grade 5 and Grade 9 sample consisted of 6,085 and 16,425 cases, respectively. Among the self-concept scales presented here, complete information is available for more than 80 % of the cases.

Since a major strength of the NEPS is its large sample size, we distinguish different school types: Hauptschule (*HS*; school for basic secondary education), Realschule (*RS*; intermediate secondary school), Gymnasium (*GY*; type of school leading to upper secondary education and Abitur), and Förderschule (*FöS*; school establishment for students whose development cannot be adequately assisted in mainstream schools on account of disability). For readability purposes and to reduce complexity for the following demonstration of analysis potential, other types of schools, such as schools with mixed student populations, were excluded.

Intercorrelations of self-concept measures

Theoretically, according to the hierarchical structure of self-concept, both dimensions of subject-specific self-concept are considered to be partially included in the general dimension of academic self-concept. This assumption turns out to be correct for both age cohorts, as Figure 1 shows. In Grade 5, the correlation between the general *academic self-concept* (*ASC*) and the *verbal self-concept* (*VSC*) is $r = .513$, while the correlation with the *mathematical self-concept* (*MSC*) is $r = .384$. For Grade 9, the pattern is very similar, even though the coefficients show slightly lower values.

Moreover, for both cohorts, we find a correlation close to zero between the two lower-level self-concept measures VSC and MSC. This indicates that the instruments are able to clearly distinguish between the two domains of verbal and mathematical skills.

Almost the same relations found independently of school type appear when distinguishing the results. Table 1 outlines the intercorrelations of the self-concept measures for each school type separately. For Grade 5, there are small differences between the school types. While neither HS nor RS nor GY shows a correlation between verbal and mathematical dimensions of self-concept, there is a correlation of $r = .247$ for FöS

Figure 1 General intercorrelations (Pearson) of self-concept measures in Grade 5 and Grade 9

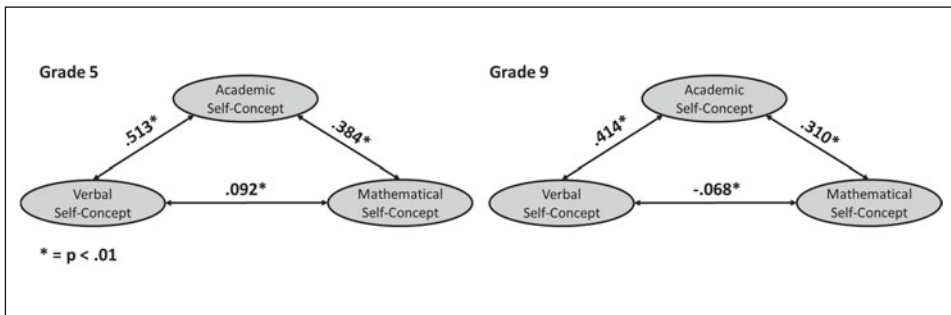


Table 1 Intercorrelations of Self-Concept Measures by School Type

School Type	Grade 5				Grade 9			
	FöS	HS	RS	GY	FöS	HS	RS	GY
VSC * MSC	.247*	n.s.	n.s.	n.s.	.112*	-.110*	-.114*	-.057*
VSC * ASC	.494*	.467*	.477*	.514*	.447*	.405*	.358*	.417*
MSC * ASC	.499*	.429*	.278*	.360*	.418*	.238*	.283*	.368*

Note. FöS = Förderschule (school establishment for students whose development cannot be adequately assisted at mainstream schools on account of disability); HS = Hauptschule (school for basic secondary education); RS = Realschule (intermediate secondary school); GY = Gymnasium (type of school leading to upper secondary education and Abitur); ASC = academic self-concept; VSC = verbal self-concept; MSC = mathematical self-concept; * = $p < .01$.

students. The relationship between the two lower-level self-concepts and the general academic dimension shows little variety across school types. The most outstanding value is the connectivity between MSC and ASC for students from RS, which is somewhat lower than for students from other school types.

The measures show slightly more variation for Grade 9. Although still close to zero, a remarkable difference between the school types can be found in the correlation between VSC and MSC: For FöS students, there is a positive correlation, whereas the other school types show a negative correlation.

When comparing Grade 5 with Grade 9, it appears that almost all correlations show lower values in the older age group. This finding will be even more interesting when the younger cohort reaches Grade 9 in a few years and longitudinal comparisons become feasible.

Mean comparison across school types

After the first impression of the intercorrelations of the self-concept measures, a look at the means seems appropriate. Table 2 displays the means for each instrument, differentiated by school type.

Within Grade 5, there is basically a slight increase of the means of all three self-concept scales across the school types, and only FöS students fall a bit outside of this pattern. In Grade 9, the picture changes: Compared with the Grade 5 means, only the VSC maintains its level. Both the ASC and (especially) the MSC are remarkably lower across all school types. Additionally, the means show less variation across the school types and now lie closer to each other. The differences presented in Table 2 were further examined with T-Tests. With few exceptions, almost all differences between the means of Grade 5 students are significant. In Grade 9, some significant coefficients can still be found, but in general, the differences are lower than the Grade 5 mean differences. Concretely, the differences between GY and the other school types remain significant, while the distance between HS and RS and the distance between

Table 2 Means of Self-Concept Measures for each School Type

School Type	Grade 5				Grade 9			
	Fös	HS	RS	GY	Fös	HS	RS	GY
Verbal Self-Concept (VSC)	3.00	2.81	2.93	3.12	2.94	2.88	2.88	3.01
Mathematical Self-Concept (MSC)	3.04	2.75	2.89	3.04	2.58	2.52	2.49	2.55
Academic Self-Concept (ASC)	3.11	3.03	3.10	3.26	2.84	2.87	2.85	2.92
N	437+	569+	993+	2150+	966+	3446+	2997+	4970+

Note. Fös = Förderschule (school establishment for students whose development cannot be adequately assisted at mainstream schools on account of disability); HS = Hauptschule (school for basic secondary education); RS = Realschule (intermediate secondary school); GY = Gymnasium (type of school leading to upper secondary education and Abitur); the "+" after each number in column N indicates that this is the **minimum** number of cases available for each scale.

HS and Fös decrease. This finding is consistent with the Big-Fish-Little-Pond effect: Until Grade 4, all students also compare themselves to students who are later separated to different school types. From Grade 5 on, their frame of reference changes, which leads to an adaptation of the self-rating after being separated into homogeneous achievement groups.

Correlations of self-concept measures with grades

The examination of the means begs the question of whether these patterns can also be detected when including grades. Table 3 shows the correlations between the three self-concept scales and academic achievement. To reflect the dimensionality of the scales, grades for the school subjects of German and mathematics were included separately and additionally averaged to take account of the hierarchy level.

All correlations are negative since lower grades indicate better achievement in the German school system. As expected, the correlations between the self-concepts and their corresponding grades show the highest connection, while the oppositional correlations between VSC and grades in mathematics and between MSC and grades in German in general is low or zero. Furthermore, all correlations between ASC and grades are lower than the correlations of subject-specific self-concepts and the grades of the corresponding subjects. Both findings can be regarded as indicators for the good separation between the different self-concept constructs. The ASC can be used when examining academic performance independent of concrete subjects, while VSC and MSC can be used for subject-specific questions.

For Grade 5, there is an erratic correlation pattern across the different school types. Students with special educational needs (Fös) mostly show the lowest correlations between self-ratings and achievements. Generally, the correlations are at a

Table 3 Correlations (Pearson) of Self-Concept Measures with Grades for each School Type

School Type	Grade 5				Grade 9			
	Fös	HS	RS	GY	Fös	HS	RS	GY
VSC * grade_G	-.274*	-.416*	-.447*	-.386*	-.470*	-.498*	-.534*	-.608*
VSC * grade_M	n. s.	n. s.	n. s.	-.098*	-.175*	n. s.	n. s.	-.100*
MSC * grade_M	-.337*	-.539*	-.534*	-.445*	-.509*	-.612*	-.638*	-.703*
MSC * grade_G	n. s.	n. s.	n. s.	n. s.	n. s.	n. s.	n. s.	-.109*
ASC * grade_G	-.163*	-.233*	-.278*	-.257*	-.262*	-.328*	-.348*	-.485*
ASC * grade_M	-.154*	-.167*	-.211*	-.200*	-.213*	-.247*	-.329*	-.455*
ASC * grade_GM	-.174*	-.248*	-.296*	-.270*	-.281*	-.342*	-.413*	-.560*

Note. Fös = Förderschule (school establishment for students whose development cannot be adequately assisted at mainstream schools on account of disability); HS = Hauptschule (school for basic secondary education); RS = Realschule (intermediate secondary school); GY = Gymnasium (type of school leading to upper secondary education and Abitur); ASC = academic self-concept; VSC = verbal self-concept; MSC = mathematical self-concept; grade_G = grade in German; grade_M = grade in mathematics; grade_GM = average grade German and mathematics; * = $p < .01$.

moderate level, and the highest correlation can be found between the mathematical self-concept and grades in math. The results for GY lie a bit underneath those of RS, partially even under the level of HS.

In Grade 9, the correlation between academic performance and self-concept is generally much stronger. Again, Fös students show lower correlations than the other school types. As before, the correlation between the MSC and grades in math is the highest. Contrary to the situation in Grade 5, the GY correlations here obtain the highest results. This finding can again be connected to the Big-Fish-Little-Pond effect: After being separated into the different school types, students with lower achievements, in particular, benefit from the new reference group, while students with higher achievements have to deal with higher competition in their new environment. After having spent four years in their new reference group, the relationship between self-concept and grades is realigned.

5 Conclusion

The self-concept measures provided by NEPS contain a great potential for many questions that have not yet been able to be answered by other datasets. As the results presented above show, the NEPS design, with its large-scale sample, can be used to distinguish different school types and still remain large enough for complex analyses. This characteristic particularly helps in deepening research on school-type-related

subgroup analyses, for example, by examining well-known phenomena such as the Big-Fish-Little-Pond effect.

On the one hand, both the instruments measuring academic self-concept as well as the timing of their usage allow for comparisons with other studies such as PISA, and on the other hand, they also allow for longitudinal comparisons that monitor self-concept development processes. The distinction between different hierarchical levels enables research located in a more general area as well as examinations of concrete subject-specific questions.

The results presented here only focus on the dimension of positive academic self-concept; however, there is greater potential within the negative dimension of self-perception and the non-academic measures. The unique structure of the NEPS, with its focus on the complete life course, enables questions focusing on the whole life-course, especially when addressing questions on educational mechanisms after leaving the homogeneous context of school.

Furthermore, the offering of other self-related concepts, such as motivation, goal attainment, and personality measured in a similar hierarchical structuring (cf. Wohlkinger et al. 2011), will also contribute to obtaining a better understanding of the interdependency of education, competence development, and self-perceptions.

The results indicate that there has to be some further analyses regarding students with special educational needs and their negative relationship between the two subject-specific self-concepts. Furthermore, there are indications that gender makes some difference, as Schilling et al. (2006) have examined. These and other topics need to be explored in further analyses.

References

- Abramson, L. Y., Seligman, M. E., & Teasdale, J. D. (1978). Learned helplessness in humans: Critique and reformulation. *Journal of Abnormal Psychology, 87*(1), 49–74.
- van Aken, M. A. G., Helmke, A., & Schneider, W. (1997). Selbstkonzept und Leistung—Dynamik ihres Zusammenspiels: Ergebnisse aus dem SCHOLASTIK-Projekt. In F. E. Weinert, & A. Helmke (Eds.), *Entwicklung im Grundschulalter* (pp. 341–350). Weinheim: Psychologische Verlags Union.
- Blossfeld, H.-P., Roßbach, H.-G., & von Maurice, J. (Eds.). (2011). Education as a Life-long Process—The German National Educational Panel Study (NEPS). [Special issue] *Zeitschrift für Erziehungswissenschaft, 14*. Wiesbaden: VS Verlag für Sozialwissenschaften.
- Bong, M., & Clark, R. E. (1999). Comparison between self-concept and self-efficacy in academic motivation research. *Educational Psychologist, 34*(3), 139.
- Calsyn, R. J., & Kenny, D. A. (1977). Self-concept of ability and perceived evaluation of others: Cause or effect of academic achievement? *Journal of Educational Psychology, 69*(2), 136–145.

- Cast, A. D., & Burke, P. J. (2002). A theory of self-esteem. *Social Forces*, 80(3), 1041–1068.
- von Collani, G., & Herzberg, P. Y. (2003a). Eine revidierte Fassung der deutschsprachigen Skala zum Selbstwertgefühl von Rosenberg. *Zeitschrift für Differentielle und Diagnostische Psychologie*, 24(1), 3–7.
- von Collani, G., & Herzberg, P. Y. (2003b). Zur internen Struktur des globalen Selbstwertgefühls nach Rosenberg. *Zeitschrift für Differentielle und Diagnostische Psychologie*, 24(1), 9–22.
- Dickhäuser, O. (2006). Fähigkeitsselfkonzepte—Entstehung, Auswirkung, Förderung. *Zeitschrift für Pädagogische Psychologie*, 20(1/2), 5–8.
- Dickhäuser, O., Schöne, C., Spinath, B., & Stiensmeier-Pelster, J. (2002). Die Skalen zum akademischen Selbstkonzept: Konstruktion und Überprüfung eines neuen Instrumentes. *Zeitschrift für Differentielle und Diagnostische Psychologie*, 23(4), 393–405.
- Ditton, H. (Ed.). (2007). *Kompetenzaufbau und Laufbahnen im Schulsystem: Ergebnisse einer Längsschnittuntersuchung an Grundschulen*. Münster: Waxmann.
- Eckert, C., Schilling, D., & Stiensmeier-Pelster, J. (2006). Einfluss des Fähigkeitsselfkonzepts auf die Intelligenz und Konzentrationsleistung. *Zeitschrift für Pädagogische Psychologie*, 20(1/2), 41–48.
- Epstein, S. (1973). The self-concept revisited: Or a theory of a theory. *American Psychologist*, 28(5), 404–416.
- Ferring, D., & Filipp, S.-H. (1996). Messung des Selbstwertgefühls: Befunde zu Reliabilität, Validität und Stabilität der Rosenberg-Skala. *Diagnostica*, 42(3), 284–292.
- Gecas, V. (1982). The self-concept. *Annual Review of Sociology*, 8, 1–33.
- Helmke, A., & van Aken, M. A. G. (1995). The causal ordering of academic achievement and self-concept of ability during elementary school: A longitudinal study. *Journal of Educational Psychology*, 87(4), 624–637.
- Jerusalem, M., & Schwarzer, R. (2006). Dimensionen der Hilflosigkeit. In A. Glöckner-Rist (Ed.), *ZUMA-Informationssystem: Elektronisches Handbuch sozialwissenschaftlicher Erhebungsinstrumente. ZIS Version 10.00*. Mannheim: Zentrum für Umfragen, Methoden und Analysen.
- Kammermeyer, G., & Martschinke, S. (2006). Selbstkonzept- und Leistungsentwicklung in der Grundschule: Ergebnisse aus der KILIA-Studie. *Empirische Pädagogik*, 20(3), 245–259.
- Kaufmann, A. (2008). *Die Rolle motivationaler Schülermerkmale bei der Entstehung sozialer Disparitäten des Schulerfolgs: Eine Längsschnittuntersuchung an Grundschulen in Bayern und Sachsen*. Berlin: Mensch-und-Buch-Verl.
- Kohn, M. L. (1981). *Persönlichkeit, Beruf und soziale Schichtung*. Stuttgart: Klett-Cotta.
- Köller, O. (2004). Konsequenzen von Leistungsgruppierungen. In *Pädagogische Psychologie und Entwicklungspsychologie* (Vol. 37). Münster: Waxmann.
- Köller, O., & Baumert, J. (2001). Leistungsgruppierungen in der Sekundarstufe I: Ihre Konsequenzen für die Mathematikleistung und das mathematische Selbstkonzept der Begabung. *Zeitschrift für Pädagogische Psychologie*, 15(2), 99–110.

- Köller, O., Trautwein, U., Lüdtke, O., & Baumert, J. (2006). Zum Zusammenspiel von schulischer Leistung, Selbstkonzept und Interesse in der gymnasialen Oberstufe. *Zeitschrift für Pädagogische Psychologie*, 20(1/2), 27–39.
- Kunter, M., Schümer, G., Artelt, C., Baumert, J., Klieme, E., Neubrand, M., ... Weiß, M. (2002). *Materialien aus der Bildungsforschung: PISA 2000. Dokumentation der Erhebungsinstrumente*. Berlin: Max-Planck-Institut für Bildungsforschung.
- Lichtlein, M. (2000). *Selbstkonzeptentwicklung in der beruflichen Erstausbildung unter besonderer Berücksichtigung motivationaler Aspekte*. In *Münchener Beiträge zur Wirtschafts- und Sozialpsychologie*. München: Utz.
- Marsh, H. W. (1987). The hierarchical structure of self-concept and the application of hierarchical confirmatory factor analysis. *Journal of Educational Measurement*, 24(1), 17–39.
- Marsh, H. W. (1990a). Causal ordering of academic self-concept and academic achievement: A multiwave, longitudinal panel analysis. *Journal of Educational Psychology*, 82(4), 646–656.
- Marsh, H. W. (1990b). The structure of academic self-concept. The Marsh/Shavelson Model. *Journal of Educational Psychology*, 82(4), 623.
- Marsh, H. W. (2005). Big-fish-little-pond effect on academic self-concept. *Zeitschrift für Pädagogische Psychologie*, 19(3), 119–127.
- Marsh, H. W., & Hau, K.-T. (2003). Big-fish-little-pond effect on academic self-concept: A cross-cultural (26-country) test of the negative effects of academically selective schools. *American Psychologist*, 58(5), 364–376.
- Marsh, H. W., and Shavelson, R. (1985). Self-Concept: Its multifaceted, hierarchical structure. *Educational Psychologist*, 20(3), 107–123.
- Robins, R. W., & Trzesniewski, K. H. (2005). Self-esteem development across the lifespan. *Current Directions in Psychological Science*, 14(3), 158–162.
- Rosenberg, M. (1965). *Society and the adolescent self-image*. Princeton, NJ: Princeton University Press.
- Rosenberg, M. (1979). *Conceiving the self*. New York: Basic Books, Inc.
- Roth, M., Decker, O., Herzberg, P. Y., & Brähler, E. (2008). Dimensionality and norms of the Rosenberg Self-esteem Scale in a German general population sample. *European Journal of Psychological Assessment*, 24(3), 190–197.
- Schilling, S. R., Sparfeldt, J. R., & Rost, D. H. (2006). Facetten schulischen Selbstkonzepts—Welchen Unterschied macht das Geschlecht? *Zeitschrift für Pädagogische Psychologie*, 20(1/2), 9–18.
- Shavelson, R. J., & Bolus, R. (1982). Self-Concept: The interplay of theory and methods. *Journal of Educational Psychology*, 74(1), 3–17.
- Shavelson, R. J., Hubner, J. J., & Stanton, G. C. (1976). Self-Concept: Validation of construct interpretations. *Review of Educational Research*, 46(3), 407–441.
- Watermann, R., Klingebiel, F., & Kurtz, T. (2010). Die motivationale Bewältigung des Grundschulübergangs aus Schüler- und Elternsicht. In K. Maaz, J. Baumert, C. Gresch, & N. McElvany (Eds.), *Bildungsforschung. Der Übergang von der Grundschule in die*

weiterführende Schule (Vol. 34, pp. 355–383). Berlin: Bundesministerium für Bildung und Forschung.

Wohlkinger, F., Ditton, H., von Maurice, J., Haugwitz, M., & Blossfeld, H.-P. (2011). Motivational concepts and personality aspects across the life course. In H.-P. Blossfeld, H.-G. Roßbach, & Maurice, J. von (Eds.), *Zeitschrift für Erziehungswissenschaft, 14. Education as a lifelong process: The German Educational Panel Study (NEPS)* (pp. 155–168): VS Verlag für Sozialwissenschaften.

About the authors

M. Bayer

Leibniz Institute for Educational Trajectories (LifBi), Bamberg.
Lutheran University of Applied Sciences, Nuremberg.

H. Ditton

Chair of Pedagogy and Research on Education and Socialization,
Ludwig-Maximilians-University of Munich, Munich.

F. Wohlkinger

Chair of Pedagogy and Research on Education and Socialization,
Ludwig-Maximilians-University of Munich, Munich.
e-mail: florian.wohlkinger@edu.lmu.de

Identifying Immigrants and Their Descendants in the National Educational Panel Study

Cornelia Kristen, Melanie Olczyk and Gisela Will

Abstract

The data gathered in the German National Educational Panel Study (NEPS) can be used to describe and analyze the education of immigrants and their offspring across different stages in the school career and the life course and to uncover the origins of ethnic educational inequalities. In order to complement this task, it is necessary to adequately identify the population of interest. For this purpose, the NEPS includes a set of potential *immigrant identifiers*. This contribution provides an overview of the various measures and illustrates alternative ways of considering immigrants and their descendants. It also addresses a selection of problems that arise when applying one operationalization as opposed to another. The focus is on the country of birth, citizenship, and language use. The analyses are based on NEPS data from three starting cohorts: Kindergarten, Grade 5, and Grade 9. The findings indicate that the size of the immigrant population varies when using different operationalizations as well as across cohorts. Assignments based on citizenship and language use yield a substantially smaller migrant population than assignments via the country of birth. Drawing on the example of the two largest immigrant groups in Germany, namely children and youth of Turkish origin and from the Former Soviet Union, it turns out that the use of certain identifiers may not be equally appropriate for different groups.

1 Introduction

The description of ethnic educational inequalities in Germany has become considerably richer in recent years. However, a comprehensive picture of achievements and attainments of immigrants and their descendants is still called for, particularly when focusing on different migrant groups and their performance throughout different

stages of the educational career (Kristen et al., 2011: 123). Using the National Educational Panel Study (NEPS) data allows for closing this gap. The NEPS offers the unique opportunity to describe the education of children from immigrant families beginning with early childhood throughout entry into preschool institutions as well as the subsequent school career up to the transition to higher education and the labor market. Most importantly, the NEPS data contain information on a variety of conditions that are specific to the educational careers of the migrant population, thus offering numerous opportunities not only for thorough and detailed descriptions but, most importantly, also for uncovering the origins of ethnic inequalities across the life course.

A central precondition for describing and explaining ethnic differences in education is the identification of the immigrant population (Olczyk, Will, & Kristen, 2014: 3). Since the adequacy of a certain measure largely depends on the specific research question, the relevant characteristics cannot be fixed a priori. Therefore, the NEPS includes a set of potential *immigrant identifiers* that may be used depending on the research interest. This distinguishes the NEPS data from most other data sources in which the information available is usually limited to a few characteristics.

In the following sections, we introduce different measures that are available in the NEPS and illustrate alternative ways of identifying immigrants and their descendants. We also discuss the advantages and limitations associated with different operationalizations. In a next step, we show how the groups identified as immigrants vary when using one definition as opposed to another (see also Gresch & Kristen, 2011). Drawing on the example of the two largest migrant groups in Germany, namely the population of Turkish origin and from the Former Soviet Union (FSU), we eventually demonstrate that the use of specific identifiers may not be equally appropriate for certain groups. Our analyses are based on NEPS data from three starting cohorts: Kindergarten, Grade 5, and Grade 9.

2 Immigrant Identifiers in the NEPS

Important measures to identify immigrants and their children in the NEPS include the country of birth, the current nationality, and, if applicable, the former nationality as well as naturalization and the immigration and residence status. Moreover, NEPS users may consider language usage in different contexts, including the language of the country of origin and destination. These instruments provide ample opportunities to study immigrants and their offspring. Most importantly, this comprehensive information is consistently collected for the six NEPS starting cohorts. In the following section, we consider each of these measures and describe typical ways of using them.

2.1 Country of Birth

The NEPS provides information on the country of birth of the target person and of his or her parents and grandparents. In most cohorts, the target person specifies the country of origin. In the two school cohorts (Grade 5 and Grade 9), parents provide this information in addition to the student. In the two youngest cohorts (Early Childhood and Kindergarten), only the parents specify the country of birth.

In general, using the country of origin allows for distinguishing between different immigrant groups and generations. Given that the NEPS collects information on the grandparents' country of birth on a large-scale basis for the first time in Germany, it is possible to identify the third generation, as well. This additional piece of information allows for fine-grained distinctions, which, depending on the research problem, may be of great relevance.

For example, using information on the country of birth of the grandparents also enables for identifying individuals as immigrant offspring in cases in which the parents belong to the second or 2.5th generation. Without the grandparent measure, the target person would be assigned to the majority. However, when considering the grandparents' country of origin, this offspring is part of the third generation. Another advantage of using information on the grandparents is that it allows for studying the offspring of interethnic couples. Usually, when only including the country of birth of the child and the parents, all cases in which one parent is born abroad and one is born in Germany would be considered interethnic. However, taking into account the grandparents' country of birth means that second or 2.5th generation parents are no longer assigned to the majority, and a relationship between this parent and a first-generation immigrant who comes from the same country of origin as the grandparents would in this case not be labeled *interethnic*.

Additional measures, such as the duration of stay, can be used to further describe the first generation. For example, it is possible to distinguish between individuals who have spent their entire school career in the country of destination and those who migrated later in life and therefore attended school in a different system. The former is often called the 1.5th generation, whereas the latter is usually assigned to the first generation.

Apart from this fine-grained view of the generation status, researchers are often interested in particular immigrant groups. The country of birth may also be used to assign individuals to certain groups. At the same time, the place of birth may not always be a sufficient piece of information, for example, when different ethnic groups have been born in the same region or country. The most prominent example, surely, is the distinction between ethnic Germans from the Former Soviet Union (the so-called *Spätaussiedler*) and other immigrants from the Former Soviet Union who do not have German ancestors. Identifying these groups requires using further measures like current nationality, which, in the case of *Spätaussiedler*, would usually be German upon arrival. It would also be possible to take the immigration and residence status into account.

2.2 Citizenship, Naturalization, and Immigration and Residence Status

From Grade 6 onwards, the target person indicates his or her citizenship/s. In the younger cohorts, this information is gathered in the parent interview. As citizenship is still the central measure mostly used in official statistics, this information ensures comparability with other data sources that are based on citizenship.

At the same time, the NEPS data comprise additional information on this topic. After the initial measurement in the first wave, respondents also indicate their nationality in subsequent waves, whereby changes in citizenship status can be considered. Moreover, adult target persons with a German passport are asked whether they have possessed this nationality since birth. If applicable, the date (i. e., the year and the month) of naturalization is documented. Thus, NEPS users can identify naturalized persons as well as the date of naturalization.

Combining information on the year of naturalization with the year of immigration provides one opportunity to address Spätaussiedler. These are first-generation migrants who possess a German nationality upon arrival or receive it shortly thereafter (Bundesministerium des Innern, 2013). This distinguishes them from other new migrants with a foreign passport who are usually not entitled to German citizenship upon arrival and have to fulfill certain conditions, such as having legitimately stayed in Germany for eight years. Thus, a short period between the date of arrival and naturalization is specific to the Spätaussiedler status.

In the Kindergarten- and Grade 5 cohort, parents are additionally asked whether their child possesses a second citizenship. This is particularly relevant in the German case considering the amended law on nationality, which became valid in 2000. Children born to foreign nationals in Germany are usually eligible for German citizenship. They obtain the German nationality in addition to that of their parents by birth. As a consequence, a large share of children born to foreign parents now have both the German and a foreign nationality.

With the NEPS data, it is also possible to identify different types of immigrants, such as refugees or migrant workers. In this case, the immigration status is of relevance. All adult target persons born abroad are asked about their immigration status. This allows for distinguishing between Spätaussiedler, asylum seekers or refugees, foreign students, migrant workers, and immigrants who come for the purpose of reuniting their family. Parents' immigration status is documented in the preschool- and school cohorts.

Furthermore, the NEPS also covers the residence status for all adult target persons and in the starting cohorts Early Childhood, Kindergarten, Grade 5, and Grade 9 for the parents of the target person. This allows for distinguishing between immigrants with a permanent residence permit and those with a temporary residence permit.

2.3 First Language and Language Use

Other characteristics sometimes used to identify immigrants and their children are the first language and language use (e. g., Bellin, Dunge, & Gunzenhauser, 2010; Kristen, 2008; Mudiappa & Kluczniok, 2015; Stanat, 2006; Van der Slik, Driessen, & De Bot, 2006; Wagner, Helmke, & Schrader, 2009). Within the NEPS, all target persons are asked which language they learned during early childhood as well as which language they usually speak at home and in other contexts. In the youngest starting cohorts (i. e., Early Childhood and Kindergarten), this information is collected via the parents. Information on language use may also help to identify different ethnic groups from the same country of origin, for example, Kurds who were born in Turkey and speak Kurdish at home.

2.4 Limitations

Despite the rich pool of information the NEPS data offers, not all groups can be identified unambiguously. For example, it is not possible to distinguish between Turks and Kurds from Turkey as both groups were born in Turkey and possess Turkish citizenship. While many of the Kurds come to Germany as political refugees and often speak Kurdish instead of Turkish, immigration status and language use may not be sufficient to capture all Kurds.

Moreover, in starting cohorts in which the samples are based on registry data, illegal immigrants are excluded by definition. However, they may be part of starting cohorts in which the samples are drawn from schools. The problem of identification nevertheless remains in these instances since the question of residence status addresses its nature instead of inquiring whether the status is legal or not.

Due to these constraints, it may not be possible to study specific groups, such as the above-mentioned Kurds or illegal migrants, as well as other ethnic minorities, for example, the Sinti and Roma. Even if they could be detected, the sizes of these groups would probably be too small to allow for meaningful analyses in most cases.

3 Data and Operationalization

In the following section, we take a closer look at three important immigrant identifiers: the country of birth, citizenship, and language use. Our analyses are based on data from the first wave of three NEPS starting cohorts: Kindergarten (NEPS Starting Cohort 2, version 2.0.0),¹ Grade 5 (NEPS Starting Cohort 3, version 2.0.0),² and

1 Doi:10.5157/NEPS:SC2:2.0.0.

2 Doi:10.5157/NEPS:SC3:2.0.0.

Grade 9 (NEPS Starting Cohort 4, version 4.0.0).³ For reasons of comparability, we use information from the parent interviews. The only exceptions pertain to citizenship and language use in Grade 9, where we need to take into account student measures. Hence, we only consider cases with both a student- and parent interview in this cohort.

First, by using the country of birth of the target person, his or her parents, and grandparents, we consider whether the individual in question stems from a migrant family. More specifically, we assign individuals to the immigrant population if the target person, at least one parent, or at least two grandparents were born abroad (for details, see Olczyk, Will, & Kristen, 2014). Moreover, we consider the generation status and differentiate between first, second, 2.5th, and third-generation individuals. The first generation is composed of target persons who were born abroad. The second generation was born in Germany to parents who were both born abroad, while in the 2.5th generation, only one parent was born in another country. Finally, in the third generation, at least two of the four grandparents were born abroad, while the target person and his or her parents were born in Germany. Obviously, further differentiations are possible (see *ibid.*). We also use the country-of-birth information to assign individuals to different immigrant groups. Later on, we combine these measures and address different generations of individuals from the Former Soviet Union and Turkey.

The second identifier is citizenship. We distinguish between target persons who possess only the German nationality, those who possess only a foreign nationality, and individuals who have both German and a foreign citizenship.

The third measure refers to language use between the target person and the parents as well as among the siblings. For each relation, it is possible to distinguish between individuals who only or mainly use German versus those who predominantly use another language. If the respondent states that he or she only or mainly speaks another language with at least one family member, this information is used—even if the person speaks predominantly German with other family members.

We drop cases with missing values. The sample size is therefore reduced by seven cases in the Kindergarten cohort, by six cases in Grade 5, and by 517 cases in Grade 9. The relatively large number of missings in Grade 9 is due item nonresponse in the student questionnaire on citizenship and language use. In total, the analyses are based on 2,333 cases in the Kindergarten cohort, 4,146 cases in Grade 5, and 8,269 cases in Grade 9.

3 Doi:10.5157/NEPS:SC4:4.0.0.

4 Results

In this section, we first focus on the overall size of the immigrant population (4.1) and then illustrate how the different operationalizations relate to each other (4.2). Thereafter, we take a closer look at the distributions for the two largest immigrant groups in Germany, that is, children and youth of Turkish origin and from families who stem from the Former Soviet Union (4.3).

4.1 The Size of the Immigrant Population in NEPS Kindergartens and Schools

Table 1 illustrates the sizes of the immigrant population according to the different operationalizations.

When considering the country of birth, the share of target persons of immigrant origin is the largest in the Kindergarten cohort at 30.4 %, followed by a share of 21.8 % in the Grade 5 cohort and 19.1 % in the Grade 9 cohort. We observe a similar pattern when assigning the third generation to the majority (not shown here), although, obviously, the percentages decrease somewhat (i. e., to 26.3 % in Kindergarten, to 18.7 % in Grade 5, and to 16.1 % in Grade 9). The majority of target persons of migrant origin belong to the second or 2.5th generation. At over 80 %, this proportion is especially large in the Kindergarten cohort. The respective shares in the older cohorts are somewhat smaller at about 70 %. In Grade 9, in contrast, the first generation, at 15.3 %, is larger than in the younger cohorts, in which the first generation amounts to 12.7 % in Grade 5 and 4.7 % in Kindergartens.

When focusing on citizenship and language use, the percentages become substantially smaller. The portion of individuals who only possess a foreign nationality ranges from roughly 2 % in the preschool cohort to 4 % in both the Grade 5 cohort and the Grade 9 cohort. The share of target persons with a German and a foreign citizenship is somewhat larger, especially in the youngest cohort, in which it adds up to 6.9 %. The overall size of the immigrant population based on citizenship lies between 8 and 10 % in all cohorts. Finally, when considering language use, we obtain an immigrant share of 14.8 % among Kindergarten children, 8.7 % among fifth graders, and 8.3 % among ninth graders.

Taken together, these numbers illustrate that the size of the immigrant population varies considerably within and between cohorts depending on the measure applied.

Table 1 The immigrant population according to different immigrant identifiers

	Kindergarten		Grade 5		Grade 9	
	<i>N</i>	%	<i>N</i>	%	<i>N</i>	%
Country of birth						
Majority	1,623	69.6	3,241	78.2	6,692	80.9
Immigrant origin (up to the 3rd generation)	710	30.4	905	21.8	1,577	19.1
Total	2,333	100.0	4,146	100.0	8,269	100.0
Generation status						
1st generation	33	4.7	115	12.7	241	15.3
2nd generation	295	41.6	259	28.6	370	23.5
2.5th generation	285	40.1	400	44.2	724	45.9
3rd generation	97	13.7	131	14.5	242	15.4
Total	710	100.1	905	100.0	1,577	100.1
Citizenship						
German citizenship only	2,117	90.7	3,762	90.7	7,587	91.8
German and foreign	161	6.9	218	5.3	357	4.3
Foreign citizenship/s only	55	2.4	166	4.0	325	3.9
Total	2,333	100.0	4,146	100.0	8,269	100.0
Language use						
German language only/mainly	1,988	85.2	3,786	91.3	7,585	91.7
Other language only/mainly	345	14.8	360	8.7	684	8.3
Total	2,333	100.0	4,146	100.0	8,269	100.0

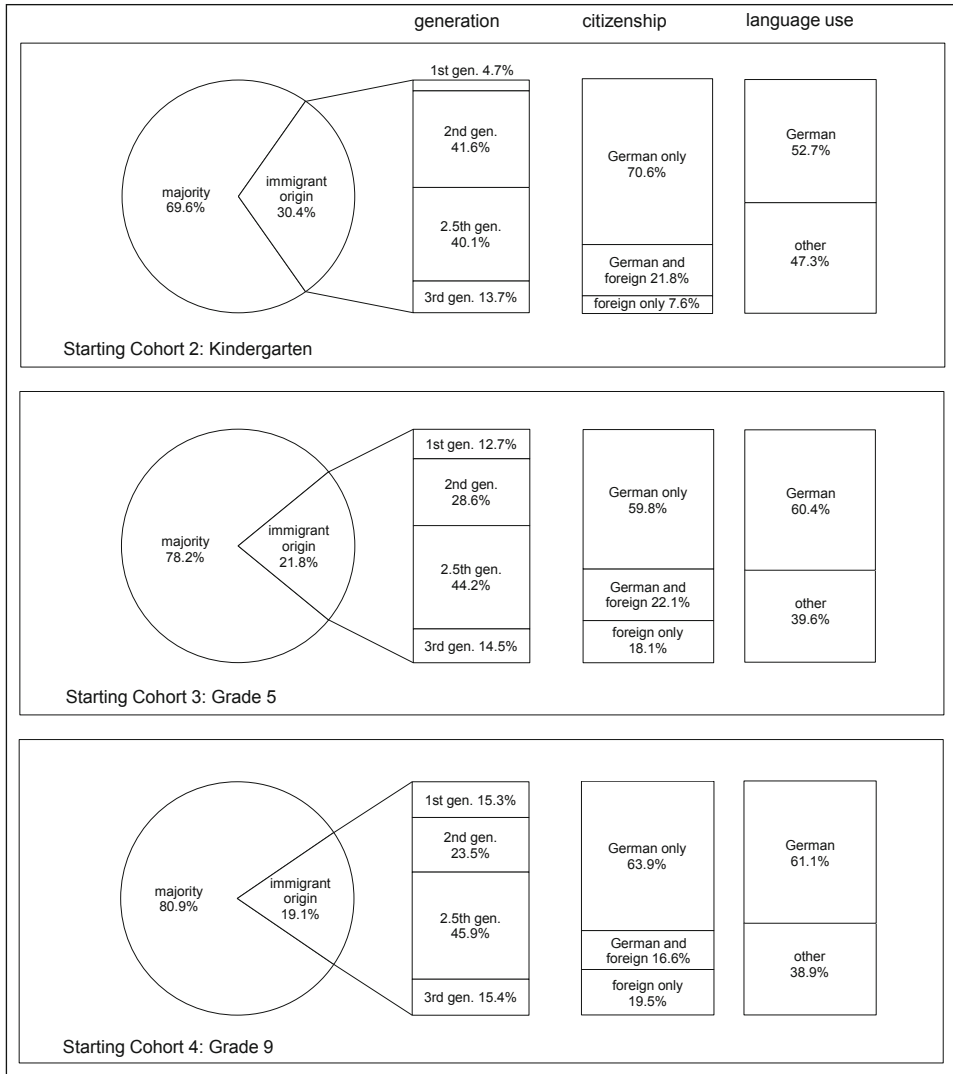
Note. Unweighted data.

4.2 How the Different Operationalizations Relate to Each Other

Figure 1 illustrates how the identification of the immigrant population via the country of birth relates to distributions according to generation status, citizenship, and language use. For each cohort, the pie charts depict the percentages of the immigrant population. For this population, the subsequent bar charts further specify the distributions for the different immigrant identifiers.

The first bar on generation status graphically illustrates the distributions described in the previous section. The second bar refers to nationality. Official statis-

Figure 1 The immigrant population according to generation status, citizenship, and language use



Note. Unweighted data.

tics still mostly rely on this characteristic and do not usually take dual citizenship into account. If, in the case of dual citizenship, individuals with a German nationality are assigned to the majority, this would imply that a much smaller share belongs to the immigrant population than would be the case if the assignment were based on the country of birth. In the youngest cohort, 92.4 % of all individuals who would be considered as persons of migrant origin via the country of birth would be part of the German majority. In the older cohorts, this share is about 10 % smaller. This difference across cohorts is mainly related to the amended citizenship law, which allows for dual citizenship by birth under certain conditions. Accordingly, the portion of children with only a foreign nationality is relatively small in the Kindergarten cohort (7.6 %) but increases in the older cohorts (to 18.1 % in Grade 5 and to 19.5 % in Grade 9) for children who were mostly born before the law changed.

When considering language use, the results indicate that the majority of immigrants in all starting cohorts only or mainly use German. The percentages range from 52.7 % in the Kindergarten cohort to 60.4 % in the Grade 5 cohort and up to 61.1 % in the Grade 9 cohort. Studies that consider language use as the key immigrant identifier therefore most probably focus on a selective population.

Table 2 provides a supplement to the graphical illustration in Figure 1. It specifies how the distributions according to generation status relate to operationalization via citizenship and language use. The numbers show once more that information on a foreign nationality is hardly suitable for detecting immigrant offspring. Foreign nationality works best for the first generation, which is most likely to possess a foreign passport. In the third generation, in contrast, having a non-German citizenship is less common. Another important finding is that the share of children and students who speak predominantly German at home increases across generations. German is the dominant language for more than 84 % of third-generation offspring. As expected, these shares decrease in the second, 2.5th, and first generation. Nevertheless, a substantive portion is apt to use German (29.9 %–45.5 %) also in the first generation.

4.3 The Immigrant Population from Turkey and the Former Soviet Union

The two largest migrant groups in Germany are individuals of Turkish origin and individuals whose families stem from the Former Soviet Union. Each group makes up a substantive share of the overall immigrant population in the three starting cohorts, ranging from 16 % to 23 %. Figure 2 illustrates the distributions for these two groups according to generation status, citizenship, and language use. Table 3 serves as a supplement to the graphical illustration.

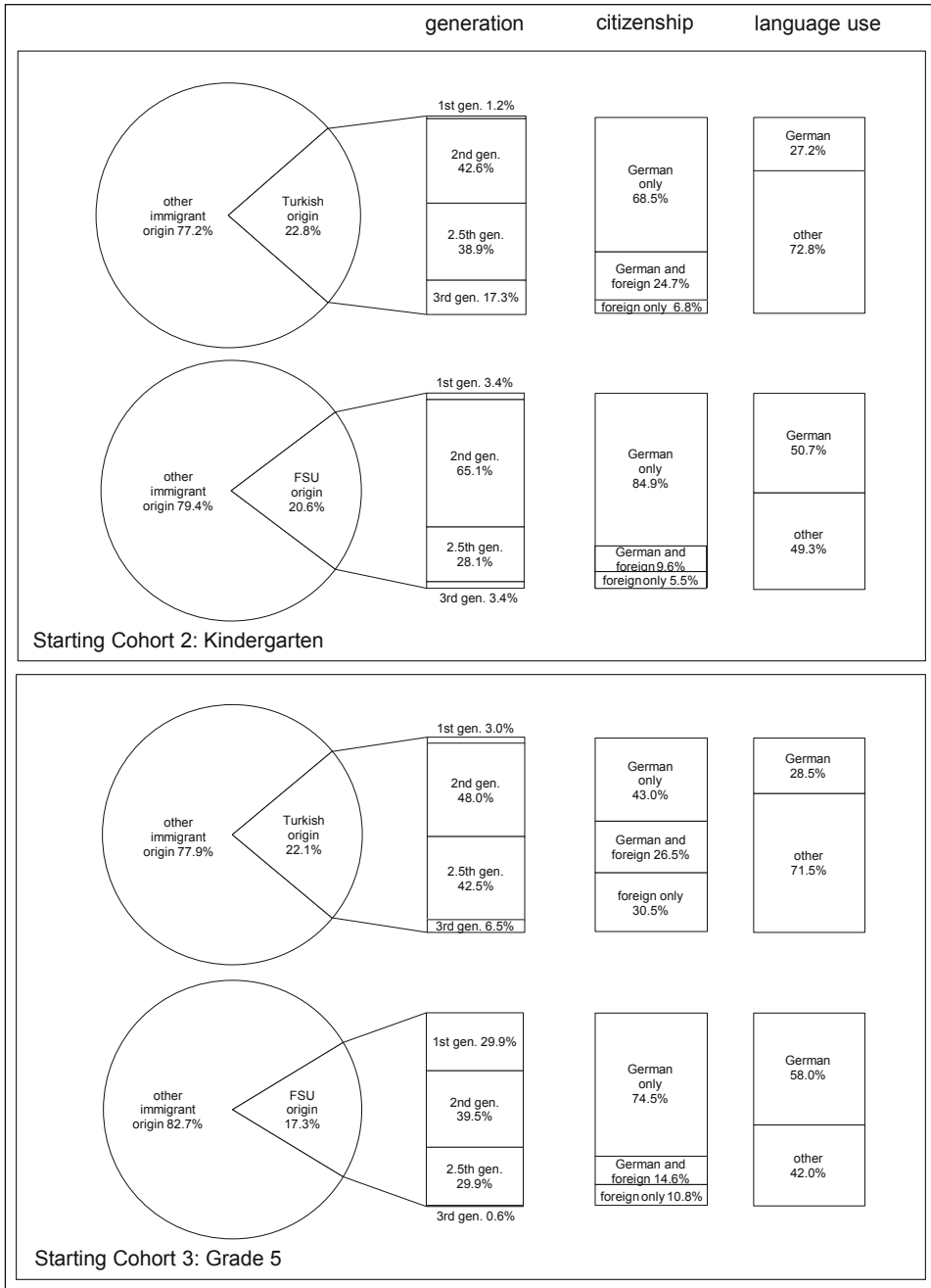
In both groups and all cohorts, most children and youth belong to the second and the 2.5th generation. In the Turkish population, the shares range from 81.5 % in the Kindergarten cohort to 90.5 % in Grade 5 and 92.4 % in Grade 9. For the offspring of

Table 2 Distributions of citizenship and language use according to generation status

	1st gen.		2nd gen.		2.5th gen.		3rd gen.	
	<i>N</i>	%	<i>N</i>	%	<i>N</i>	%	<i>N</i>	%
<i>Kindergarten</i>								
Citizenship								
German citizenship only	10	30.3	209	70.8	209	73.3	73	75.3
German and foreign citizenship	8	24.2	57	19.3	68	23.9	22	22.7
Foreign citizenship/s only	15	45.5	29	9.8	8	2.8	2	2.1
Total	33	100.0	295	100.0	285	100.0	97	100.0
Language use								
German language only/mainly	15	45.5	112	38.0	165	57.9	82	84.5
Other language only/mainly	18	54.5	183	62.0	120	42.1	15	15.5
Total	33	100.0	295	100.0	285	100.0	97	100.0
<i>Grade 5</i>								
Citizenship								
German citizenship only	42	36.5	156	60.2	247	61.8	96	73.3
German and foreign citizenship	36	31.3	43	16.6	98	24.5	23	17.6
Foreign citizenship/s only	37	32.2	60	23.2	55	13.8	12	9.2
Total	115	100.0	259	100.0	400	100.0	131	100.0
Language use								
German language only/mainly	48	41.7	110	42.5	277	69.3	112	85.5
Other language only/mainly	67	58.3	149	57.5	123	30.8	19	14.5
Total	115	100.0	259	100.0	400	100.0	131	100.0
<i>Grade 9</i>								
Citizenship								
German citizenship only	104	43.2	221	59.7	477	65.9	206	85.1
German and foreign citizenship	60	24.9	48	13.0	134	18.5	20	8.3
Foreign citizenship/s only	77	32.0	101	27.3	113	15.6	16	6.6
Total	241	100.0	370	100.0	724	100.0	242	100.0
Language use								
German language only/mainly	72	29.9	162	43.8	504	69.6	225	93.0
Other language only/mainly	169	70.1	208	56.2	220	30.4	17	7.0
Total	241	100.0	370	100.0	724	100.0	242	100.0

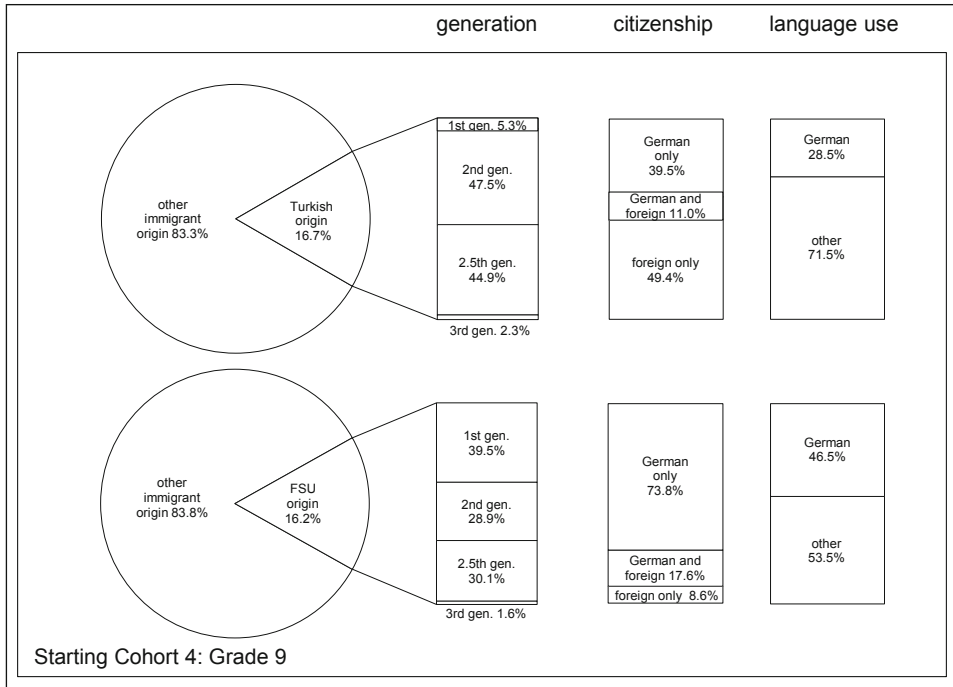
Note. Unweighted data.

Figure 2 Immigrant populations stemming from Turkey and the former Soviet Union according to generation status, citizenship, and language use



Note. Unweighted data.

Figure 2 (continued) Immigrant populations stemming from Turkey and the former Soviet Union according to generation status, citizenship, and language use



Note. Unweighted data.

FSU immigrants, the percentages are larger in Kindergartens (at 93.2%) and smaller in the two older cohorts (i.e., 69.4% in Grade 5 and 59.0% in Grade 9). The more recent history of immigration from the FSU since the 1990s compared with the labor immigration and subsequent family reunion of individuals from Turkey since the 1960s is also clearly visible when looking at the percentages of first- and third-generation offspring. Given the relatively longer migration history among Turks, the third generation amounts to a substantive share of 17.3% among Kindergarten children and decreases over cohorts to 6.5% in Grade 5 and 2.3% in Grade 9. The first generation in all cohorts, in contrast, is rather small in the Turkish group. For individuals of FSU origin, a reversed picture emerges: The third generation hardly exists, whereas the first generation is of considerable size in the two school cohorts, with 39.5% in Grade 9 and 29.9% in Grade 5.

When considering citizenship, the share of children who only possesses the German nationality is large among FSU offspring (i.e., 84.9% in Kindergarten, 74.5% in Grade 5, and 73.8% in Grade 9). Presumably, many of these children are Spätaussiedler or descendants of them. In the Turkish population, the percentages of Ger-

Table 3 The population of Turkish origin and from the former Soviet Union

	Kindergarten		Grade 5		Grade 9	
	<i>N</i>	%	<i>N</i>	%	<i>N</i>	%
<i>Turkish origin</i>						
Generation status						
1st generation	2	1.2	6	3.0	14	5.3
2nd generation	69	42.6	96	48.0	125	47.5
2.5th generation	63	38.9	85	42.5	118	44.9
3rd generation	28	17.3	13	6.5	6	2.3
Total	162	100.0	200	100.0	263	100.0
Citizenship						
German citizenship only	111	68.5	86	43.0	104	39.5
German and foreign citizenship	40	24.7	53	26.5	29	11.0
Foreign citizenship/s only	11	6.8	61	30.5	130	49.4
Total	162	100.0	200	100.0	263	100.0
Language use						
German language only/mainly	44	27.2	57	28.5	75	28.5
Other language only/mainly	118	72.8	143	71.5	188	71.5
Total	162	100.0	200	100.0	263	100.0
<i>From the Former Soviet Union</i>						
Generation status						
1st generation	5	3.4	47	29.9	101	39.5
2nd generation	95	65.1	62	39.5	74	28.9
2.5th generation	41	28.1	47	29.9	77	30.1
3rd generation	5	3.4	1	0.6	4	1.6
Total	146	100.0	157	100.0	256	100.0
Citizenship						
German citizenship only	124	84.9	117	74.5	189	73.8
German and foreign citizenship	14	9.6	23	14.6	45	17.6
Foreign citizenship/s only	8	5.5	17	10.8	22	8.6
Total	146	100.0	157	100.0	256	100.0
Language use						
German language only/mainly	74	50.7	91	58.0	119	46.5
Other language only/mainly	72	49.3	66	42.0	137	53.5
Total	146	100.0	157	100.0	256	100.0

Note. Unweighted data.

man nationals are smaller, ranging from 68.5 % in Kindergarten to 43.0 % in Grade 5 and 39.5 % in Grade 9. In the school cohorts, the proportions of children and youth with only a foreign citizenship are considerably larger in the Turkish group (at 30.5 % in Grade 5 and 49.4 % in Grade 9) compared with the percentages among students of FSU origins (10.8 % in Grade 5 and 8.6 % in Grade 9). Given the amended citizenship law, most kindergarten children in both migrant groups are German nationals (94.5 % among FSU offspring and 93.2 % among children from Turkish families). The most important conclusion that can be drawn from these results is probably that immigrant offspring would be more often assigned to the majority in the FSU group than in the Turkish group when applying nationality as the key identifier. Thus, using citizenship yields different distributional outcomes for the two groups.

Another discrepant pattern appears in the findings on language use. About half of all FSU children speak only or mainly German at home (i. e., 50.7 % in Kindergarten, 58.0 % in Grade 5, and 46.5 % in Grade 9). The situation is reversed for children of Turkish origin: The majority speaks only or mainly another language with at least one family member (72.8 % in Kindergarten, 71.5 % in the Grade 5, and 71.5 % in the Grade 9 cohort).

5 Conclusions

The NEPS provides a set of measures that can be used to identify immigrants and their children. The most important ways of assigning individuals to certain groups include operationalization based on the country of birth, citizenship, and language.

The findings from the Kindergarten-, Grade 5-, and Grade 9 cohort illustrate that the size of the immigrant population varies considerably for different immigrant identifiers. It is largest when focusing on the country of birth. In Kindergarten, about 30 % of all children are of migrant origin, whereas the percentages in the two older cohorts add up to about 20 %. When using assignments based on citizenship or language use instead of the country of birth, the migrant population shrinks substantially. These measures capture only about 50 % of the initially identified immigrants; most of them belong to the first and second generation.

Another important result is that not all identifiers seem to be equally adequate at detecting specific groups. While the population from the FSU would mostly be assigned to the majority when considering citizenship or language use, this is not the case to the same extent for the Turkish group. Hence, there is good reason to suspect that applying the same operationalization to these two groups does not allow for capturing them in a similar manner. More generally, certain ways of identifying immigrants and their offspring may be more or less selective with regard to the population of interest. Discrepancies of this kind are also related to the specific migration histories of the groups in question, for example, regarding the point in time at which individuals indicate German citizenship, which is clearly distinct for groups such as the

Spätaussiedler. It is necessary to keep these distinctive features in mind when studying the integration patterns of different migrant groups.

At the same time, it is clear that the adequacy of a certain operationalization depends on the research interest. The NEPS provides the necessary tools; however, the decision to consider one assignment as opposed to another is obviously up to the researcher.

References

- Bellin, N., Dunge, O., & Gunzenhauser, C. (2010). The importance of class composition for reading achievement: Migration background, social composition and instructional practices. An analysis of the German 2006 PIRLS data. In M. von Davier, & D. Hastedt (Eds.), *IERI Monograph Series: Issues and methodologies in large-scale assessment* (Vol. 3, pp. 9–34). Hamburg: IER Institute.
- Bundesministerium des Innern (Ed.). (2013). *Migrationsbericht des Bundesamtes für Migration und Flüchtlinge im Auftrag der Bundesregierung: Migrationsbericht 2011*. Nürnberg: Bundesamt für Migration und Flüchtlinge.
- Gresch, C., & Kristen, C. (2011): Staatsbürgerschaft oder Migrationshintergrund? Ein Vergleich unterschiedlicher Operationalisierungsweisen am Beispiel der Bildungsbeurteilung. *Zeitschrift für Soziologie*, 40(3), 208–227.
- Kristen, C. (2008). Schulische Leistungen von Kindern aus türkischen Familien am Ende der Grundschulzeit: Befunde aus der IGLU-Studie. In F. Kalter (Hrsg.), *Migration und Integration. Kölner Zeitschrift für Soziologie und Sozialpsychologie* (Sonderband 48, pp. 230–251). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Kristen, C., Edele, A., Kalter, F., Kogan, I., Schulz, B., Stanat, P., & Will, G. (2011). The education of migrants and their children across the life course. In H.-P. Blossfeld, H.-G. Roßbach, & J. von Maurice (Eds.), *Zeitschrift für Erziehungswissenschaft, 14. Education as a lifelong process: The German National Educational Panel Study (NEPS)* (pp. 121–138). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Mudiappa, M., & Kluczniok, K. (2015). Visits to cultural learning places in the early childhood. *European Early Childhood Education Journal* 23(2), 200–212.
- Olczyk, M., Will, G., & Kristen, C. (2014). *Immigrants in the NEPS: Identifying generation status and group of origin*. (NEPS Working Paper No. 41a). Bamberg: University of Bamberg, National Educational Panel Study.
- Stanat, P. (2006). Schulleistungen von Jugendlichen mit Migrationshintergrund: Die Rolle der Zusammensetzung der Schülerschaft. In J. Baumert, P. Stanat, & R. Watermann (Eds.), *Herkunftsbedingte Disparitäten im Bildungswesen: Differenzielle Bildungsprozesse und Probleme der Verteilungsgerechtigkeit. Vertiefende Analysen im Rahmen von PISA 2000* (pp. 189–219). Wiesbaden: VS Verlag für Sozialwissenschaften.

- Van der Slik, F. W. P., Driessen, G. W. J. M., & De Bot, K. L. J. (2006). Ethnic and socioeconomic class composition and language proficiency: A longitudinal multilevel examination in Dutch elementary schools. *European Sociological Review*, 22(3), 293–308.
- Wagner, W., Helmke, A., & Schrader, F.-W. (2009). Die Rekonstruktion der Übergangsempfehlung für die Sekundarstufe I und der Wahl des Bildungsgangs auf der Basis des Migrationsstatus, der sozialen Herkunft, der Schulleistung und schulklassenspezifischer Merkmale. In J. Baumert, K. Maaz, & U. Trautwein (Eds.), *Zeitschrift für Erziehungswissenschaft, 12. Bildungsentscheidungen* (pp. 183–204). Wiesbaden: VS Verlag für Sozialwissenschaften.

About the authors

C. Kristen
Department of Sociology, esp. Analysis of Social Structures,
University of Bamberg.

M. Olczyk
Leibniz Institute for Educational Trajectories (LifBi),
Bamberg.
e-mail: melanie.olczyk@lifbi.de

G. Will
Leibniz Institute for Educational Trajectories (LifBi),
Bamberg.

Measuring Health in a Longitudinal Education Study

Johann Carstensen, Anja Gottburgsen and Monika Jungbauer-Gans

Abstract

When analyzing health in an educational study, there are some methodological aspects and problems that must be considered. In this paper, we address questions of data quality in the measurement of health outcomes. It is possible that data quality can be biased by social desirability since questions on health (e. g., on eating disorders or body height and weight) are fairly sensitive items, and accordingly, the impact of the privacy of the setting increases with the sensitivity of the questions. Therefore, we expect mode effects resulting from the way the data are collected. Following a methodological discussion of these issues, empirical analyses are presented. We compare the measuring of body height, weight, BMI, and the likelihood of having an eating disorder in the NEPS with data from reference studies (KiGGS and GEDA from 2010) carried out by the Robert Koch Institute. To conduct the analysis of BMI, we use the Kindergarten cohort, the ninth graders, and the adults' cohort. The eating disorder scale is compared for ninth graders only. The results show some differences between NEPS data and the reference data, which point towards an influence of the interview situation. In about half of our comparisons, no significant deviations between the datasets can be found. A short section describes some further thoughts on endogeneity problems.

1 Introduction

Social epidemiological studies clearly indicate that education, just like income and occupational status, strongly correlates with health. The higher the educational qualification, the lower are mortality and morbidity (Cutler & Lleras-Muney, 2012; Mielck, 2008; OECD, 2006). Higher education has positive effects on occupational conditions and financial returns, on psychosocial resources, and, to a large extent, on health be-

havior, which then results in “better” health (Mackenbach, 2006; Ross & Wu, 1995). Focusing on education as an important indicator of social status to evaluate the social gradient of health is advantageous for adults. Education and social status are acquired early in life and remain relatively stable over the course of individuals’ lives. However, if their social status changes due to unemployment, this may also be a result of illness. Therefore, studies analyzing the effect of social status on health regularly face the question of cause and effect, or rather, the problem of endogeneity (Siegrist & Marmot, 2006; Kimbro, Bzostek, Goldman, & Rodríguez, 2008).

Why measure health in an education study? The fact that a poor state of health or health-related risk behavior can have detrimental effects on educational achievement (the acquisition of competencies and qualifications, grades, as well as long-term outcomes such as the acquisition of academic degrees) has long been neglected, especially in German research. Recent international overviews of a number of longitudinal studies by Suhrcke and de Paz Nieves (2011), Dadaczynski (2012), and Basch (2011) indicate such a correlation with respect to physical fitness, overweight/obesity, sleeping disorders, and risk behavior (smoking, malnutrition), whereas the findings on mental health (anxiety, depression, ADHD, behavioral disorders) are heterogeneous. There is increasing evidence that health is important for educational results, but it is clearly confounded with other variables, such as parents’ social status (Basch, 2011), gender (Dadaczynski, 2012), and ethnic/cultural origin (Kimbro et al., 2008).

Health depends to a large extent on developmental age and stage: To examine the relationship of cause and effect, long time periods must be taken into consideration (Dragano & Siegrist, 2009; Power & Kuh, 2006). For example, depending on social status, genetic determinants, conditions of socialization, and healthcare, it is possible for the effects of a low birthweight to first manifest themselves in late adulthood—a fact that makes it very difficult to identify causal relationships (Dragano & Siegrist, 2009). The multisequential cohort design applied in the National Educational Panel Study (NEPS) spans the entire life course, from birth to late adulthood. Pooling data from different cohorts offers insights into the development of health under the prerequisite that the measurement be comparable for different age groups.

However, the comparison of different age groups also causes methodological problems, as the example of the body mass index for assessing overweight/obesity shows. For adults, definitions of overweight (Body Mass Index/BMI > 25) and obesity (BMI > 30) have been accepted worldwide. For children and adolescents, however, changes in body mass due to age, developmental stage, and gender do not allow for rigid threshold values, as in adulthood. Therefore, the definitions of underweight and overweight are usually based on age- and gender-related percentiles of a reference population (Cole, Bellizzi, Flegal, & Dietz, 2000; Kurth & Schaffrath-Rosario, 2007). The investigations in the framework of the NEPS are thus faced with age-related norming problems. These problems can be solved by validating the quality of data in comparison with the data of the Robert-Koch Institute, which serve as a reference system.

The NEPS offers a unique opportunity to examine the reciprocal causality between education (or other factors such as socioeconomic status) and health (Basch, 2011; Suhrcke & de Paz Nieves, 2011). A number of methodological aspects and problems must be considered when analyzing health in the framework of an educational study such as the NEPS. These are, for example, questions concerning the already-mentioned endogeneity, the dependence of health on development and age, as well as surveying effects that result from the sensitivity of health-related questions (i. e., weight and height, self-rated health, eating disorders) and from the respective mode of surveying. In the following section, we first discuss methodological issues, that is, effects of the mode of data collection and social desirability. After this, we address these questions of data quality empirically by using NEPS data. The next section describes some further thoughts on endogeneity problems and how they might be solved.

2 Effects of the Mode of Data Collection and Social Desirability

In the NEPS, questions on health (e. g., on eating disorders or body height and weight) are fairly sensitive items (Tourangeau, Rips, & Rasinski, 2000). Answering them could be unpleasant or embarrassing, and respondents might therefore refrain from answering (“non-response”) or modify their answers in the direction of social desirability to present themselves in a more favorable light (“social desirability bias”). Effects of social desirability can lead to systematic distortions of responses, for example, to an overreporting of socially approved behaviors (e. g., voting) and an underreporting of socially disapproved behaviors (e. g., using illicit drugs) (Groves, Fowler, Lepkowski, Singer, & Tourangeau, 2004; Tourangeau & Smith, 1996). The quality of data is highly dependent on the way in which it was gathered in respect to both completeness and the extent of interference of the social desirability bias.

The respondents’ degree of privacy depends on whether the data were collected in an interview (face-to-face vs. telephone), in a written survey (paper and pencil interview/PAPI vs. online), or self-administered vs. non self-administered (computer-assisted personal interview/CAPI vs. computer-assisted self interview/CASI) (Groves et al., 2004; Tourangeau & Smith, 1996). The impact of the privacy of the setting increases with the sensitivity of the questions. This, in turn, increases the effects of the mode of interrogation on the answering behavior of respondents in that questions may not be answered at all or in a biased way. Compared with self-administered procedures, interviews yield a high completeness of data but seem to evoke stronger effects of the social desirability bias (Groves et al., 2004; Tourangeau & Smith, 1996).

The effects of social desirability and of the mode of data collection must be considered with respect to body height and, particularly, to weight. Many studies have shown that there is a substantial difference between objective measures and survey data concerning body weight and height, which is additionally mediated by the concrete interview situation (e. g., Béland & St-Pierre, 2008; Shields, Grober, & Tremblay,

2008; see also Glaesmer & Brähler, 2002; Kroh, 2004 for Germany). Critical influences on measurement error are the anonymity of the interview situation, the interviewer, repeated measures, and the tendency to provide round numbers in an interview (Kroh, 2004). Compared with measurements, there is evidence for systematic misjudgments of body weight (underreporting) in self-administered formats (Visscher, Viet, Kroesbergen, & Seidell, 2006) depending on the characteristics of the interviewer and his or her presence (Kroh, 2004). All these studies point towards the fact that interview-assessed information on body weight and height delivers data that result in a lower BMI than does using objective measures. This means that respondents overestimate their height and underestimate their weight. In the NEPS, data on BMI are gathered via different self-administered or interviewer-based modes of data collection. Therefore, the quality of data on this item is also tested through comparison with measurements by medical personnel in order to quantify the bias caused by self-reporting and to identify potential correcting factors. We compare our subjective data below to objective data from an external source for the younger cohorts and to other survey data for the adult cohort.

3 Empirical Analysis of Data Quality

3.1 Body Mass Index

The BMI is an anthropometric measure for the relation of body weight to height. In the NEPS, it is calculated from the self-reported information of these values. The formula is given by $BMI = \frac{m}{l^2}$, where m is the body mass measured in kilograms and l is the body height measured in meters. The underlying information of the BMI is, as stated above, highly prone to mode-effects. Having a low BMI appears to be more socially desirable than having a BMI that indicates obesity. Which survey mode is used to ask the respondents about their body weight and height is therefore expected to matter.

We use NEPS data from Starting Cohorts 2 and 4 (NEPS SC2, version 1.0.0,¹ and SC4, version 1.0.0²), from Kindergarten and Grade 9, respectively, to compare this information with data from the German Health Survey for Children and Adolescent (*Studie zur Gesundheit von Kindern und Jugendlichen in Deutschland*, KiGGS).³ The KiGGS study was conducted by the German Robert Koch Institute from 2003 to 2006. In comparison with the NEPS, the KiGGS is wholly a health survey that contains not only interviews of children and parents, but also medical examinations and blood

1 Doi:10.5157/NEPS:SC2:1.0.0.

2 Doi:10.5157/NEPS:SC4:1.0.0.

3 Public Use File KiGGS, The German Health Survey for Children and Adolescents 2003–2006, Robert Koch Institute, Berlin (Germany), 2008.

tests. Thus, in this survey, body height and weight are measured with tape and scales. Although there are still other sources of error, the KiGGS could thereby rule out all the potential errors resulting from an interview situation. For this reason, we take the KiGGS data to compare with our data collected through survey interviews. The KiGGS data are collected via a two-stage cluster sampling procedure in which a fixed number of children of every age-group are sampled. The response rate of the KiGGS is 66.6 %. Item-nonresponse on the measures we are interested in is practically non-existent based on the data collection method described above.

The NEPS contains a multi-informant perspective. In younger age groups, most of the information on target children is gained through parent interviews. As the children grow older, they are interviewed in person. The information from Starting Cohort 2 is based on reports of the parents collected via computer-assisted telephone interviews (CATI), whereas the data from ninth graders stems from paper-and-pencil interviews of the target persons, which were administered in a class setting. The sample of the Starting Cohort 2 is drawn using a rather complex procedure: To account for the institutional level, in a first step, daycare facilities are sampled, whereby the probability of going to certain NEPS schools is also considered. The response rate is rather high (82.3 %). Item-nonresponse on anthropometric measures is moderate, with body height being the most frequently unanswered question (2.1 %). Starting Cohort 4 is sampled using a stratified cluster sample in which schools of different types are first drawn corresponding to the frequency of the school types in Germany. Within these schools, school classes are drawn using simple random sampling. All students of one school class are then surveyed. The response rate here is lower than in Starting Cohort 2 (58.96 %). In this cohort, item-nonresponse is a serious problem, especially concerning the question about body weight (11.7 %). This aspect points towards a social desirability issue, which is especially existent in this age-group.

As stated above, the World Health Organization's (WHO) definitions of obesity and underweight based on the BMI are not appropriate for the application to under-age respondents. Therefore, it is common to rely on a reference population and define obesity and underweight by age-specific percentiles. We use the age-specific BMI reference percentiles by Kromeyer-Hauschild et al. (2001).⁴ Every Person with a BMI > 97 % of the age-specific reference population is defined as obese, every person with a BMI < 3 % of the age-specific distribution is defined as strongly underweight, and so forth.

Table 1 shows proportions of the BMI classes with corresponding confidence intervals in brackets for the NEPS Starting Cohort 2⁵ and the respective age group from the KiGGS. Design weights were used for both datasets. It is clearly visible that the group of normal-weight participants is relatively small in the NEPS compared with the data from the Robert Koch Institute. In total, the share of overweight and obese

4 < P3 and P3 to < P10, and > P90 to P97 and > P97.

5 Fifteen cases were excluded from the analyses due to implausibility.

Table 1 BMI Categories in NEPS and KiGGS (Kindergarten)

	NEPS SC 2 (Kindergarten)			KiGGS (age 4–6)		
	Male	Female	Total	Male	Female	Total
Strongly underweight	0.1097 (0.0878, 0.1361)	0.1281 (0.1029, 0.1583)	0.1189 (0.1018, 0.1385)	0.0183 (0.0114, 0.0292)	0.0174 (0.0108, 0.0278)	0.0179 (0.0128, 0.0249)
Underweight	0.0646 (0.0485, 0.0854)	0.0906 (0.0711, 0.1147)	0.0777 (0.0646, 0.0931)	0.0446 (0.0331, 0.0598)	0.0409 (0.0293, 0.0569)	0.0428 (0.0343, 0.0533)
Normal weight	0.7007 (0.6643, 0.7348)	0.7031 (0.6663, 0.7374)	0.7019 (0.6763, 0.7263)	0.8596 (0.8353, 0.8809)	0.8599 (0.8354, 0.8812)	0.8597 (0.8428, 0.8751)
Overweight	0.0781 (0.0588, 0.1032)	0.0438 (0.0309, 0.0617)	0.0609 (0.0488, 0.0758)	0.0565 (0.0428, 0.0743)	0.0471 (0.035, 0.0629)	0.0519 (0.0424, 0.0634)
Obese	0.0469 (0.0333, 0.0658)	0.0344 (0.023, 0.0512)	0.0406 (0.0313, 0.0526)	0.0209 (0.0135, 0.0323)	0.0348 (0.0248, 0.0487)	0.0277 (0.0212, 0.0362)
	1	1	1	1	1	1
	N = 2242			N = 2194		

Note. Reference percentiles: Kromeyer-Hauschild et al., (2001). Confidence intervals in brackets.

participants (10%) is quite close to that of the reference population by Kromeyer-Hauschild et al. (2001), but it differs when analyzed separately by gender. For girls, the accumulated share of overweight persons is lower than expected (7.8%), which is not substantively different from the KiGGS data, while for boys, it is higher (12.5%) than in the reference population and also higher than in the KiGGS data. The far more noticeable problem appears in the other direction: The shares of underweight and strongly underweight boys and girls in the NEPS are very high. We come back to this point in the regression analysis below.

Table 2 holds a similar listing for Starting Cohort 4⁶ of the NEPS compared with the group of 12-to-18-year-olds in the KiGGS. As above, these data were weighted with survey design weights.

At first glance, these data do not pose as much of a problem as do those from the younger cohort. The rates of underweight and strongly underweight participants are

6 Sixty-four cases were excluded from the analyses due to implausibility.

Table 2 BMI Categories in NEPS and KiGGS (Grade 9)

	NEPS SC4 (Grade 9)			KiGGS (age 12–18)		
	Male	Female	Total	Male	Female	Total
Strong underweight	0.0231 (0.0195, 0.0273)	0.036 (0.0313, 0.0415)	0.0292 (0.0262, 0.0325)	0.0183 (0.0114, 0.0292)	0.0174 (0.0108, 0.0278)	0.0179 (0.0128, 0.0249)
Underweight	0.046 (0.0408, 0.0519)	0.0746 (0.0677, 0.082)	0.0595 (0.0552, 0.0641)	0.0446 (0.0331, 0.0598)	0.0409 (0.0293, 0.0569)	0.0428 (0.0343, 0.0533)
Normal weight	0.8027 (0.7922, 0.8129)	0.8234 (0.8129, 0.8335)	0.8125 (0.8051, 0.8197)	0.8596 (0.8353, 0.8809)	0.8599 (0.8354, 0.8812)	0.8597 (0.8428, 0.8751)
Overweight	0.0821 (0.0752, 0.0896)	0.0346 (0.0301, 0.0398)	0.0598 (0.0554, 0.0644)	0.0565 (.0428, 0.0743)	0.0471 (0.035, 0.0629)	0.0519 (0.0424, 0.0634)
Obese	0.046 (0.041, 0.0516)	0.0314 (0.027, 0.0363)	0.0391 (0.0357, 0.0428)	0.0209 (0.0135, 0.0323)	0.0348 (0.0248, 0.0487)	0.0277 (0.0212, 0.0362)
	1	1	1	1	1	1
	N = 12071			N = 5723		

Note. Reference percentiles: Kromeyer-Hauschild et al., (2001). Confidence intervals in brackets.

again slightly higher in the NEPS than in the KiGGS measurement for both girls and boys. The prevalence of accumulated underweight is about 3 percentage points higher in the NEPS data in total. The rates for girls, however, are noticeably higher in the NEPS. This overreporting is in line with expectations from the social desirability theory regarding the increase in eating disorders and disturbed body images within the age group in question. Following this interpretation, girls would tend to “correct” their body height and weight to a desirable ratio, which would lead to a higher share of abnormal BMIs.

The upper range of the scale deals with smaller irregularities. In the NEPS, males especially tend to report a higher prevalence of obesity than what is measured by the Robert Koch Institute in the same age group, whereas females report slightly lower rates of overweight and obesity. Since the tendency of the boys to report higher BMI is contrary to the assumptions made concerning the social desirability bias, this is a rather paradoxical finding. Besides various sources of measurement error, the fact that the KiGGS data are from the period of 2003 to 2006 while the NEPS data are from

2011 must be taken into account. The increase in overweight and obesity as well as in eating disorders, in turn leading to more extreme body proportions, could be at least partially responsible for the depicted image.

For the adult population, we can rely on a much simpler classification of underweight, overweight, and obesity following the definitions of the WHO (2000). The threshold values are as follows: underweight (< 18.5), normal weight ($18.5\text{--}24.9$), overweight ($25\text{--}29.9$), Obese Class I ($30\text{--}34.9$), Obese Class II ($35\text{--}39.9$), and Obese Class III (≥ 40) (WHO, 2000).

In the case of the adults, our reference data stem from the German Health Update (*Gesundheit in Deutschland aktuell*, GEDA 2010),⁷ again carried out by the Robert Koch Institute. The GEDA is a telephone survey with random-digit dial sampling. Contrary to the KiGGS, there is not any form of objective measurement of anthropometric data. The information is self-rated, as in the NEPS. A possible difference here could be the context of a health survey, which probably increases the acceptance of health-related questions. On the other hand, there could be a form of self-selection in advance so that people with serious health problems or undesirable health behavior would not participate in the survey. Sampling in the GEDA is conducted by the so-called Gabler-Häder method, a special form of random-digit dialing for telephone surveys. Since it operates through cold contacting, the response rate is expectably low (28.9%). The data of the fourth wave of Starting Cohort 6⁸ was collected by a stratified two-stage sampling procedure using data from communities' registration offices. Anthropometric questions were only administered to panel respondents who had participated in the previous waves. The response rate for this group in the fourth wave was 78%.

The Robert Koch Institute provides design weights with the data as well as population-based calibration weights. We use the calibration weights in parallel with those of the NEPS in Starting Cohort 6, which were generated using calibration factors from the German micro census 2011.

The listed proportion of BMI classifications shows some differences, especially around the middle of the BMI distribution. In the NEPS, a smaller proportion of individuals are of normal weight than in the GEDA study. The latter, in addition, provides slightly higher rates of underweight individuals. In contrast, people in the NEPS obviously report higher BMIs. The proportion of overweight and obese participants in the NEPS is higher than in the GEDA study in all obesity classes.

After inspecting the descriptive results, we applied regression models using the mode of data collection (survey data vs. objective measure) as a dummy variable. Because the sampling procedures differ in the NEPS and the KiGGS with respect to GEDA 2010, it was impossible to pool the data across surveys and still perform weighted analyses. However, we were able to control for all variables relevant to sam-

7 Public Use File GEDA 2010, Robert Koch Institute, Berlin (Germany), 2012.

8 Doi:10.5157/NEPS:SC6:5.1.0

Table 3 BMI Categories in NEPS and GEDA 2010 (Adults)

	NEPS SC6			GEDA 2010 (age 18–69)		
	Male	Female	Total	Male	Female	Total
Underweight	0.0034 (0.0015, 0.008)	0.0227 (0.0165, 0.0312)	0.013 (0.0096, 0.0174)	0.0084 (0.0064, 0.011)	0.0374 (0.0335, 0.0418)	0.0226 (0.0204, 0.0251)
Normal weight	0.3644 (0.3459, 0.3833)	0.4907 (0.4716, 0.5098)	0.4267 (0.4134, 0.4402)	0.4028 (0.3898, 0.416)	0.5567 (0.5447, 0.5686)	0.4784 (0.4694, 0.4874)
Overweight	0.4634 (0.4438, 0.4831)	0.3152 (0.2952, 0.3358)	0.3903 (0.3758, 0.4049)	0.4299 (0.4164, 0.4434)	0.2623 (0.2518, 0.2731)	0.3476 (0.3389, 0.3564)
Obese class I	0.1281 (0.1157, 0.1417)	0.1213 (0.1093, 0.1345)	0.1248 (0.1158, 0.1344)	0.1246 (0.1155, 0.1343)	0.0982 (0.0908, 0.1061)	0.1116 (0.1057, 0.1179)
Obese class II	0.0305 (0.0246, 0.0378)	0.0352 (0.0288, 0.043)	0.0328 (0.0284, 0.038)	0.0263 (0.0221, 0.0314)	0.031 (0.0266, 0.036)	0.0286 (0.0255, 0.0321)
Obese class III	0.0101 (0.0065, 0.0158)	0.0149 (0.01, 0.022)	0.0125 (0.0093, 0.0166)	0.008 (0.0057, 0.0111)	0.0144 (0.0116, 0.018)	0.0111 (0.0093, 0.0134)
	1	1	1	1	1	1
	N = 8780			N = 19136		

Note. Categories defined by the WHO (2000). Confidence intervals in brackets.

ple weights in the case of the KiGGS and for most of the relevant sampling variables in case of GEDA 2010. Since GEDA 2010 is a telephone survey using random-digit dialing, the number of telephone landlines per household was used to calculate survey weights to account for different inclusion probabilities. This information is lacking in the NEPS data, and we were thus not able to account for this difference in sampling frames.

Table 4 shows the results of four regression models using body height, body weight, BMI, and the squared difference of the BMI from the ideal BMI as dependent variables for Starting Cohort 2. The ideal BMI was calculated using the P50 percentile of the reference sample by Kromeyer-Hauschild et al. (2001). It can be seen that the reason for the unusual distribution of the BMI shown above lies in the information about body weight since the body height data show no significant difference to the

Table 4 OLS Regression Models for Height, Weight, and BMI (Kindergarten)

SC 2 vs. KiGGS	Body height (cm)	Body weight (kg)	BMI	BMI (sq. diff. to ideal BMI)
Mode of data collection (1 = objective measurement)	-0.121 (-0.737)	0.627*** (-6.786)	0.456*** (-7.990)	-1.779*** (-4.643)
Constant	77.392*** (-116.207)	6.238*** (-14.273)	14.884*** (-60.178)	0.746 (-0.435)
R^2	0.350	0.197	0.020	0.006
N	4672	4677	4626	4626

Note. OLS-Regressions with clustered standard errors, t-values in brackets. Control variables: age (exact), region (East vs. West Germany), gender, nationality (German).

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

measured data from the KiGGS. Parents in the NEPS report a body weight of Kindergarten children that is on average 0.63 kg lower than what is measured in the KiGGS. The mode of data collection is significant for both BMI and the squared deviance from ideal BMI. The latter is an indicator for parents in the NEPS who report extreme values of body weight and height slightly more often, leading to a higher frequency of extreme BMIs in total. A possible explanation of the underreporting of weight could be an answering bias towards lower body weight mainly affecting parents of children with regular BMIs. It is thus not possible to rule out that this effect is caused by problems in the data collection via CATI or differences in unit-nonresponse.

In Starting Cohort 4, a somewhat different image appears. Participants report highly significant higher body height and significantly higher body weight in the NEPS. The effect of mode on body height points in the expected direction since a tall body height appears to be more socially desirable than a short height. Contrary to the results for the Kindergarten cohort, where the typical image of underreporting body weight or overreporting body height is shown, the ninth graders in the NEPS report higher weight than those in the KiGGS reference data. The third model indicates that these two contradicting effects cancel each other out, resulting in equal BMI values. However, in the last model, it is obvious that there is still some impact. Since the squared deviance does not differentiate between negative and positive values and accounts more for extreme values, it can be shown that the NEPS and the KiGGS differ mostly in the more extreme values of the BMI. In total, the results for Starting Cohort 4 do not seem to be mainly driven by social desirability since deviations from normal weight appear more frequently in both directions. Other forms of measurement error or sample selection might play a larger role here.

Table 5 OLS Regression Models for Height, Weight, and BMI (Grade 9)

SC4 vs. KiGGS	Body height (cm)	Body weight (kg)	BMI	BMI (sq. diff. to ideal BMI)
Mode of data collection (1 = objective measurement)	-3.558*** (-20.631)	-0.791* (-2.276)	0.055 (-0.321)	-120.896*** (-3.752)
Constant	131.481*** (-132.373)	8.886*** (-4.355)	12.570*** (-12.536)	118.455 (-0.627)
R^2	0.295	0.104	0.009	0.001
N	12991	12445	12380	12380

Note. OLS-regressions with clustered standard errors, t-values in brackets. Control variables: age (exact), region (East vs. West Germany), gender, nationality (German).

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

Lastly, we fitted the same four regression models for the adult cohort, comparing them to the data from GEDA 2010. Here, we calculated the ideal BMI for the fourth regression model using the group mean value of the normal weight range ($BMI_{ideal} = 21.7$).

Since both data bases are survey data, we do not expect them to differ as much as the comparisons above. Both datasets contain data (mostly) collected via tele-

Table 6 OLS Regression Models for Height, Weight, and BMI (Adults)

SC6 vs. GEDA 2010	Body height (cm)	Body weight (kg)	BMI	BMI (sq. diff. to ideal BMI)
Mode of data collection (1 = GEDA2010)	-0.574*** (-6.657)	-1.058*** (-5.671)	-0.183** (-3.126)	-1.680 (-1.495)
Constant	177.703*** (701.604)	85.082*** (159.259)	26.885*** (158.156)	45.993*** (16.332)
R^2	0.522	0.270	0.098	0.022
N	28113	27671	27662	27662

Note. OLS-regressions with clustered standard errors, t-values in brackets. Control variables: number of adult household members, nationality (German), region (7 clusters based on Bundesländer), gender, education (years of schooling based on CASMIN), interaction age*education, interaction age*gender.

* $p < 0.05$, ** $p < 0.05$, *** $p < 0.001$

phone interviews. The regression models, however, show significant effects for all parameters except for the squared difference from the ideal BMI. People report greater heights and heavier body weights in the adult cohort of the NEPS than in the GEDA. The values do result in unequal BMIs that still differ by 0.18 units between the two studies. Again, we have higher reported BMIs in the NEPS, while the difference in squared deviances from the ideal BMI is not significant. This finding indicates that participants in the NEPS do not report extreme values more often than in the GEDA. The differences we found are, in this case, not as easy to interpret as above. As already stated, they could be a result of a change in the true value of the population—since the relevant NEPS study was carried out one year after the GEDA 2010—or of artifacts resulting from differences in the measurement process (interview form) or the sampling procedure and response error. After all, the results do not point towards increased problems concerning the social desirability bias since respondents in SC6 reported figures leading to even higher BMIs than in GEDA. This finding, as in Starting Cohort 4, points more towards different forms of measurement error or sample selection than towards social desirability. Moreover, a possible study effect based on the main topics of the surveys would be expected to show a reversed effect since acceptance of health-related intrusive questions should be higher in a health survey and answers should thus be less affected by social desirability.

3.2 Eating Disorder: SCOFF

“Eating disorders” are measured in the NEPS with the help of the sick-control-one stone-fat food scale (SCOFF scale), a well-validated screening instrument developed for the clinical assessment of such behavior. The measurement of risky eating behavior belongs to the aforementioned sensitive questions in the NEPS and can be affected by both the social desirability bias and mode effects, as well as their interaction. The NEPS data were gathered via self-administered surveys in the classroom (PAPI). This interview situation could have affected response behavior. Although children in the test situation are reassured that neither teachers nor interviewers will see their answers on the test and that they cannot be viewed by other students, the group context could create a form of social control that leads to a kind of uninformed social desirability. Therefore, the quality of NEPS data is checked in comparison with KIGGS data, which were also collected via self-reporting and in written form, but in an individual setting (Kurth, 2007).

The SCOFF screening tool (Morgan, Reid, & Lacey, 1999) was developed not as a survey instrument, but as a clinical screening tool for the identification of possible cases of bulimia and anorexia nervosa. It therefore has a high sensitivity (e.g., Perry et al., 2002) but is nevertheless often used as a survey instrument due to its simple application and robust psychometric properties. However, there is, of course, a difference between those forms of application in terms of data quality and measurement

error. The SCOFF questionnaire consists of five questions that can be answered with yes and no. Two or more positive answers indicate a possible case of anorexia nervosa or bulimia nervosa.

We use NEPS data from Starting Cohort 4 (NEPS SC4, version 1.0.0).⁹ Below, we again compare our SCOFF data to data from the KiGGS survey from the Robert Koch Institute to make statements about data quality. In comparison with the analyses above, there is no objective measurement here in either of the surveys. Both surveys were carried out as self-administered PAPI questionnaires. However, a possible difference could result from the specific context: NEPS is mainly an educational study with a focus on school- and education-related topics, whereas the KiGGS contains mostly questions about individual health. For the NEPS, this could result in a minor acceptance of sensitive items about health. Furthermore, the targets are interviewed in classrooms surrounded by their peers, which could evoke measurement biases.

Table 7 shows the weighted proportions of suspected cases of eating disorders obtained by the SCOFF questionnaire for females and males in the respective groups. The KiGGS finds a proportion of 22 % in total providing at least two positive answers and therefore indicating a possible case of anorexia nervosa or bulimia nervosa. As expected, the difference between boys and girls is considerable: Girls are twice as likely to display indications of possible anorexia or bulimia as are boys. This relation also holds in the NEPS data, but at a higher level: Here, 26 % of all participants give two or more positive answers. 35 % of the female participants are identified with a possible eating disorder.

Table 7 SCOFF in NEPS and KiGGS (Grade 9)

	NEPS SC4			KiGGS (age 12–18)		
	Male	Female	Total	Male	Female	Total
Suspicion	0.181 (0.1716, 0.1910)	0.355 (0.3432, 0.3674)	0.266 (0.2585, 0.2742)	0.146 (0.1326, 0.1605)	0.301 (0.2823, 0.3205)	0.222 (0.2100, 0.2339)
No suspicion	0.819 (0.8090, 0.8284)	0.645 (0.6326, 0.6568)	0.734 (0.7258, 0.7415)	0.854 (0.8395, 0.8674)	0.699 (0.6795, 0.7177)	0.778 (0.7661, 0.7900)
	1	1	1	1	1	1
	N = 12071			N = 5723		

Note. Suspected cases of eating disorders (two and more positive answers). Confidence intervals in brackets.

9 Doi:10.5157/NEPS:SC4:1.0.0.

Table 8 Logistic Regression Model for SCOFF (Grade 9, AME)

SC4 vs. KiGGS	SCOFF: suspected case of eating disorder
Mode of data collection (1 = KiGGS)	−0.026*** (−4.049)
Pseudo R^2	0.038
<i>N</i>	13416

Note. Logistic regression, coefficients: average marginal effects, z-values in brackets. Control variables: age (exact), region (East vs. West Germany), gender, nationality (German).

* $p < 0.05$, ** $p < 0.05$, *** $p < 0.001$

The same picture is given by a logistic regression using the pooled data of the KiGGS and the NEPS. Participants of the NEPS show on average a 2.6 % higher probability of answering the SCOFF in a manner that indicates the possible prevalence of bulimia nervosa or anorexia nervosa. This relationship is highly significant.

There are various reasons to explain this occurrence. As in all comparisons made in this paper, one possible explanation is that sample selection works differently in a health survey than in an educational survey. People with undesirable health behavior could more frequently be cases of unit nonresponse in a health survey, whereas these same people may not necessarily react to unexpected health questions in an educational survey through item nonresponse. In this case, a systematic sample selection bias could explain the higher prevalence of adverse health behavior found in the NEPS data. Again, another possible explanation is the time lag between both studies. Eating disorders have often been found to be on the rise amongst teenagers and adolescents. As for obesity, the differences in the data could be caused by real differences in the populations.

Lastly, we would like to stress that the analyses above cannot account for possible errors resulting from the sample selection and unit nonresponse. Questions such as these can only be addressed in mode effects studies using, for example, experimental methods.

4 Educational effects on health?

In this section, we do not intend to analyze a substantial question or present empirical results about the effects of education on health; rather, we wish to point to methodological issues that may prove helpful should one pursue this objective. The NEPS Pillar 5—Returns to Education Over the Life Course is predominantly concerned with the impact of education on other spheres of life, such as health, participation, deviant behavior, and family generation and fertility (Gross, Jobst, Jungbauer-Gans, &

Schwarze, 2011). While the direction of causal effects seems to be clear for social and political participation and family generation, the opposite influence of health on educational achievement can also be assumed, as explained in the introductory section above. Empirical evidence in medical sociology justifies supposing the fact that education affects health via a couple of intervening social mechanisms that are effective over the whole life course, a fact that substantiates long-term effects, in particular. On the other hand, health status can be important for educational achievements while attending school. In statistics, this problem is referred to as “simultaneous causality,” or “endogeneity” in a narrower sense (Engle, Henry, & Richard, 1983; Proppe, 2009). Endogeneity in the broader sense includes (1) omitting important independent variables, that is, unobserved heterogeneity; (2) biased measurement of variables; (3) serial autocorrelation and lagged dependent variables; (4) self-selection problems; and (5) the aforementioned simultaneous causality (Proppe, 2009). As far as researchers are interested in the direction of causal effects, reverse causality should be ruled out by statistical measures to avoid biased and inconsistent results.

Several methods of causal analysis have been discussed in Legewie (2012). For analyzing causality, randomized experiments—where samples are distributed randomly over treatment and control groups—are preferred. Such an approach is not feasible in the framework of the NEPS for ethical reasons, particularly with regard to health issues. Unobserved heterogeneity caused by time-invariant individual-level variables can be overcome by using fixed effects estimators with panel data. Using this approach, the effects of incidences of serious illnesses on the development of competencies may be analyzed; however, this is not the case for the effect of inherent disabilities or disabilities already in existence since these cases are not taken into account in fixed-effects models. If the causal effect of education on health is analyzed, fixed-effects models are useful as long as the problematic processes take place within the observation window of the study. However, in the case of long-term effects, this is very unlikely (e.g., competencies affect health knowledge that correlates to health-promoting behavior; after several years, inadequate behavior may result in illnesses). Medical knowledge about the physiological and psychological mechanisms and their time frame could be useful here. For example, how long does it take on average until a high workload causes burnout symptoms? This example also shows that thorough theorizing is necessary to include all preconditions (here: high effort and ambitions) and necessary control variables. Education and health are both highly correlated to social background, which captures material and immaterial living conditions and environmental influences in childhood to a certain extent. These confounding variables that can affect both educational success and health should be measured and included as control variables. It is also conceivable that features of the learning environment or of the class context are relevant. Legewie (2012) proposes the use of a school fixed effects model to take care of self-selection in schools in order to assess the effect of class composition on achievement based on the assumption that students are randomly assigned to classrooms within schools.

Simultaneous causality can be considered within an interdependent equation system (von Auer, 2011). Identified equation systems can be estimated using an indirect least square estimation method described in the econometric literature (von Auer, 2011). Underidentified equation systems cannot be estimated, whereas overidentified equation systems can be solved using a two-stage least square estimation (instrument variable estimation).

A common method for dealing with endogeneity is the instrumental variables approach (Angrist & Krueger, 2001). In this approach, the endogenous variable that correlates to the error term in the regression equation is replaced by an instrument that does not correlate to the error term but is ideally highly correlated to the independent variable in question. The estimation takes place in a two-stage least square procedure. In the first stage, the instrument is estimated using additional instrument variables. In the second stage, the instrument replaces the endogenous variable. The instrument variable approach leads to unbiased results if the instrument is not correlated to the error term. The higher the correlation of the instrument and the original variable, the more efficient the estimation is (Proppe, 2009). The crucial problem of this method is finding appropriate instrument variables. Natural experiments can assist in finding instrument variables (Wooldridge, 2002). A natural experiment is given if an exogenously defined mechanism causes variation in an endogenous process. In the NEPS, information about military service or educational grants according to the Federal Education and Training Assistance Act (*Bundesausbildungsförderungsgesetz*/BAFöG) is collected to serve as instrument variables.

5 Discussion

This chapter deals with some of the methodological questions of measuring health within an educational survey. In particular, it analyzes how social desirability in interaction with the mode of data collection might influence data quality. Taken together, the analyses show that in four out of ten comparisons, the results in the NEPS differ from other data in the expected way assuming a tendency towards social desirability. In two cases, no significant differences can be found, and in four cases, the discrepancy from the reference data is contrary to the expected direction. However, in the case of ninth graders, the discrepancy is consistent with higher body heights reported in the same dataset. The same argument can be applied to adults for whom weight is again reported higher in the NEPS data—also resulting in a higher BMI—than in GEDA. If respondents had reported their masses in a biased way, they would not have reported taller heights and higher weights, but rather taller heights and lower weights. Another puzzling result is the high proportion of underweight children in Kindergarten in the NEPS. Either the proportion of children who are underweight grew in the time between both studies, or parents of these children did not take part in a health survey conducted by a state authority, such as the KiGGS. It should also be taken into

account that the data from the parent interviews was collected via CATI, which provides different sources of error than do paper-and-pencil interviews. Future waves of the NEPS will contain additional ranges that check not only for implausible values during the interview, but also for unlikely values.

At present, it is difficult to say within what time span health will show effects on educational success and competencies, or over what period of time educational achievement will impact on health development during the life course. For Germany, the NEPS may yield an increase in knowledge that is comparable with the knowledge that long-term British or American cohort studies, such as the National Child Development Study (NCDS since 1958 in Great Britain) and the Children and Young Adults of the National Longitudinal Survey of Youth (NLSY79, since 1979 in the USA), have acquired.

The NEPS offers the unique opportunity to identify the reciprocal interrelations of education and health in the life course. This seems all the more important as social status is of central importance for both education and health. Depending on political will, institutions of the educational system (day nursery, nursery school, school) could promote health and prevention not only in the framework of programs (such as “Gesunde Schule/Healthy School”, “Klasse 2000/Grade 2000”), but also in the framework of all-day schools by offering a healthy breakfast or lunch and a number of physical activities. This not only could improve individuals’ quality of life and chances of success, but would also be likely to decrease health expenses for society at large.

References

- Angrist, J. D., & Krueger, A. B. (2001). Instrumental variables and the search for identification: From supply and demand to natural experiments. *Journal of Economic Perspectives*, 15(4), 69–85.
- Basch, C. E. (2011). Healthier students are better learners: A missing link in school reforms to close the achievement gap. *Journal of School Health*, 81(10), 593–598.
- Béland, Y., & St-Pierre, M. (2008). Mode effects in the Canadian Community Health Survey: A comparison of CATI and CAPI. In J. M. Lepkowski (Ed.), *Advances in telephone survey methodology* (pp. 297–311). Hoboken, NJ: John Wiley & Sons.
- Blossfeld, H.-P., Roßbach, H.-G., & von Maurice, J. (Eds.) (2011). Education as a life-long process: The German National Educational Panel Study (NEPS) [Special Issue]. *Zeitschrift für Erziehungswissenschaft*, 14. Wiesbaden: VS Verlag für Sozialwissenschaften.
- Cole, T. J., Bellizzi, M. C., Flegal, K. M., & Dietz, W. H. (2000). Establishing a standard definition for child over-weight and obesity worldwide: International survey. *British Medical Journal*, 320(7244), 1–6.

- Cutler, D. M., & Lleras-Muney, A. (2012). *Education, and health: Insights from international comparisons*. (NBER Working Paper No. 17738). Cambridge, MA: National Bureau of Economic Research.
- Dadaczynski, K. (2012). Stand der Forschung zum Zusammenhang von Gesundheit und Bildung. *Zeitschrift für Gesundheitspsychologie*, 20, 141–153.
- Dragano, N., & Siegrist, J. (2009). Die Lebenslaufperspektive gesundheitlicher Ungleichheit. In M. Richter, & K. Hurrelmann (Eds.), *Gesundheitliche Ungleichheit: Grundlagen, Probleme, Perspektiven* (pp. 181–194). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Engle, R. F., Henry, D. F., & Richard, J.-F. (1983). Exogeneity. *Econometrica*, 51(2), 277–304.
- Glaesmer, H., & Brähler, E. (2002). Schätzung der Prävalenz von Übergewicht und Adipositas auf der Grundlage subjektiver Daten zum Body-Mass-Index (BMI). *Das Gesundheitswesen*, 64(3), 133–138.
- Gross, C., Jobst, A., Jungbauer-Gans, M., & Schwarze, J. (2011). Educational returns over the life course. In H.-P. Blossfeld, H.-G. Roßbach, & J. von Maurice (Eds.), *Zeitschrift für Erziehungswissenschaft*, 14. *Education as a lifelong process: The German National Educational Panel Study (NEPS)* (pp. 139–154). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Groves, R. M., Fowler F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2004). *Survey methodology*. New York: Wiley.
- Kimbro, R. T., Bzostek, S., Goldman, N., & Rodríguez, G. (2008). Race, ethnicity, and the education gradient in health. *Health Affairs*, 27, 361–372.
- Kroh, M. (2004). *Intervieweffekte bei der Erhebung des Körpergewichts: Die Qualität von umfragebasierten Gewichtsangaben* (DIW-Diskussionspapier No. 439). Berlin: German Institute for Economic Research (DIW Berlin).
- Kromeyer-Hauschild, K., Wabitsch, M., Geller, F., Ziegler, A., Geiß, H. C., Hesse, V., & Hebebrand, J. (2001). Perzentile für den Body Mass Index für das Kindes- und Jugendalter unter Heranziehung verschiedener deutscher Stichproben. *Monatsschrift Kinderheilkunde*, 149(8), 807–818.
- Kurth, B. M. (2007). Der Kinder- und Jugendgesundheitsurvey (KiGGS): Ein Überblick über Planung, Durchführung und Ergebnisse unter Berücksichtigung von Aspekten eines Qualitätsmanagements. *Bundesgesundheitsblatt—Gesundheitsforschung—Gesundheitsschutz*, 50, 533–546.
- Kurth, B. M., & Schaffrath-Rosario, A. (2007). Die Verbreitung von Übergewicht und Adipositas bei Kindern und Jugendlichen in Deutschland. *Bundesgesundheitsblatt—Gesundheitsforschung—Gesundheitsschutz*, 50, 736–743.
- Legewie, J. (2012). Die Schätzung von kausalen Effekten: Überlegungen zu Methoden der Kausalanalyse anhand von Kontexteffekten in der Schule. *Kölner Zeitschrift für Soziologie und Sozialpsychologie*, 64(1), 123–153.
- Mackenbach, J. P. (2006). *Health inequalities: Europe in profile. An independent expert report commissioned by the UK presidency of the EU*. London: Department of Health.

- Mielck, A. (2008). Soziale Ungleichheit und Gesundheit in Deutschland. Die internationale Perspektive. *Bundesgesundheitsblatt—Gesundheitsforschung—Gesundheitsschutz*, 51, 345–352.
- Morgan, J. F., Reid, F., & Lacey, J. H. (1999). The SCOFF questionnaire: Assessment of a new screening tool for eating disorders. *British Medical Journal*, 319(7223), 1467–1468.
- Siegrist, J., & Marmot, M. (2006). Social inequalities in health: Basic facts. In J. Siegrist, & M. Marmot (Eds.), *Social inequalities in health. New evidence and policy implications* (pp. 1–25). Oxford: Oxford University Press.
- OECD. (2006). What does education do to our health? In OECD (Ed.), *Measuring the effects of education on health and civic engagement* (pp. 355–363). Paris: OECD.
- Perry, L., Morgan, J., Reid, F., Brunton, J., O'Brien, A., Luck, A., & Lacey, H. (2002). Screening for symptoms of eating disorders: Reliability of the SCOFF screening tool with written compared to oral delivery. *International Journal of Eating Disorders*, 32(4), 466–472.
- Power, C., & Kuh, D. (2006). Life course development of unequal health. In J. Siegrist, & M. Marmot (Eds.), *Social inequalities in health. New evidence and policy implications* (pp. 27–54). Oxford: Oxford University Press.
- Proppe, D. (2009). Endogenität und Instrumentenschätzer. In S. Albers, D. Klapper, U. Konradt, A. Walter, & J. Wolf (Eds.), *Methodik der empirischen Forschung* (3rd ed., pp. 253–266). München: Gabler.
- Public Use File GEDA 2010, Robert Koch Institute, Berlin (Germany) 2012.
- Public Use File KiGGS, The German Health Survey for Children and Adolescents 2003–2006, Robert Koch Institute, Berlin (Germany), 2008.
- Ross, C. E., & Wu, C. (1995). The links between education and health. *American Sociological Review*, 60(5), 719–745.
- Shields, M., Grober, S. C., & Tremblay, M. S. (2008). Effects of measurement on obesity and morbidity. *Health Reports*, 19(2), 77–84.
- Suhrcke, M., & de Paz Nieves, C. (2011). *The impact of health and health behaviours on educational outcomes in high-income countries: A review of the evidence*. Copenhagen: WHO.
- Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge: Cambridge University Press.
- Tourangeau, R., & Smith, T. W. (1996). Asking sensitive questions: The impact of data collection mode, question format, and question context. *The Public Opinion Quarterly*, 60(2), 275–304.
- Visscher, T. L. S., Viet, A. L., Kroesbergen, I. H., & Seidell, J. C. (2006). Underreporting of BMI in adults and its effect on obesity prevalence estimations in the period 1998 to 2001. *Obesity*, 14(11), 2054–2063.
- von Auer, L. (2011). *Ökonometrie. Eine Einführung* (5th ed.). Berlin: Springer.
- WHO. (2000). *Obesity: Preventing and managing the global epidemic. Report of a WHO consultation* (Technical Report Series No 894). Geneva: WHO.
- Wooldridge, J. M. (2002). *Econometric analysis of cross section and panel data*. Cambridge: B & T.

Acknowledgement

This paper uses data from the National Educational Panel Study (NEPS):

Starting Cohort 2—Kindergarten, doi:10.5157/NEPS:SC2:1.0.0.

Starting Cohort 4—Grade 9, doi:10.5157/NEPS:SC4:1.0.0.

Starting Cohort 6—Adults, doi:10.5157/NEPS:SC6:5.1.0.

The NEPS data collection is part of the Framework Programme for the Promotion of Empirical Educational Research, funded by the German Federal Ministry of Education and Research and supported by the Federal States.

About the authors

J. Carstensen

DZHW – German Centre for Higher Education Research and Science Studies

e-mail: carstensen@dzhw.eu

A. Gottburgsen

DZHW – German Centre for Higher Education Research and Science Studies

M. Jungbauer-Gans

DZHW – German Centre for Higher Education Research and Science Studies

The Standard Stress Scale (SSS): Measuring Stress in the Life Course

Christiane Gross and Katharina Seebaß

Abstract

This contribution presents the Standard Stress Scale (SSS), a new scale that has been specially developed to meet the requirements of multicohort panel studies—such as the National Educational Panel Study (NEPS)—that refer to the whole life course. Accordingly, the SSS is consistently applicable for different age groups from 14 years old onwards and is also suitable for a wide range of people, irrespective of their stage in life and employment situation. The items are applicable to (university) students; employed, unemployed, and self-employed people; housewives and -husbands; old-age pensioners; and so forth. To obtain the final 11-item Standard Stress Scale (SSS), 35 questions regarding stressful life situations, social stress, daily distress, anxiety about the future, and other stresses and strains were developed following the theoretical approach of the effort-reward imbalance model (ERI) and the demand-control model. These 35 items were pretested with different subsamples—such as students in different school types, university students, and adults—using self-administered questionnaires. The total sample of the pretest includes 372 respondents. All of the 35 original questions had a small item-nonresponse rate and a good variance among respondents. Using factor analyses, the questions with the highest factor loading in each of the dimensions were used to represent the final 11-item SSS. In some cases, when the questions with the highest loading did not perform well in the cognitive pretest, the item with the second-highest loading was chosen instead. Although the most distinct items were selected, the final 11 items of the SSS show good reliability values. The Cronbach's Alpha values vary in a range in all subsamples from 0.58 for the unemployed to 0.66 for students. In addition, further analyses show a high correlation of the final SSS with self-rated health. The use of the SSS is free of charge but has to be cited using this publication.

1 Introduction

Stress is one of the main determinants of health status (Backé et al., 2012; Steptoe, 1991). Therefore, an instrument to adequately measure stress is of prime interest not only in public health research, but also for the examination of educational returns. School and workplace requirements are both essential sources of stress, and stress levels can also be affected by unemployment.

Providing excellent data on nonmonetary returns to education—such as health—is one focus of Pillar 5 (Returns to Education Over the Life Course) of the National Educational Panel Study (NEPS). The NEPS aims to use a constant scale that meets the standards of survey methodology to measure stress for different age groups and living conditions. As none of the existing scales meets these requirements, we have developed the Standard Stress Scale (SSS), which is applied by the NEPS but can also be used in further surveys. The SSS is applicable for different age groups (14 years old and above) and is also suitable for all sorts of people, irrespective of their stage in life and employment situation. The SSS was used in the NEPS Starting Cohort 3—Grade 5 (in Wave 4), Starting Cohort 4—Grade 9 (in Waves 5 & 6), Starting Cohort 5—First-Year Students (in Wave 6), and Starting Cohort 6—Adults (in Wave 4).

We first present previous stress scales to underline the need for the development of a new instrument to measure stress in the life course (Section 2). Then we outline the theoretical dimensions of stress on which the Standard Stress Scale is based (Section 3) and introduce the methods used to develop the scale (Section 4) as well as the results of the cognitive pretest and factor analyses along with an explanation of how to build a stress index (Section 5). Finally, we show some attributes of the resulting stress index based on the SSS (Section 6).

2 Previous Stress Scales

A variety of previous instruments to measure stress, available in a German version, are summarized in Table 1.

- a) Possibly the most popular instrument is the “Effort-Reward Imbalance Scale (ERI)” (Siegrist, 1996; Siegrist et al., 2004), which is based on the theoretical concept of reciprocity. The model of effort-reward assumes that negative emotions occur when the effort made by a person is much higher than the reward the person receives, meaning that the main principle of reciprocity has been violated. Although the original scale was established to measure stress in the workplace only, Siegrist and colleagues developed further scales that focus on school (Li et al., 2010), housework among women (Sperlich et al., 2012, Sperlich et al., 2013), social relationships, and reciprocity (Chandola et al., 2007). The strength of the ERI-Scale—being well adapted for specific life circumstances, such as being an em-

- ployee or a student—is a vital handicap to its application in multicohort panel studies as well as to general cross-sectional surveys. There is no scale dedicated to the unemployed, self-employed, pensioners, or housewives and -husbands. Apart from this, no version of an ERI-scale is applicable from school age through to old age.
- b) The second stress scale—“Skala sozialer Stressoren am Arbeitsplatz” (Frese & Zapf, 1987)—is also limited to measuring stress in the workplace and, in particular, problems within teams in the workplace.
 - c) “The Social Readjustment Rating Scale (SRRS)” provided by Holmes and Rahe (1967) focuses on the number and impact of life-change events and is not limited to employees. However, the SRRS is a product of its time containing items such as “wife begins or stops work,” which are addressed at heterosexual men only.
 - d) The “Stress-Reaktivitäts-Skala (SRS)” by Schulz, Jansen, and Schlotz (2005) is applicable for adult populations only and is mainly used in clinical research to evaluate coping strategies used for stressful situations.

The last two scales are for universal use:

- e) The “Trierer Inventar zur Erfassung von chronischem Stress (TICS)” by Schulz, Schlotz, and Becker (2004) covers six dimensions of chronic stress: excess work, dissatisfaction with work, social strains, lack of social approval, anxiety, and incriminatory memories. Although the issue of “work” is very present here, the items could also be used for other subgroups when interpreting “work” in a wider sense. Nevertheless, the scale does not meet the standards of survey methodology since it has items with two dimensions.
- f) The “Perceived Stress Questionnaire (PSQ)” by Levenstein et al. (1993) is also available in a German version (Fliege et al., 2001; Fliege et al., 2005). It focuses on stress as a result of perceived strains. The German version has been validated with a sample of women after having given birth or having had a miscarriage and a sample of students of medicine (Fliege et al., 2001), as well as in a general household survey (Kocalevent et al., 2011). To date, there is no validation or cognitive pretest for school-aged children. In addition, 30 items is a large number for a survey.

As previous scales do not meet the requirements of the NEPS (a constant scale for many cohorts and all life situations of adults with a small number of items covering many dimensions of stress and having no methodological flaws), we developed the SSS for use in NEPS- and other general surveys. The use of the SSS is free of charge, but using the scale without citing this publication is strictly forbidden. A reference to this article is obligatory for any further use of single items out of the SSS or the whole SSS instrument.

Table 1 Previous Stress Scales

Scale	Theoretical focus/dimensions of stress	Target population	# Items (short version)	References	Comments
(a) Effort-reward imbalance scale (ERI)	Imbalance of effort and reward	Employees, school students, university students	23 (scale for employees, including overcommitment) (10)	Siegrist (1996), Siegrist et al. (2004), Siegrist et al. (2008)	Suitable for surveys with special groups, no constant scale for all subgroups, lack of instruments for special subgroups (such as the unemployed)
(b) Skala sozialer Stressoren am Arbeitsplatz	Social stress in the workplace	Employees	17 (8)	Fries and Zapf (1987)	Suitable for measuring employees' stress levels, especially social problems in teams in the workplace
(c) The Social Readjustment Rating Scale (SRRS)	Number and severity of life-change events	Adults	43	Holmes and Rahe (1967)	For (male) adults only, obsolete items addressed at heterosexual men (such as "wife begins or stops work")
(d) Stress-Reaktivitätsskala (SRS)	Stress reactivity, individual coping strategies, "the extent to which a person is likely to show emotional or physical reactions to a stressful event" (Bolger & Zuckerman 1995: 890)	Adults	29	Schulz et al. (2005)	For adults only, focus on clinical research
(e) Trierer Inventar zur Erfassung von chronischem Stress (TICS)	Six dimensions of chronic stress: excess of work, dissatisfaction with work, social strains, lack of social approval, anxiety, and inflammatory memories	Universal	57 (12)	Schulz et al. (2004)	Items with two dimensions do not meet the standards of survey methodology
(f) Perceived Stress Questionnaire (PSQ)	Stress as a representation of perceived strains	Universal	30	Levenstein et al. (1993); for German adaptation: Fliege et al. (2001, 2005)	No validation for school-aged children, no short version available

3 Dimensions of Stress in the Standard Stress Scale

Our theoretical concept of chronic stress is essentially based on the two most popular models in stress research: the demand-control model (Karasek & Theorell, 1990) and the effort-reward imbalance model (Siegrist, 1996; Siegrist et al., 2004). Both models were developed to measure stress in the workplace. The demand-control model assumes “a high risk of psychological strain and physical illness” when “psychological demands” are high and “decision latitude (control)” is low (Karasek & Theorell, 1990: 32). The inverse situation with low demands and high control would lead to high learning motivation (Karasek & Theorell, 1990). The ERI-model is based on the concept of reciprocity and postulates negative affections when efforts being made are high and rewards in terms of low income, low social approval, and so forth, are low (Siegrist, 1996; Siegrist et al., 2004).

We adopted the theoretical dimensions of these models, such as overcommitment and social approval of the ERI-model as well as the control component of the demand-control model. Based on these dimensions, we developed new items that are suitable for all subgroups, independent of their employment status and school attendance. The subdimensions of stress and the corresponding items of the original 35-item battery (see Figure 1) are presented in Table 2.

Several items were developed for each dimension, resulting in a 35-item scale. Each item was answered using a 5-point Likert scale ranging from 1 “not at all” to 5 “completely.”

To validate these items and generate a short version, the following methods were used.

Table 2 Dimensions of Stress

Subdimensions	Items (see Table 3)
Overcommitment, workload	1, 3, 5, 7, 10
Enjoyment of work, self-realization, empowerment	2, 4, 6, 8, 9, 11
Social distress, social support, social approval	12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 34
Recreational capacities, exhaustion	23, 24, 26, 27, 28, 29, 30, 32
Anxiety about the future, uncertainty	25, 31, 33, 35

Figure 1 Original 35-Item Stress Questionnaire

On an average day, to what extent would you agree with the following statements? Wenn Sie an einen normalen Tag denken, inwiefern treffen folgende Aussagen auf Sie zu?	
[On a 5-point scale of: not at all, very little, neutral, somewhat, to a great extent.] (5-stufige Antwortskala: trifft überhaupt nicht zu, trifft eher nicht zu, teils-teils, trifft eher zu, trifft voll und ganz zu)	
1	I have more tasks to do than I can handle. <i>Ich habe mehr Aufgaben zu bewältigen als ich leisten kann.</i>
2	Generally, I am very satisfied with the results of my actions. <i>In der Regel bin ich mit den Ergebnissen meiner Tätigkeiten sehr zufrieden</i>
3	I often feel like I am in a rat race. <i>Ich fühle mich oft wie ein Hamster in Rad.</i>
4	If I do not enjoy doing something, I usually do not have to do it. <i>Wenn mir eine Tätigkeit keinen Spaß macht, muss ich sie in der Regel auch nicht tun.</i>
5	If I do not take charge, no one else will. <i>Wenn ich mich nicht um die Dinge kümmere macht es niemand.</i>
6	I am in control of many aspects of my life. <i>Ich kann viele Dinge in meinem Leben selbst bestimmen.</i>
7	I usually get left with whatever still needs to be done. <i>Meistens bleibt die Arbeit dann doch wieder an mir hängen.</i>
8	I am often completely frustrated. <i>Ich bin oft völlig frustriert.</i>
9	I enjoy the tasks and duties of an ordinary day. <i>Die Aufgaben an einem gewöhnlichen Tag bereiten mir Freude.</i>
10	I could use more time for my daily duties than I have. <i>Ich bräuchte mehr Zeit für die täglichen Tätigkeiten als ich habe.</i>
11	What I do is meaningful. <i>Ich übe sinnvolle Tätigkeiten aus.</i>
12	My friends expect more of me than I can give them. <i>Meine Freunde erwarten mehr von mir als ich ihnen geben kann.</i>
13	My family brings me more joy than hassle. <i>Meine Familie bereitet mir viel mehr Freude als Ärger.</i>
14	I have great friends. <i>Ich habe tolle Freunde.</i>
15	I am often treated unfairly. <i>Ich werde oft unfair behandelt.</i>
16	I do not meet the expectations of my family. <i>Ich erfülle die Erwartungen meiner Familie nicht.</i>
17	I often deal with people who stress me out. <i>Ich habe viel mit Menschen zu tun, die mich stressen.</i>
18	I often feel lonely. <i>Ich fühle mich oft einsam.</i>
19	Most people admire how I manage my life. <i>Die meisten bewundern mich dafür, wie ich mein Leben meistere.</i>
20	My performance is properly appreciated. <i>Meine Leistungen werden angemessen gewürdigt.</i>
21	No matter what happens, I won't be left alone with problems. <i>Egal was passiert, ich werde mit Problemen nicht allein gelassen.</i>
22	There are people that I can count on. <i>Es gibt Menschen, auf die ich mich verlassen kann.</i>
23	I usually have restful sleep. <i>In der Regel habe ich einen erholsamen Schlaf.</i>
24	I often brood over problems. <i>Ich grübele oft.</i>
25	Presumably, my life situation will worsen. <i>Vermutlich wird sich meine Lebenssituation verschlechtern.</i>
26	Generally, I solve problems well. <i>Im Allgemeinen kann ich Probleme gut lösen.</i>
27	It is easy for me to relax. <i>Ich kann gut abschalten.</i>
28	After a normal day, I feel happy. <i>Nach einem normalen Tag fühle ich mich glücklich.</i>
29	I spend a lot of time thinking about problems. <i>Ich denke viel über Probleme nach.</i>
30	After a normal day, I feel exhausted. <i>Nach einem normalen Tag fühle ich mich erschöpft.</i>
31	I worry a lot about my future. <i>Ich mache mir viel Sorgen um meine Zukunft.</i>
32	After two days off, I feel fully refreshed. <i>Nach zwei freien Tagen, fühle ich mich völlig erholt.</i>
33	I am afraid of what my life will be like in three years. <i>Ich habe Angst davor, wie mein Leben in drei Jahren aussehen könnte.</i>
34	I worry a lot about the people around me. <i>Ich mache mir viel Sorgen um meine Mitmenschen.</i>
35	I look forward to the future. <i>Ich freue mich auf die Zukunft.</i>

4 Methods

To reach the goal of creating a short scale to measure diverse dimensions of stress, it was critical to select the right items—comprehensive for all subgroups of respondents—from the original 35-item stress battery. Therefore, we conducted *cognitive pretests* to guarantee comprehensibility and *factor analyses* to separate dimensions of stress and to choose the most discriminating items for the short version of the SSS. Before referring to these two methods, we first describe the pretest subsamples. Pretests were conducted via paper-and-pencil-interviewing (PAPI) using the 35-item battery of the SSS (see Figure 1) in the following locations:

- 1) Respondents were interviewed while visiting the *registration office* of Nuremberg and waiting for their turn. Because usability among all age groups and employment statuses was especially important for the pretest, the city hall seemed to be a good setting. These interviews were conducted on different days of the week in June and August 2011.
- 2) A *university* sample of bachelor students (second semester) at the Department of Social Economics at the University of Erlangen-Nuremberg was interviewed in a class setting in the summer term of 2011.
- 3) The *school* sample contains five classes attending a “Gymnasium” [a type of school leading to upper secondary education and the Abitur] in the City of Kiel ($n = 110$) and two classes of a “Berufsfachschule” [full-time vocational school] in the city of Ludwigshafen ($n = 31$) that cover a wide range of levels of competencies. The Gymnasium sample consists of two classes in Grade 9 and one class each in Grades 10, 11, and 12. The subsample in Berufsfachschule covers two first-year classes. The students of these two classes are strive for vocational degrees as lacquerers and painters and often have a low level of competencies within Berufsfachschulen. The students of the whole school sample were 14 years old and above at the time of interviewing and were also interviewed in a class setting.

Table 3 provides an overview of the different subsamples that were realized by location.

Table 3 Subsamples of the Pretests

Location of pretest	<i>N</i>	%
Registration office	159	42.7
University	72	19.4
School	141	37.9
Total	372	100.00

Table 4 Status of Participants

Status	N	%	Aggregated status	N	%
Full-time employed	71	19.1	Employed	90	24.2
Part-time employed	19	5.1			
University student	86	23.1	University student	86	23.1
School student	150	40.3	School student	150	40.3
Housewife, -husband	12	3.2	Other	46	12.4
Retired	18	4.8			
Unemployed	10	2.7			
Other	6	1.6			
Total	372	100.0	Total	372	100.0

Although the settings of the subsamples were rather specific, a wide range of people in different educational and occupational statuses was able to be realized (see Table 4). For further analyses, the status is aggregated into four groups (see Columns 3 to 6 in Table 4).

Because of the focus on students in schools and universities, the age distribution among the respondents tends towards the younger age groups (see Table 5); nevertheless, the number of older people in the pretest should still be sufficient for the analyses.

Factor analyses are generally used to uncover structural dimensions within the data and extract factors for further use when generating an index (Backhaus et al., 2003). In addition, factor analyses can be conducted to reduce complex data structure by identifying important items within the data (Costello & Osborne, 2005; Wolff

Table 5 Age of Respondents

Age group	N	%
Under 18	113	30.8
18–25 years	140	38.2
25–45 years	62	16.9
45–65 years	40	10.9
Over 65	12	3.7
Total	367	100.0

& Bacher, 2010). For each extracted factor, the included items load differently on the factor. In our analyses, the item with the highest factor loading was considered the best representative item for this factor. The item with the second-highest factor loading was used instead when the item with highest loading did not perform well in the cognitive pretest (for further details, see Section 5.2). Therefore, the final instrument is based on those items retrieved from factor analyses that best represent the factors. The factor analyses were performed using the principal component method with varimax rotation. They were carried out by using both the whole sample as well as subsamples by employment status of the participants (see aggregated status in Table 4).

The *cognitive pretests* were mainly targeted at the comprehensibility of the items' wording. Questions during the interviews in school- and university classes were noted and analyzed. The respondents in the registration-office sample were able to address their questions directly to the interviewer, who was instructed to note their questions. All questionnaires contained an open question at the end that asked for feedback on the questionnaire and on problems of comprehensibility. The next section shows the result of the respondents' questions and remarks as well as the results of the factor analyses.

5 Results

The selection of the final items was dependent on several criteria: no (or very few) missing values, high variance in the answers (meaning that Categories 1 to 5 were chosen as far as possible) (see Section 5.1), no cognitive problems with the wording of the item (see Section 5.2), and finally, high factor loadings on the item (see Section 5.3).

5.1 Descriptive Results

The descriptive results of the analysis show good variance of every item. Each category was answered at least twice. The number of missing values is reasonable. Only two items have an item-nonresponse rate greater than 2% (Item v03 and Item v27) (compare Table 6).

Table 6 Descriptive Statistics of 35-Item Stress Questionnaire

Variable	<i>N</i>	<i>M</i>	<i>SD</i>	Min	Max
v01	370	2.81	1.03	1	5
v02	369	3.54	0.89	1	5
v03	352	2.47	1.16	1	5
v04	369	2.43	1.13	1	5
v05	371	3.01	1.10	1	5
v06	370	3.86	0.91	1	5
v07	369	2.98	1.03	1	5
v08	369	2.18	1.09	1	5
v09	370	3.22	0.90	1	5
v10	368	3.23	1.18	1	5
v11	367	3.62	0.97	1	5
v12	370	2.15	1.06	1	5
v13	367	3.85	1.15	1	5
v14	369	4.33	0.94	1	5
v15	364	2.22	0.92	1	5
v16	368	2.03	1.12	1	5
v17	372	2.66	1.12	1	5
v18	369	2.03	1.11	1	5
v19	366	3.17	1.02	1	5
v20	368	3.29	0.91	1	5
v21	370	3.65	1.05	1	5
v22	370	4.44	0.87	1	5
v23	371	3.29	1.13	1	5
v24	365	3.31	1.07	1	5
v25	368	2.11	0.98	1	5
v26	370	3.84	0.78	1	5
v27	362	3.26	1.13	1	5
v28	367	3.52	0.93	1	5
v29	370	3.55	1.06	1	5
v30	371	3.06	1.03	1	5
v31	370	3.11	1.15	1	5
v32	369	3.29	1.17	1	5
v33	369	2.45	1.20	1	5
v34	370	3.12	1.00	1	5
v35	369	3.73	0.97	1	5

5.2 Cognitive Pretests

The cognitive pretest among both school and university students showed that six questions were not comprehensible for some respondents and were therefore not considered for the final version of the SSS. The cognitive pretests revealed comprehension problems with single words or the wording of some items:

- With Item v03, students did not understand the meaning of the saying (“I often feel like I am in a rat race.”) well. This item refers to someone who keeps on running without moving on and without being able to stop.
- The negative connotation of Item v08 (“I am often completely frustrated.”) was criticized by students.
- Regarding Item v13 (“My family brings me more joy than hassle.”), some respondents remarked that they do not have a family and therefore could not answer the question.
- Item v19 (“Most people admire how I manage my life.”) was criticized by school students, in particular. They argued that “sometimes you do not know what other people think of you” and that it is therefore impossible to answer the question accurately.
- The shortcoming of Item v24 (“I often brood over problems.”) was respondents’ lack of knowledge of the German term for “to brood” (“grübeln”).
- Item v27 (“It is easy for me to relax.”) confused students with the ambivalent meaning of “abschalten” (which can mean both “relax” and “switch off”). Students mostly thought of switching off technical equipment, such as computers, smartphones, or televisions.

As a result of these comprehension problems, Items v03, v08, v13, v19, v24, and v27 were not considered for the final instrument, regardless of what their performance in the factor analyses was like. In addition, the shortcomings of Items v03 and v27 showed up in the descriptive analysis through a high item-nonresponse.

5.3 Factor Analyses

Factor analyses were carried out with subsamples of employed people, university students, school students, and others (see Table 7). Using this design allows us to meet the needs of a multi-cohort study in which scales have to function for all subgroups. Every subsample led to slightly different results concerning the number of factors extracted. This is mainly due to the fact that factor analysis is sensitive to sample size in general and also that it is an exploratory method (see Costello & Osborne, 2005).

In Table 7, the items with the highest loadings on the factors are presented. Depending on the subsample, 9, 10, or 11 factors were retrieved. We decided on an

Table 7 Results of Factor Analyses on the Subsamples

Status	Number of observations	Number of factors	Items with highest factor loading
Employed	90	11	v04, v05, v11, v18, v20, v22, v27, v29, v30, v31, v35
University student	86	11	v04, v05, v11, v18, v20, v22, v27, v29, v30, v31, v35
Student	150	10	v04, v07, v10, v12, v14, v16, v19, v24, v31, v35
Other	46	11	v04, v05, v11, v18, v20, v22, v27, v29, v30, v31, v35
Total	372	9	v02, v04, v07, v10, v16, v22, v24, v33, v35
Final Scale			v04, v05, v11, v18, v20, v22, v23, v29, v30, v33, v35

11-factor solution and chose the items with the highest or second-highest loadings. Three out of our four subsamples in Table 7 work with 11 factors; therefore, we decided to use 11 items to represent the factors in the final SSS.

In the cognitive pretesting, Item v27 led to misunderstanding and shows a rather high number of missing values (2.8 %); therefore, in the final scale, Item v27 was replaced by v23, which had the second-highest loading in most of the factor analyses. Because of the similar wording of Items v31 and v35, Item v31 was replaced by v33, which always had the second-highest loading on the specific factor.¹ All groups seem to lead to similar results concerning the items with the highest factor loading. Only students seem to show a slightly different pattern; however, when also considering the items that have the second-highest loading in the factor analysis of students² (v02, v05, v06, v13, v22, v28, v29, v33), the results match better with those of the other subgroups.

6 Characteristics of the Instrument

The final instrument, which includes the diverse dimensions of stress, consists of 11 items of the initial 35-item stress battery. These items meet the preconditions of selection (low missing values, high variance, no cognitive problems, and good representation of a stress dimension). The final instrument is a short battery of 11 questions concerning the general life situation of the respondents that can be combined in a stress index. The chosen questions cover all subdimensions of stress (cf. Table 2), with one item representing over-commitment/workload (Item 5), two questions re-

1 A detailed methodological report on the results of the factor analyses can be obtained from the authors (Gross & Seebaß, 2012).

2 In order to keep information content high, only factor loadings > 0.5 are considered.

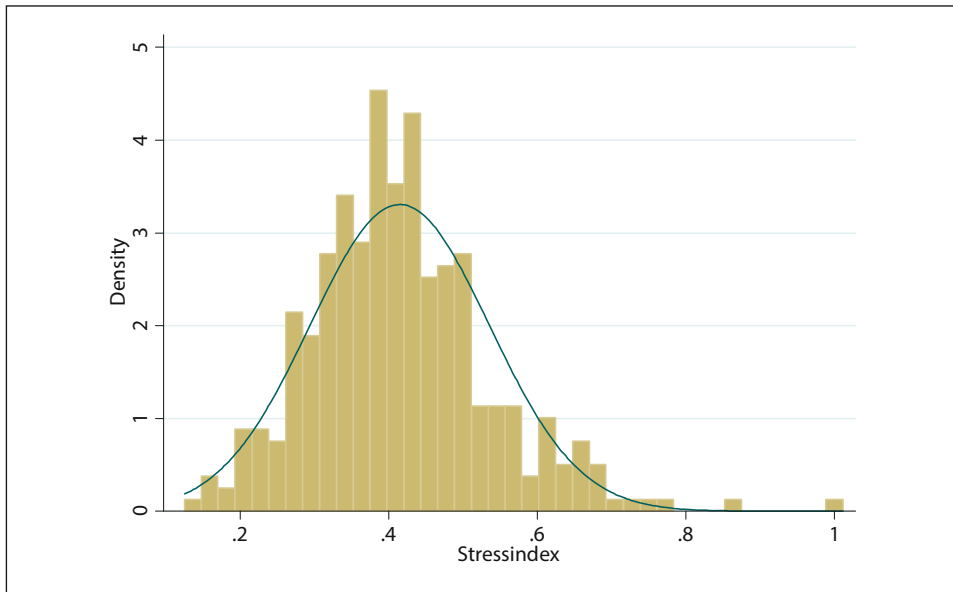
Figure 2 The 11 Items for Measuring the Standard Stress Scale (SSS)

<p>Nun interessieren wir uns dafür, wie es Ihnen ganz allgemein geht. Denken Sie dabei bitte an alle Lebensbereiche. Inwiefern treffen folgende Aussagen auf Sie zu?</p> <p>[We are now interested in how you are in general. Please think of all areas of life. To what extent do the following statements apply to you?]</p>	
<p>(5-stufige Antwortskala: trifft gar nicht zu; trifft eher nicht zu; teils, teils; trifft eher zu; trifft völlig zu) [5-point scale: not at all, very little, neutral, somewhat, to a great extent]</p>	
1	Wenn mir eine Tätigkeit keinen Spaß macht, muss ich sie in der Regel auch nicht tun. (question 4) [If I do not enjoy doing something, I usually do not have to do it.]
2	Wenn ich mich nicht selbst um etwas kümmere, tut es keiner. (question 5) [If I do not take charge, no one else will.]
3	Ich übe sinnvolle Tätigkeiten aus. (question 11) [What I do is meaningful.]
4	Ich fühle mich oft einsam. (question 18) [I often feel lonely.]
5	Meine Leistungen werden angemessen gewürdigt. (question 20) [My performance is properly appreciated.]
6	Es gibt Menschen, auf die ich mich verlassen kann. (question 22) [There are people that I can count on.]
7	In der Regel habe ich einen erholsamen Schlaf. (question 23) [I usually have restful sleep.]
8	Ich denke viel über Probleme nach. (question 29) [I spend a lot of time thinking about problems.]
9	Nach einem normalen Tag fühle ich mich erschöpft. (question 30) [After a normal day, I am feel exhausted.]
10	Ich habe Angst davor, wie mein Leben in drei Jahren aussehen könnte. (question 33) [I am afraid of what my life will be like in three years.]
11	Ich freue mich auf die Zukunft. (question 35) [I look forward to the future.]

lated to enjoyment of work/self-realization/empowerment (Items 4 and 11), three items considering social distress/social support/social approval (Items 18, 20, and 22), three items representing recreational capacities/exhaustion (Items 23, 29, and 30), and two questions covering anxiety about the future/uncertainty (Items 33 and 35).

To build the 0-1-standardized SSS index, the following procedure is used:

- Recode Items 1, 3, 5, 6, 7, and 11 so that a high value indicates a stressful issue.
- Generate a new variable by adding the 11 answer values, subtracting 11 (minimum), and dividing by 44 (maximum after subtraction). Alternatively, you can use the routines implemented in your statistics software.
- When missing values occur, adjust the procedure (for 2 missing values, subtract 9 and divide by 36, etc.). For a high number of missing values, balance the pros and cons for your purpose of having a missing value for the whole index or an index that does not represent all stress dimensions.

Figure 3 Distribution of the Stress Index

The SSS index should have a possible range from 0 to 1, with 1 indicating a maximum of stress and 0 a minimum of stress. With the data from our pretest sample, the SSS index shows a good fit to a normal distribution (see Figure 3), which is a great advantage when using parametric methods of data analysis.

6.1 Reliability

Although the most distinct items were selected, the final 11 items of the SSS show good reliability values. Within the subpopulations, Cronbach's alpha ranges from between 0.58 for the "others" category and 0.66 for school students. The alpha for the total sample with 0.62 is still satisfactory (see Table 8).

6.2 Criterion Validity

Stress scales are usually validated by showing a strong association between the stress index and self-rated health (Li et al., 2010; Niedhammer et al., 2004; Siegrist et al., 2008). The explanatory power of the SSS has also been examined for subjective health status. Within the pretest, students were asked to rate their personal subjective health status (ranging from very good to very bad on a 5-point Likert scale). The stress index

Table 8 Reliability in the Subsamples

Status	<i>N</i>	Cronbach's α
Employed	90	0.65
School student	150	0.66
University student	86	0.60
Other ^a	46	0.58
Total	372	0.62

^a Includes: unemployed, retired, housewife/-husband, maternity leave, etc.

Table 9 Logistic Regression on Subjective Health among Subsample of School Students

	Marginal Effects (z-value)
Gender (1 = female)	-0.06 (-0.88)
Age (in years)	0.00 (0.23)
Stress scale	-1.33 (-4.16)***
<i>N</i>	124
Pseudo R^2	0.26

is highly significant in explaining health. The higher the measured stress, the lower the likelihood of having a subjective (very) good health status is (see Table 9). This result supports the high usability of the SSS.

7 Conclusion

The SSS index has very positive attributes for further use in multivariate analyses: It is almost normally distributed, has a good reliability in spite of covering all main stress dimensions, and has a high association with self-rated health. Moreover, the SSS has been pretested among different populations, from adolescents through to retirees. Therefore, the SSS is highly suitable for applications in general-population surveys as well as in panel studies among heterogeneous subgroups.

However, the items are not adapted to specific life contexts, such as working conditions, school environment, and so forth, so other scales, such as the effort-reward imbalance scale, are likely to be more appropriate for special-issue surveys without the requirements of a constant instrument for all life situations.

References

- Backé, E.-M., Seidler, A., Latza, U., Rossnagel, K., & Schumann, B. (2012). The role of psychosocial stress at work for the development of cardiovascular diseases: A systematic review. *International Archives of Occupational and Environmental Health*, *85*, 67–79. doi:10.1007/s00420-011-0643-6
- Backhaus, K., Erichson, B., Plinke, W., & Weiber, R. (2003). *Multivariate Analysemethoden. Eine anwendungsorientierte Einführung*. Heidelberg: Springer.
- Chandola, T, Marmot, M., & J. Siegrist (2007). Failed reciprocity in close social relationships and health: Findings from the Whitehall II study. *Journal of Psychosom Research*, *63*, 403–411. doi: 10.1016/j.jpsychores.2007.07.012
- Costello, A. B., & Osborne, J. W. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research & Evaluation*, *10*(7), 1–9.
- Gross, C., & Seebaß, K. (2012). *Eine neue Skala zur Messung von Stress im Lebensverlauf. Qualitätsbericht zu Erhebungsdesign und Methodik für NEPS*. Unpublished manuscript.
- Fliege, H., Rose, M., Arck, P., Levenstein, S., & Klapp, B. F. (2001). Validierung des “Perceived Stress Questionnaire” (PSQ) an einer deutschen Stichprobe. *Diagnostica*, *47*, 142–152. doi:10.1026//0012-1924.47.3.142
- Fliege, H., Rose, M., Arck, P., Walter, O. B., Kocalevent, R. D., Weber, C., & Klapp, B. F. (2005). The Perceived Stress Questionnaire (PSQ) reconsidered: Validation and reference values from different clinical and healthy adult samples. *Psychosomatic Medicine*, *67*, 78–88. doi:10.1097/01.psy.0000151491.80178.78
- Frese, M., & Zapf, D. (1987). Eine Skala zur Erfassung von sozialen Stressoren am Arbeitsplatz. *Zeitschrift für Arbeitswissenschaften*, *41*(3), 134–141.
- Holmes, T. H., & Rahe, R. H. (1967). The social readjustment rating scale. *Journal of Psychosomatic Research*, *11*, 213–218. doi:10.1016/0022-3999(67)90010-4
- Karasek, R., & Theorell, T. (1990). *Healthy work, stress, productivity, and the reconstruction of working life*. New York: Basic Books.
- Kocalevent, R.-D., Hinz, A., Brähler, E., & Klapp, B. F. (2011). Regionale und individuelle Faktoren von Stresserleben in Deutschland: Ergebnisse einer repräsentativen Befragung mit dem Perceived Stress Questionnaire (PSQ). *Gesundheitswesen*, *73*, 829–834. doi:10.1055/s-0030-1268445
- Levenstein, S., Prantera C., Varvo V., Scribano, M. L., Berto, E., Luzi, C., & Andreoli, A. (1993). Development of the Perceived Stress Questionnaire: A new tool for psychosomatic research. *Journal of Psychosomatic Research*, *37*, 19–32. doi:022-3999/93
- Li, J., Shang, L., Wang, T., & Siegrist, J. (2010). Measuring effort—reward imbalance in school settings: A novel approach and its association with self-rated health. *Journal of Epidemiology*, *20*, 111–118. doi:10.2188/jea.JE20090057
- Niedhammer, I., Tek, M.-L., Starke, D., & Siegrist, S. (2004). Effort—reward imbalance model and self-reported health: Cross-sectional and prospective findings from

- the GAZEL cohort. *Social Science & Medicine*, 58, 1531–1541. doi:10.1016/S0277-9536(03)00346-0
- Schulz, J., Jansen, L. J., & Schlotz, W. (2005). Stressreaktivität: Theoretisches Konzept und Messung. *Diagnostica*, 51, 124–133. doi:10.1026/0012-1924.51.3.124
- Schulz, P., Schlotz, W., & Becker, P. (2004). *TICS Trier Inventar zum chronischen Stress*. Göttingen: Hogrefe.
- Siegrist, J. (1996). *Soziale Krisen und Gesundheit. Eine Theorie der Gesundheitsförderung am Beispiel von Herz-Kreislauf-Risiken im Erwerbsleben*. Göttingen: Hogrefe.
- Siegrist, J., Starke, D., Chandola, T., Godin, I., Marmot, M., Niedhammer, I., & Peter, R. (2004). The measurement of effort—reward imbalance at work: European comparisons. *Social Science & Medicine*, 58, 1483–1499. doi:10.1016/S0277-9536(03)00351-4
- Siegrist, J., Wege, N., Pühlhofer, F., & Wahrendorf, M. (2008). A short generic measure of work stress in the era of globalization: Effort—reward imbalance. *International Archives of Occupational and Environmental Health*, 82, 1005–1013. doi:10.1007/s00420-008-0384-3
- Sperlich S., Peter, R. & Geyer, S. (2012). Applying the effort-reward imbalance model to household and family work: A population-based study of German mothers. *BMC Public Health*, 12, 1–12. doi: 10.1186/1471-2458-12-12
- Sperlich, S., Arnhold-Kerri, S., Siegrist, J., & Geyer, S. (2013). The mismatch between high effort and low reward in household and family work predicts impaired health among mothers. *European Journal of Public Health*, 23, 893–898. doi: 10.1093/eurpub/cks134
- Step toe, A. (1991). The links between stress and illness. *Journal of Psychosomatic Research*, 35, 633–644. doi:022-3999/91
- Wolff, H.-G., & Bacher, J. (2010). Hauptkomponentenanalyse und explorative Faktorenanalyse. In C. Wolf, & H. Best. (Eds.), *Handbuch der sozialwissenschaftlichen Datenanalyse* (S. 333–366). Wiesbaden: Springer.

About the authors

C. Gross
University of Hanover, Hanover
c.gross@ish.uni-hannover.de

K. Seebaß
University of Erlangen-Nuremberg, Nuremberg.
e-mail: katharina.seebass@fau.de

Validity of Survey Data of Students with Special Educational Needs—Results From the National Educational Panel Study

Lena Nusser, Jana Heydrich, Claus H. Carstensen,
Cordula Artelt and Sabine Weinert

Abstract

Within the German National Educational Panel Study (NEPS), $N = 578$ students in Grade 5 and $N = 1,186$ students in Grade 9 with special educational needs in the area of learning (SEN-L) took part in feasibility studies examining how to include students with special needs in large-scale assessments like the NEPS (Heydrich, Weinert, Nusser, Artelt, & Carstensen, 2013). Among other things, written questionnaires were administered to the participating students. Alongside gaining insight into students' perspectives on educationally relevant questions, the information given by the students is also important in case of nonparticipating parents and thus of missing information on family backgrounds from the parents. Former research could show that secondary-school students without SEN living at home with their parents are reliable proxy reporters for their parents' socioeconomic status and familial background. However, there is no database showing that this conclusion can be generalized to students with SEN-L. Thus, we asked whether the administered student questionnaire validly assessed the social background of these students. In addition to a thorough descriptive analysis of missing data as an indicator of the response behavior of students with SEN-L, the validity of students' answers was also tested by matching the parents' data with the students' responses in order to identify accuracy using a chance-corrected agreement coefficient. Students with SEN-L responded validly and accurately to certain questions, while other items resulted in low completion rates and reduced validity of the students' reports.

1 Introduction

When exploring individual educational pathways, as is done in educational panel studies, it is essential to gain a detailed view of the target person and their respective educational contexts. This requires a variety of reliably and validly assessed context and background information about and from the target persons. Within the National Educational Panel Study (NEPS; Blossfeld, Roßbach, & von Maurice, 2011), the collection of context information is accomplished through written questionnaires as well as personal and telephone interviews (Frahm et al., 2011). Next to surveying the target person, further context persons, such as parents, teachers, and principals, are asked to participate in the survey to gain a broad spectrum of relevant information following a multi-informant perspective for several items. In addition to reporting about the target persons' own experiences, appraisals, and further personal information, the participants are asked for statements about third persons. Information about the socioeconomic status and ethnical origin of parents and family, in particular, are retrieved via these proxy-reports, as is also done in several other studies (e. g., PISA 2000; OECD, 2002). These variables are important in case parents do not participate in the study and thus do not provide the relevant information.

Former research indicates that linguistic skills and levels of cognitive development may affect the validity of self-reports. Other pivotal factors for valid data are mental representations of the requested topics as well as the relevance of the question content for the respondent (Fuchs, 2009; Looker, 1989). These aspects raise the question of whether children whose cognitive and linguistic abilities are still developing are able to provide valid and reliable information. However, while it can be assumed that adolescents are generally able to answer a questionnaire, it is to be expected that data quality for children under 14 years of age is comparatively lower (Fuchs, 2009). Until now, whether these findings hold true for the group of students with special educational needs in the area of learning (SEN-L) has remained uncertain. Focusing on this special target population, there is no data to maintain such a conclusion. When considering the validity of responses of students with SEN-L, additional factors, such as reduced attentional resources and delayed cognitive and language development, have to be considered (Schröder, 2000).

This chapter sheds some light on the validity of the survey data that was collected within the NEPS from students with SEN-L.

2 Current State of Research

Surveys are an essential part of research for many scientific disciplines. About 90% of collected data derive from surveys (Bortz & Döring, 2006). The number of adolescents and children surveyed has been increasing over the past decades. This trend may be accounted for by two facts: On the one hand, research interest has shifted

more and more to the children themselves as autonomous human beings, to their environments, and to their living conditions. On the other hand, children are often used as proxy reporters, for example, to provide details on the socioeconomic status of their parents (Kränzl-Nagl & Wilk, 2000; Scott, 1997).

2.1 Theoretical Context

Written questionnaires, like other forms of surveys, pose certain demands for the respondent and depend on his or her ability and willingness to reply (Scholl, 2003). The respondent has to pass through a cognitive question-answer process that represents a complex interaction between the respondent and the survey instrument (Fuchs, 2004). Tourangeau's (2000) cognitive model of response behavior assumes four stages of answering questions: comprehension of the question, retrieval of the relevant information, judgment regarding the completeness of the information, and editing a response. Based on Tourangeau's model, Krosnick (2000) established a theory identifying two response behaviors. In contrast to an optimal answering process as described by the four steps above, Krosnick specifies an alternative response behavior called *satisficing*, meaning that not all cognitive steps are conducted, and instead, the first acceptable response alternative is chosen. The likelihood of the occurrence of satisficing is related to three factors. Specifically, difficult tasks or items tend to lead to satisficing for respondents with comparatively lower cognitive abilities and less motivation.

Furthermore, each phase of the cognitive-response behavior can be afflicted with stage-specific errors on behalf of the respondent, such as limited comprehension of the question or lacking mental representations that may lead to invalid answers and reduced data quality. There is also evidence that item characteristics, such as the phrasing of questions and items (Benson & Hocevar, 1985; de Leeuw, Borgers, & Smits, 2004), the number of response categories (Borgers & Hox, 2001), the position and order within the questionnaire (Fuchs, 2004), and the salience for the respondent (Looker, 1989; Lipski, 2000) may impact the reliability of data.

2.2 Surveying Children and Adolescents

In general, studies on the validity of students' responses in surveys judge their answers to be predominantly useful. More specifically, adolescents at secondary school who live at home with their parents have been shown to give reliable proxy reports of their parents' socioeconomic status and familial background (Looker, 1989).

Maaz, Kreuter, and Watermann (2006) analyzed the validity of responses collected from 15-year-old adolescents in Germany. The congruency between the students' and their parents' responses to questions regarding parental education and achieved certificates was examined by the agreement coefficient Cohen's kappa (Cohen, 1960).

The results showed a high amount of conformity between the two reports as well as recognizable differences depending on the type of school attended (for mothers' school-leaving qualification: $K = .50-.80$; for fathers' school-leaving qualification: $K = .42-.67$). Assuming that attending a certain type of school correlates with the cognitive performance of the students, the results indicate that better cognitive abilities may lead to more valid reports on parental education and that lower cognitive abilities may lead to more difficulties in providing correct answers. These findings suggest that even adolescents are not necessarily able to give correct responses regarding their parents' education. However, West, Sweeting, and Speed (2001) showed that 11-year-old children were able to report correctly on their parents' occupation. They found high or very high levels of agreement ($K = .69$ for fathers' occupation; $K = .82$ for mothers' occupation) in addition to low nonresponse rates. Nevertheless, these findings have to be put in context since oral interviews in one-on-one settings were used for the assessments.

Obviously, linguistic demands and complexity of items in a written questionnaire may lead to a challenging answering process. However, compared with item characteristics, child characteristics and abilities seem to play an even more prominent role (Bell, 2007). Borgers, de Leeuw, and Hox (2000), for example, showed that individual differences in reading comprehension of children from the age of 7 to 8 significantly impact on response rates and the consistency of responses. Other results indicate that limited reading competence sometimes influences the response validity of negatively phrased items (Marsh, 1986), thus showing that item and person characteristics interact.

Although research projects in Germany have collected information from students with SEN-L via written questionnaires (Lehmann & Hoffmann, 2009; Wocken, 2005), experiences in surveying this group of students (especially in large-scale assessments) is still rather limited in Germany.

2.3 Students with Special Educational Needs in the Area of Learning

Comprising 40% of all students with SEN, those with SEN-L constitute by far the largest group of students with SEN in Germany (Autorengruppe Bildungsberichterstattung, 2014). It is a highly heterogeneous group with very heterogeneous competence profiles (Antor & Bleidick, 2001). Students with SEN-L have severe and extensive deficits in the accomplishment of cognitive performance requirements lasting over a period of time. Constraints are primarily found in the acquisition of cognitive-verbal and abstract content (Grünke, 2004). These children's ability to cope with learning requirements can be characterized by using and applying fewer strategies for gathering and processing relevant information (Grünke, 2004; Klauer & Lauth, 1997). Working memory and attention span are expected to be comparatively restricted, which may result in difficulties following instructions (Schmetz, 1999). Children

who are assigned to special schools for students with SEN-L usually have difficulties in reading and writing, which impact on various learning areas (Valtin & Sasse, 2012).

3 Research Question

To gain valid and comparable data within large-scale assessments, standardized administrations of tests and questionnaires are implemented. Considering the characteristics of students with SEN-L, it is worthwhile to ask how they cope with constraints and conditions of standardized surveys. Kränzl-Nagl and Wilk (2000) emphasize the challenges of standardized surveys for children whose cognitive development might be called delayed. Aiming to investigate response validity in written questionnaires for students with SEN-L, the following questions were addressed:

- 1) Does response validity differ between students with SEN-L at special schools and students without SEN at regular schools?
- 2) Are there specific differences in the content of the items that are more or less valid for students with SEN-L?
- 3) Does the validity of the responses depend on students' age?
- 4) Are there changes in sustained attention across a given questionnaire that might influence response validity?

With respect to partially limited cognitive abilities, it can be expected that students with SEN-L might provide less valid data than general-education students without SEN-L (Borgers & Hox, 2001; Fuchs, 2009). According to previous research, older students—both general-education students and students attending special schools—are anticipated to be more likely to provide valid information compared with their younger peers. To investigate whether students with SEN-L are attentive and able to provide answers throughout a written questionnaire, we observed their performance throughout the advancing questionnaire, expecting a decline in completion rates.

4 Method

4.1 Sample

This study uses data from the NEPS Starting Cohort 3 and NEPS Starting Cohort 4.¹ Within the two cohorts, a series of feasibility studies was conducted including stu-

1 This paper uses data from the National Educational Panel Study (NEPS): Starting Cohort 3–5th Grade, doi:10.5157/NEPS:SC3:1.0.0. From 2008 to 2013, NEPS data were collected as part of the Framework Programme for the Promotion of Empirical Educational Research funded by the Ger-

dents at special schools: Students with SEN-L were oversampled in Grades 5 and 9. $N = 587$ Grade-5 students with SEN-L were on average $M_{\text{age}} = 11.44$ ($SD_{\text{age}} = 0.65$) years old, and 44.1% were female. The sample of students with SEN-L in Grade 9 comprised $N = 1,186$ students with a mean age of $M_{\text{age}} = 15.55$ ($SD_{\text{age}} = 0.64$) years, and 44.4% were female. As a reference group, data from students attending regular schools were used (Grade 5: $N = 5,208$, $M_{\text{age}} = 10.95$ [$SD_{\text{age}} = 0.52$]; Grade 9: $N = 14,540$, $M_{\text{age}} = 15.19$ [$SD_{\text{age}} = 0.64$]).

In addition to the students filling out an extensive student questionnaire, the parents of participating students were asked to take part in the study. The parent interview was implemented as a computer-assisted telephone interview (Frahm et al., 2011). About 51.4% of parents of Grade-5 students at special schools and 47.0% of parents of students with SEN-L in Grade 9 participated in the study. The participation rate for parents of general-education students was higher (69.3% for Grade 5; 54.3% for Grade 9). Due to varying participation rates, some analyses were restricted to a fraction of the sample.

4.2 Design

With respect to students with SEN-L, the survey follows an experimental design. Specific accommodations for students with SEN-L were implemented to possibly increase and test for aspects of validity. The questionnaire's design is adjusted in terms of (a) length, (b) selected contents, (c) sequence of administrated items, and (d) mode of presentation.

(a) Considering a limited attention span of students with SEN-L, the written questionnaire was reduced in length. Since students with SEN-L attending Grade 5 were surveyed on two days, the written questionnaire was split into two parts. Overall, the amount of items was reduced by 23% compared with the instrument for regular schools. All students attending Grade 9 were also surveyed on two days: one in fall 2010, and one in spring 2011. Overall, the questionnaires for the sample of students with SEN-L at special schools were shortened by 53% in comparison with the regular survey instrument (Skopek, Pink, & Bela, 2012a, 2012b). Each shortened survey instrument was designed to require about 20 minutes.

(b) The questionnaires were arranged to cover a broad spectrum of subjects. For instance, general information about the familial background, socioeconomic status, ethnic origin, and language use was surveyed (Kristen et al., 2011; Stocké, Blossfeld, Hoening, & Sixt, 2011). In addition, the selected content addressed reading engagement as well as nonformal/informal learning environment, school achievement, and

Table 1 Experimental Design

Experimental group: Forward	BI +	m1 + m2 + m3 (+ m4)
Experimental group: Backward	BI +	(m4 +) m3 + m2 + m1

computer usage (Frahm et al., 2011). The selection and compilation of the items was guided by thematic salience for the students as well as by the linguistic and cognitive requirements of the questions.

(c) Moreover, we anticipated that—in the course of the procedure—the attentiveness of students with SEN-L would decrease substantially so that the validity of individual responses might be affected. To identify and test for potential effects of item position, the design allotted a rotation of content-bound modules (m1–m4) in two experimental versions. It is important to note that the module on *basic information* (BI), including questions concerning the ethnical and social origin, was not touched by this variation. The questionnaire was administered in group settings. All testing groups were randomly assigned to one of the experimental conditions *forward* or *backward*, that is, to the original or a reversed sequence of modules (see Table 1).

(d) With respect to the expected partially limited reading fluency and comprehension, it is questionable whether students with SEN-L were able to answer the provided questionnaire in a straightforward manner without any assistance. To circumvent the effects of reading restrictions, the National Center for Educational Outcomes (NCEO) recommended the use of ‘read aloud’ as an essential adaptation when evaluating students with SEN-L (Koretz & Barton, 2003). Therefore, the questions and items were presented orally, that is, they were read aloud by the interviewer using a predefined script. The effect of reading aloud on the validity of responses is not addressed in this chapter (see Gresch, Strietholt, Kandera, & Solga in this volume for an analyses and comparison of these data with regular-school students attending *Hauptschule*).

4.3 Measures and Procedures

Several methods were employed to investigate the validity of the data reported by students with SEN-L and to approach the questions raised above. For a direct assessment of the validity of the students’ data, the parents’ data—which were not always available—were matched with the students’ responses in order to identify congruencies and accuracy. Therefore, a coefficient based on the percentage of factual conformity of two reports is calculated. The chance-corrected agreement coefficient Cohen’s kappa is a standardized measure of agreement that accounts for the expected proportion of agreements by chance (Wirtz & Caspar, 2002). In general, a kappa value > .75 is suggested to indicate very high agreement, while a kappa between .6 and .75 indi-

cates good agreement (Fleiss & Cohen, 1973). Kappa values between .4 and .6 are regarded as acceptable depending on the specific research subject under study (Wirtz & Caspar, 2002). It is important to note that identical reports of students and their parents do not necessarily imply valid and meaningful data. However, the consistency of independently collected information can be seen as an indication of the plausibility of both the students' as well as the parents' reports.

As further indicators for validity, Bell (2007) suggests inspecting inconsistencies of individual response patterns. Particular focus lies on rates of missing values, such as invalid responses and nonresponses, to detect specific content-related refusals or difficulties. By comparing the observed patterns of nonresponse within and across the two experimental conditions, we analyze positional effects related to decreasing attention. This also provides hints as to whether response behavior is more likely to be related to the specific questions and topics or to the item position within the questionnaire.

5 Results and Analysis

In this section, first results regarding the direct measurement of validity are reported, followed by a description of missing values addressing the question of sustained attention throughout the questionnaire.

The direct measurement of validity operationalized via the agreement coefficient Cohen's kappa is only possible for a subset of all administrated items—namely those requesting facts such as ethnic origin and native language. Looking at the kappa values for these items, a distinctive pattern can be observed (see Table 2).

The agreement coefficients vary between the samples of students at regular schools and students at special schools, as well as between the two age-groups. Altogether, the coefficients follow comparable patterns. Items asking about the country of birth of the students themselves, as well as that of their parents, reach high and very high conformity, respectively. However, values of kappa decline for items regarding the third generation. Not only does the congruency between students' and parents' reports decline, but the completion rate of the items also decreases. The rates of missing values rise from less than 3 % up to almost 40 % for these particular items (see Table 3). However, this increasing rate of item nonresponse corresponds to the administrated order of the items, and it seems rather connected to the content of the questions. The response rates for items regarding the native language of both the child and the parents are higher. The agreement-coefficients over $K = .9$ for the samples in both cohorts indicate high validity for these items. Since these questions permit multiple responses for people growing up multilingually, the chances for congruency are higher and also account for high Kappa values.

Notably, the majority of the agreement coefficients are higher for the sample of general-education students in comparison with students with SEN-L. Exceptions are

Table 2 Agreement Coefficient for Ethnic Origin and Native Language

Variables	Grade 5: Special schools	Grade 5: Regular schools	Grade 9: Special schools	Grade 9: Regular schools
Country of birth	.710	.852	.836	.872
Country of birth: Mother	.668	.871	.889	.884
Country of birth: Father	.695	.853	.717	.879
Country of birth: Maternal grandmother	.620	.581	.388	.547
Country of birth: Maternal grandfather	.458	.506	.264	.511
Country of birth: Paternal grandmother	.230	.557	.075	.408
Country of birth: Paternal grandfather	.376	.555	.17	.421
Native language	.945	.951	.939	.954
Native language: Mother	.949	.969	.963	.963
Native language: Father	.911	.957	.906	.958

Table 3 Proportion of Item Nonresponse for Ethnic Origin and Native Language

Variables	Grade 5: Special schools	Grade 5: Regular schools	Grade 9: Special schools	Grade 9: Regular schools
Country of birth	2.3 %	1.2 %	1.4 %	0.7 %
Country of birth: Mother	7.5 %	3.2 %	7.0 %	2.1 %
Country of birth: Father	14.2 %	5.9 %	11.5 %	4.2 %
Country of birth: Maternal grandmother	21.5 %	11.3 %	17.7 %	5.8 %
Country of birth: Maternal grandfather	29.9 %	16.0 %	22.4 %	8.6 %
Country of birth: Paternal grandmother	32.8 %	14.6 %	26.8 %	9.8 %
Country of birth: Paternal grandfather	39.1 %	18.9 %	28.6 %	11.7 %
Native language	2.1 %	2.7 %	2.8 %	0.9 %
Native language: Mother	6.4 %	3.1 %	4.4 %	1.9 %
Native language: Father	10.3 %	4.2 %	9.0 %	3.8 %

Table 4 Agreement Coefficient for Parental Education and Occupation in Grade 9

Variables	Special schools	Regular schools
Highest education qualification: Mother	.288	.526
Highest education qualification: Father	.348	.466
Employment: Mother	.476	.537
Employment: Father	.604	.483
Vocational position: Mother	.262	.435
Vocational position: Father	.370	.565
Occupation: Mother	.504	.586
Occupation: Father	.501	.511

items with a general high congruency, such as native language and the country of birth of the child and parents. Comparing the two cohorts, the coefficients are nearly identical for the sample of general-education students, while for the sample of students with SEN-L, age seems to have an effect. Students attending Grade 9 at special schools reply less validly to various items regarding ethnic origin compared with students attending Grade 5 at special schools or students attending general-education schools.

Students in Grade 9 were also asked about their parents' educational qualifications as well as their employment status and occupation (see Table 4). Overall, these items show lower but partially acceptable congruency according to Wirtz and Caspar (2002). These items seem to cause more difficulties for students to respond validly. For students with SEN-L, particularly low agreement coefficients are found for the questions of the highest education qualification and the vocational position of both parents. With one exception (item: current employment of father), the students at regular schools achieve higher congruency with their parents' reports. About one fourth of the fathers of students attending special schools are reported to be without employment. However, less than 10% of students attending general educational schools report that their fathers are unemployed.

The challenges that questions regarding the educational careers of parents may produce can also be detected by looking at the item nonresponse. For the item of the father's highest educational qualification, rates of missing values are up to 70% for students with SEN-L (see Table 5). Low completion rates reduce the sample considerably so that results are only meaningful for a subsample of the students at special schools.

Regarding the length and volume of the questionnaires, positional effects of the items were observed. As can be seen in Table 6 and Table 7, the amount of item-non-

Table 5 Proportion of Item Nonresponse for Parental Education and Occupation

Variables	Special schools	Regular schools
Highest education qualification: Mother	63.1 %	24.0 %
Highest education qualification: Father	70.8 %	30.0 %
Employment: Mother	10.7 %	6.7 %
Employment: Father	18.5 %	9.1 %
Vocational position: Mother	45.3 %	26.5 %
Vocational position: Father	41.7 %	27.0 %
Occupation: Mother	56.2 %	34.8 %
Occupation: Father	56.7 %	37.6 %

response varies between the two experimental conditions *forward* and *backward*. The anticipated gradual increase of item-nonresponse in the progress of the survey instrument is not observed. The response patterns indicate that missing values are not directly associated with the position of items within the questionnaire. In fact, the occurrence of reduced completion rates does not seem to depend on the position of the item within the questionnaire, but rather on item content. Since less completion rates occur in both groups for the identical items, subject-specific causes can be assumed.

Thus, subjects dealing with reading engagement and nonformal/informal learning environment tend to lead to item-nonresponse for fifth graders. Modules 2 and 3, which are concerned with familial learning environment and school achievement as well as the quality of instruction, show the lowest rates of missing values in both experimental groups.

Students attending Grade 9 at special schools show lower completion rates in comparison with students at general-education schools. However, there are few dif-

Table 6 Proportion of Item Nonresponse for Modules 1–4 in Grade 5

	Forward	Backward	Number of items
Module 1: Reading engagement	8.2 %	10.3 %	17
Module 2: Familial learning environment, School achievement	5.3 %	7.3 %	19
Module 3: Quality of instruction	5.0 %	7.7 %	15
Module 4: Nonformal/informal learning environment	10.4 %	11.3 %	9

Table 7 Proportion of Item Nonresponse for Modules 1–3 in Grade 9

	Forward	Backward	Number of items
Module 1: Reading engagement	10.0 %	10.8 %	11
Module 2: School achievement, Nonformal/informal learning environment	10.6 %	11.9 %	15
Module 3: Computer usage	7.1 %	10.1 %	22

ferences between modules and topics, regardless of item position. Both groups show the lowest rates of missing values for the subject of computer usage.

6 Discussion

In this chapter, the question of whether the use of a survey questionnaire for students with SEN-L in Grade 5 and Grade 9 is valid was addressed. The results on congruency between the students' and parents' reports as well as the completion rates have revealed some challenges regarding surveying students with SEN-L. Analyses have shown that the response data can be considered valid in respect to particular questions. For some subjects, students with SEN-L are capable of responding validly and accurately. Other items lead to difficulties that can result in low completion rates and reduced validity of the students' reports. The absence of mental representations, for example, those regarding the place of birth of grandparents (especially for students who have a background of migration) may yield problems. However, questions concerning native language—a salient feature of daily communication within the family—lead to higher agreement between the students' and parents' reports. In contrast, agreement coefficients for the socioeconomic status of parents illustrate the constraints of using students with SEN-L as proxy reporters.

However, the matching child's and parent's data does not necessarily indicate validity, and the accuracy of parents' reports is not automatically evident. Nevertheless, congruency can be taken as important evidence regarding the value of the students' data. Additionally, parents' information is not available for all students, and it is therefore only possible to gain information for a fraction of the sample. It is rather important to collect proxy reports from the children themselves, particularly for students whose parents did not take part in the survey. The question of validity is especially relevant for this subsample.

In general, students with SEN-L produce higher rates of missing values, which restricts the calculation of agreement coefficients to a subsample of students. Hence, the interpretation of the results is limited. Reduced completion rates also circumvent a valid and comprehensive description of the entire sample. Reporting on the ethnic

origin of students at special schools and their familial background based on the students' questionnaire alone is not feasible. Systemic variations regarding selective non-responses may have effects on further analyses (Kreuter, Maaz, & Watermann, 2004).

Considering the length of the administrated questionnaires, they seem to be suitable for students with SEN-L since the occurrence of item nonresponse has proven to be primarily linked to the content of the questions. The mode of reading aloud may support a continuous response behavior. However, even this administration mode does not lead to complete item response for these questions, which may create difficulties because of content and wording. The mechanism behind the reduced completion rate for specific items needs to be addressed in further analyses.

The approach of post-testing as described in this chapter can only take some aspects into account. The response behavior of editing an answer or not is an important and obvious indicator. However, it is not possible to examine cognitive processes that may lead to certain response behaviors. To understand more about the challenging issues regarding the validity of collected data, further aspects need to be considered. The effects of the respondents' cognitive abilities and the interaction with item characteristics as stated in various studies need to be examined in more detail.

Although the validity of the collected data from students with SEN-L seems evident for various items, caution should still be taken when working with the survey data.

References

- Antor, G., & Bleidick, U. (Eds.). (2001). *Handlexikon der Behindertenpädagogik: Schlüsselbegriffe aus Theorie und Praxis*. Stuttgart: Kohlhammer.
- Autorengruppe Bildungsberichterstattung (2014). *Bildung in Deutschland 2014. Ein indikatorengestützter Bericht mit einer Analyse zur Bildung von Menschen mit Behinderungen*. Bielefeld: Bertelsmann Verlag.
- Bell, A. (2007). Designing and testing questionnaires for children. *Journal of Research in Nursing*, 12(5), 461–469.
- Benson, J., & Hocevar, D. (1985). The impact of item phrasing on the validity of attitude scales for elementary school children. *Journal of Educational Measurement*, 22(3), 231–240.
- Blossfeld, H.-P., Roßbach, H.-G., & von Maurice, J. (Eds.) (2011). Education as a life-long process: The German National Educational Panel Study (NEPS) [Special issue]. *Zeitschrift für Erziehungswissenschaft*, 14. Wiesbaden: VS Verlag für Sozialwissenschaften.
- Borgers, N., de Leeuw, E., & Hox, J. (2000). Children as respondents in survey research: Cognitive developmental and response quality. *Bulletin de Méthodologie Sociologique*, 66(1), 60–75.

- Borgers, N., & Hox, J. (2001). Item nonresponse in questionnaire research with children. *Journal of Official Statistics*, 17(2), 321–335.
- Bortz, J., & Döring, N. (2006). *Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler*. Heidelberg: Springer Medizin Verlag.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), 37–46.
- De Leeuw, E., Borgers, N., & Smits, A. (2004). Pretesting questionnaires for children and adolescents. 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. 409–429). New York: John Wiley & Sons.
- Fleiss, J. L., & Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, 33(3), 613–619.
- Frahm, S., Goy, M., Kowalski, K., Sixt, M., Strietholt, R., Blatt, I., ... Kandera, M. (2011). Transition and development from lower secondary to upper secondary school. In H.-P. Blossfeld, H.-G. Roßbach, & J. von Maurice (Eds.), *Zeitschrift für Erziehungswissenschaft, 14. Education as a lifelong process: The German National Educational Panel Study (NEPS)* (pp. 217–232). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Fuchs, M. (2004). Kinder und Jugendliche als Befragte. *ZUMA-Nachrichten*, 28(54), 60–88.
- Fuchs, M. (2009). The reliability of children's survey responses. The impact of cognitive functioning on respondent behavior. In Statistics Canada (Ed.), *Symposium 2008: Data collection: Challenges, achievements and new directions* (pp. 1–8). Ottawa: Stat-Can.
- Grünke, M. (2004). Lernbehinderung. In G. W. Lauth, M. Grünke, & J. Brunstein (Eds.), *Interventionen bei Lernstörungen* (pp. 65–77). Göttingen: Hogrefe.
- Heydrich, J., Weinert, S., Nusser, L., Artelt, C., & Carstensen, C. H. (2013). Including students with special educational needs into large-scale assessments of competencies: Challenges and approaches within the German National Educational Panel Study (NEPS). *Journal for Educational Research Online*, 5(2), 217–240.
- Klauer, K. J., & Lauth, G. W. (1997). Lernbehinderungen und Leistungsschwierigkeiten bei Schülern. In F. E. Weinert (Ed.), *Psychologie des Unterrichts und der Schule* (Enzyklopädie der Psychologie, Serie Pädagogische Psychologie, Bd. 3, pp. 701–738). Göttingen: Hogrefe.
- Koretz, D. M., & Barton, K. E. (2003). *Assessing students with disabilities: Issues and evidence* (CSE Technical Report No. 587). Los Angeles, CA: University of California, Center for the Study of Evaluation.
- Kränzl-Nagl, R., & Wilk, L. (2000). Möglichkeiten und Grenzen standardisierter Befragungen unter besonderer Berücksichtigung der Faktoren soziale und personale Wünschbarkeit. In F. Heinzel (Ed.), *Methoden der Kindheitsforschung* (pp. 59–76). München: Weinheim.
- Kreuter, F., Maaz, K., & Watermann, R. (2004). Der Zusammenhang zwischen Qualität von Schülerangaben zur sozialen Herkunft und den Schulleistungen. In K.-S. Rehberg

- (Ed.), *Soziale Ungleichheit—Kulturelle Unterschiede, Verhandlungen des 32. Kongresses der Deutschen Gesellschaft für Soziologie in München 2004* (pp. 3465–3478). Frankfurt: Campus.
- Kristen, C., Edele, A., Kalter, F., Kogan, I., Schulz, B., Stanat, P., & Will, G. (2011). The education of migrants and their children across the life course. In H.-P. Blossfeld, H.-G. Roßbach, & J. von Maurice (Eds.), *Zeitschrift für Erziehungswissenschaft, 14. Education as a lifelong process: The German National Educational Panel Study (NEPS)* (pp. 121–137). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Krosnick, J. (2000). The threat of satisficing in surveys: The shortcuts respondents take in answering questions. *Survey Methods Newsletter, 20*(1), 4–8.
- Lehmann, R., & Hoffmann, E. (2009). Anlage und Durchführung der Untersuchung. In R. Lehmann, & E. Hoffmann (Eds.), *BELLA. Berliner Erhebung arbeitsrelevanter Basiskompetenzen von Schülerinnen und Schülern mit Förderbedarf "Lernen"* (pp. 17–29). Münster: Waxmann.
- Lipski, J. (2000). Zur Verlässlichkeit der Angaben von Kindern bei standardisierten Befragungen. In F. Heinzel (Ed.), *Methoden der Kindheitsforschung. Ein Überblick über Forschungszugänge zur kindlichen Perspektive* (pp. 77–86). München: Weinheim.
- Looker, E. D. (1989). Accuracy of proxy reports of parental status characteristics. *Sociology of Education, 62*(4), 257–276.
- Maaz, K., Kreuter, F., & Watermann, R. (2006). Schüler als Informanten? Die Qualität von Schülerangaben zum sozialen Hintergrund. In J. Baumert, P. Stanat, & R. Watermann (Eds.), *Herkunftsbedingte Disparitäten im Bildungswesen: Differenzielle Bildungsprozesse und Probleme der Verteilungsgerechtigkeit. Vertiefende Analysen im Rahmen von PISA 2000* (pp. 31–59). Weinheim: VS Verlag für Sozialwissenschaften.
- Marsh, H. W. (1986). Negative item bias in rating scales for preadolescent children: A cognitive developmental phenomenon. *Developmental Psychology, 22*(1), 37–49.
- OECD (2002). *PISA 2000 technical report*. Retrieved from <http://www.pisa.oecd.org/dataoecd/53/19/33688233.pdf>
- Schmetz, D. (1999). Förderschwerpunkt Lernen. *Zeitschrift für Heilpädagogik, 4*, 134–143.
- Scholl, A. (2003). *Die Befragung*. Konstanz: UVK-Verlag.
- Schröder, U. (2000). *Lernbehindertenpädagogik: Grundlagen und Perspektiven sonderpädagogischer Lernhilfe*. Stuttgart: Kohlhammer.
- Scott, J. (1997). Children as respondents: Methods for improving data quality. In L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz, & D. Trewin (Eds.), *Survey measurement and process quality* (pp. 331–350). New York: Wiley.
- Skopek, J., Pink, S., & Bela, D. (2012a). *Data manual. Starting Cohort 3—From lower to upper secondary school. NEPS SC 3 1.0.0* (NEPS Research Data Paper). Bamberg: University of Bamberg, National Educational Panel Study.
- Skopek, J., Pink, S., & Bela, D. (2012b). *Starting Cohort 4: 9th grade (SC4). SUF-Version 1.0.0. Data Manual* (NEPS Research Data Paper). Bamberg: University of Bamberg, National Educational Panel Study.

- Stocké, V., Blossfeld, H.-P., Hoenig, K., & Sixt, M. (2011). Social inequality and educational decisions in the life course. In H.-P. Blossfeld, H.-G. Roßbach, & J. von Maurice (Eds.), *Zeitschrift für Erziehungswissenschaft, 14. Education as a lifelong process: The German National Educational Panel Study (NEPS)* (pp. 103–119). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Tourangeau, R. (2000). *The psychology of survey response*. Cambridge: University Press.
- Valtin, R., & Sasse, A. (2012). Schriftspracherwerb. In W. Schrader, & F. B. Wember (Eds.), *Didaktik des Unterrichts im Förderschwerpunkt Lernen* (pp. 179–190). Stuttgart: Kohlhammer.
- West, P., Sweeting, H., & Speed, E. (2001). We really do know what you do: A comparison of reports from 11 year olds and their parents in respect of parental economic activity and occupation. *Sociology, 35*(2), 539–559.
- Wirtz, M., & Caspar, F. (2002). *Beurteilerübereinstimmung und Beurteilerreliabilität*. Göttingen: Hogrefe.
- Wocken, H. (2005). *Andere Länder, andere Schüler? Vergleichende Untersuchungen von Förderschülern in den Bundesländern Brandenburg, Hamburg und Niedersachsen* (Forschungsbericht). Retrieved from http://www.mbj.s.brandenburg.de/sixcms/media.php/5527/wocken_ergebnis-heft.pdf

About the authors

C. Artelt
Department of Educational Research,
University of Bamberg, Bamberg.

C. H. Carstensen
Psychology and Methods of Educational Research,
University of Bamberg, Bamberg.
Leibniz Institute for Educational Trajectories (LifBi), Bamberg.

J. Heydrich
University of Bamberg, Bamberg.

L. Nusser
Department of Psychology I: Developmental Psychology,
University of Bamberg, Bamberg
e-mail: lena.nusser@uni-bamberg.de

S. Weinert
Department of Psychology I: Developmental Psychology,
University of Bamberg, Bamberg

The Conceptualization, Development, and Validation of an Instrument for Measuring the Formal Learning Environment in Higher Education

Hildegard Schaeper and Thomas Weiß

Abstract

Our article describes the conceptualization and measurement of the formal learning environment in higher education that was used in the substudy “Stage 7—From Higher Education to the Labor Market” of the National Educational Panel Study (NEPS). On the basis of a coherent conceptual framework adopted throughout the NEPS, we developed a parsimonious questionnaire that proved to be a valid and reliable instrument for measuring central dimensions of the process quality of higher education, namely structure, support, challenge, and orientation. This article presents the results of the reliability and validity analyses and identifies areas for further improvements.

1 Introduction

Nowadays, the fact that contextual factors play a significant role in educational decision-making and competence development is not disputed. Therefore, the National Educational Panel Study (NEPS) places special emphasis on learning environments and captures the most relevant dimensions that are expected to impact on learning and the educational career. According to a widespread distinction, educational contexts can either be formal, nonformal, or informal (see Bäumer, Preis, Roßbach, Stecher, & Klieme, 2011). In our paper, we focus on formal learning environments, that is, organized educational settings in educational and other organizations. Typically, certification is also a constituent part of the concept of formal learning environments. In the NEPS, however, this element is neglected in favor of stringency and of also being able to include typical formal learning environments in Kindergartens and companies (see Bäumer et al., 2011).

Despite the significance attached to the institutional context, coherent and theory-based conceptualizations of German higher education institutions as formal learning environments are rare. The few existing models—two of them are briefly discussed in Section 2.1—adopt a multilayer perspective and either derive relevant context dimensions tentatively or in a theory-driven manner.

In German instructional research, a different approach has been proposed. This approach, which is described in more detail in Section 2.2, was developed for conceptualizing and analyzing formal learning environments at schools and distinguishes four dimensions: structure, support, challenge, and orientation (SSCO). This so-called SSCO model has been adopted throughout the NEPS as a theoretical basis for measuring the process quality of any learning environment (Bäumer et al., 2011).

Combining the SSCO approach and the multilayered structure of learning environments leads to a highly differentiated model that becomes even more complex when different perspectives on learning environments, namely objectivist and subjectivist approaches, are taken into account (see Section 2.3). We used this model as a general framework for capturing the formal learning environment in higher education and for developing a questionnaire to be used in the NEPS study “Stage 7—From Higher Education to the Labor Market.”

The measurement instrument captures the perceived learning environment at the level of the degree program and underwent a rigorous assessment and selection process (see Sections 3 and 4). It was developed in several steps and in close cooperation with NEPS Pillar 2—Education Processes in Life-Course-Specific Learning Environments. Starting with a relatively large initial pool of items (see Section 3), we finally arrived at a parsimonious instrument of 42 items that was used in the main study of the NEPS Starting Cohort 5—First-Year Students and proved to adequately represent the SSCO model (see Section 4).

Although the psychometric properties of the questionnaire are satisfying and we exploit other data sources to represent additional layers, facets, and perspectives, we still see desiderata. These desirable or necessary requirements for an encompassing measurement of the formal learning environment in higher education are discussed in Section 5.

2 Conceptualizations of Learning Environments

According to constructivist learning theories, learning is a context-bound, social process of active construction. In this perspective, the learning situation with other involved actors; the physical, social, and organizational conditions; and learning opportunities is as important as the learner with his/her activities, individual characteristics, and time spent studying. As early as 1762, Rousseau acknowledged the significance of the environment for human development and education: “We are born sensitive and from our birth onwards we are affected in various ways by our environment”

(Rousseau, 1921, p. 7). As late as the second half of the last century, first attempts to systematically conceptualize and measure German institutions of higher education as learning environments were published. Until now, though, only a few additional conceptualizations have been proposed in Germany.

In contrast to the state of research in Germany (and in Europe), studying the impact of colleges and universities on students has a long tradition in the USA (see Dippelhofer-Stiem, 1986), and the body of literature in this area is vast. Providing a complete account of the work done on this topic is, however, beyond the scope of this paper (for the USA, see the comprehensive overview of three decades of research given by Pascarella and Terenzini (1991, 2005)). In lieu of reviewing this research, which is often a-theoretical and partly yields contradictory empirical results (Dippelhofer-Stiem, 1986), we focus on two theory-driven approaches to German institutions of higher education as learning environments: the multilayer models of Dippelhofer-Stiem (1983, 1986) and Wosnitza (2007).

2.1 Institutions of Higher Education as Formal Learning Environments: Approaches in Germany

The conceptualizations of German higher education institutions as a formal learning environment proposed by Dippelhofer-Stiem (1983, 1986) and Wosnitza (2007) both refer to Bronfenbrenner (1979). His ecological systems theory conceives of the environment topologically as “a nested arrangement of concentric structures, each contained within the next” (Bronfenbrenner, 1979, p. 22) and distinguishes among four system levels: the *microsystem* (“pattern of activities, roles, and interpersonal relations ... in a given setting with particular physical and material characteristics” (Bronfenbrenner, 1979, p. 22)), the *mesosystem* (“interrelations among two or more settings in which the developing person actively participates”; “system of microsystems” (Bronfenbrenner, 1979, p. 25)), the *exosystem* (external microsystems in which events occur that influence the immediate setting (see Bronfenbrenner, 1979, p. 25)), and the *macrosystem* (common patterns of micro-, meso-, and exosystem characteristics in a given culture or subculture (see Bronfenbrenner, 1979, p. 26)). Later, Bronfenbrenner (1986) added the *chronosystem*—a term that includes the dimension of time and that refers both to the individual’s movement through different systems (transitions, life course) and to historical changes of the environments.

In applying this multilevel perspective of hierarchically ordered layers to German higher education, Dippelhofer-Stiem (1983, 1986) and Wosnitza (2007) propose similar but slightly different models: Dippelhofer-Stiem distinguishes between (a) the higher education system as being embedded in a national framework, (b) the university or college (institutional level), (c) the subject area and department, (d) the course, and—we return to this issue in Section 2.3—(e) the individual. Wosnitza adopts a broader definition of the macrosystem and uses the term to describe the reality out-

side the higher education institution in general, for example, the society, the culture, and the regional context. Like Dippelhofer-Stiem, Wosnitza places the higher education institution in the exo-level, followed by the field of study or degree program (mesosystem). The lowest level is the teaching-learning unit within a course.

Apart from the structure of nested environmental levels, both authors introduce an additional component to their models: dimensions covering and structuring the relevant aspects of the different environmental levels. Dippelhofer-Stiem (1983, 1986), who is primarily interested in the process of socialization in higher education, takes the aims and intentions of higher education and the prospective outcomes of students' socialization as a starting point and derives four dimensions: (a) academic freedom, (b) interdisciplinarity, (c) communication and participation, and (d) practice and social relevance. Inspired by Lewin's (1936) distinction between quasi-physical, quasi-social, and quasi-conceptual facts, Wosnitza (2007) starts out with different types of objects effective in the environment and identifies the dimensions (a) of material-physical aspects of learning environments, (b) of social aspects, and (c) of formal aspects. Unfortunately, Wosnitza does not define these categories explicitly. This is particularly disadvantageous in the case of formal aspects. Only by reading his empirical studies is it possible to get an idea of what is meant. It turns out that formal aspects are heterogeneous and include diverse attributes, such as practice orientation, interdisciplinarity, and modes of teaching at the micro-level; organization of the degree program and information at the meso-level; counseling services, leisure activities, opening hours, and tuition fees at the exo-level; and financial support and employment opportunities at the macro-level (Wosnitza, 2007, pp. 148–149).

Dippelhofer-Stiem (1983, 1986), as well as Wosnitza (2007), addresses the question of perspective, that is, the point of view from which the environment is measured. Both authors distinguish between (a) a subjectivist approach, which aims at measuring the environment as perceived by the actors involved (especially by the students), and (b) an objectivist approach, which attempts "to describe the environment as if from the outside" (Dippelhofer-Stiem, 1986, p. 476) to produce intersubjectively verifiable data or to assess the 'potential' environment that exists independent of the individual perception.

2.2 Support, Structure, Challenge, and Orientation: Four Basic Dimensions of the Quality of Learning Environments

In contrast to the more or less inductive approach to identifying relevant dimensions of learning environments adopted by Dippelhofer-Stiem (1983, 1986) and Wosnitza (2007), Klieme, Lipowsky, Rakoczy, and Ratzka (2006) start from pedagogical-psychological theories and general instructional concepts and link them to empirically confirmed effects on learning outcomes. In doing so, they identify three basic dimensions of the process quality of the learning environment in schools: structure, support,

and challenge. Radisch, Stecher, Klieme, and Kühnbach (2007) add orientation as a fourth dimension. They argue that research always has to consider structure, support, challenge, and orientation as quality attributes of learning environments in the classroom, in extracurricular activities, in out-of-school activities, and in the family, and to determine their potential for educational processes (Radisch et al., 2007). These dimensions are referred to as SSCO in the NEPS, and they guide the measurement of the process quality of different formal, nonformal, and informal learning environments (Bäumer et al., 2011).¹

In higher education (as elsewhere), the structural dimension is concerned with the degree of clarity, organization, transparency, stability, and safety of learning opportunities (e. g., rules, learning conditions, study requirements, and expectations). *Support* involves helping students to develop competencies, to gain a certain degree of autonomy, and to cope with study requirements or social integration. *Challenge* focuses on cognitive activation as a means of promoting deep understanding and preventing inert knowledge. *Orientation* refers to “shared values and norms of the actors, coherence among actors, general attitudes and orientations related to educational processes, [and] attitudes toward attributions of academic achievements” (Bäumer et al., 2011, p. 94). In higher education, this dimension includes, for example, practice orientation, research orientation, interdisciplinarity, achievement orientation, and the emphasis placed on internationalization (Aschinger et al., 2011).

While these dimensions cover central properties of the process quality of education, a fifth and sixth dimension address the input quality and the context in the educational effectiveness framework proposed by Scheerens and Bosker (1997; see also Klieme & Rakoczy, 2008). In the NEPS, we focus on two dimensions: *structural characteristics* (“comparatively persistent general conditions for educational processes” (Bäumer et al. 2011, p. 95), such as material, financial, and human resources) and *contextual characteristics* (“framing conditions of the learning environment” (Bäumer et al. 2011, p. 95)), which include regional characteristics, such as economic structure and population as well as settlement characteristics. In our adapted model of the formal learning environment in higher education, we combine the two categories into one and call it *structural opportunities and restrictions* (SOR).

1 Here, the learning environment approach meets prominent conceptualizations of educational quality and educational effectiveness, which distinguish structural, process, and outcome quality (see the general framework of quality suggested by Donabedian (1980)) or—according to the framework for educational effectiveness research proposed by Scheerens and Bosker (1997)—the dimensions of input, process, outcome, and context.

2.3 The Multilayered SSCO-SOR Approach: An Integrated Model of Higher Education as a Formal Learning Environment

In our own proposal for a conceptualization of the formal learning environment in higher education, we integrate the multilayer perspective suggested by Dippelhofer-Stiem (1983, 1986) and Wosnitza (2007) and the SSCO or SSCO-SOR approach of the NEPS and add the dimension of perspective (like Wosnitza, 2007) and an additional layer (similar to Dippelhofer-Stiem, 1983, 1986) (Figure 1).

In accordance with Dippelhofer-Stiem (1983, 1986) and Wosnitza (2007), we conceptualize the formal learning environment in higher education as a nested arrangement of contextual levels. The individual student is located at the lowest level or in the center of the learning environment. In incorporating this level, we follow the *opportunity-use model* proposed by Fend (2008), who argues—in line with interactionist sociological, psychological, and educational theories—that the quantity and quality of learning opportunities represent the one side of the coin, and the “user,” with his/her individual characteristics and the way he/she uses the learning opportunities, represents the other side. The learner is, of course, always part of the educational context, and his/her individual attributes (e. g., cognitive competencies, motivation, and interests) and use of learning opportunities (e. g., time spent on studying) are systematically measured in the NEPS.

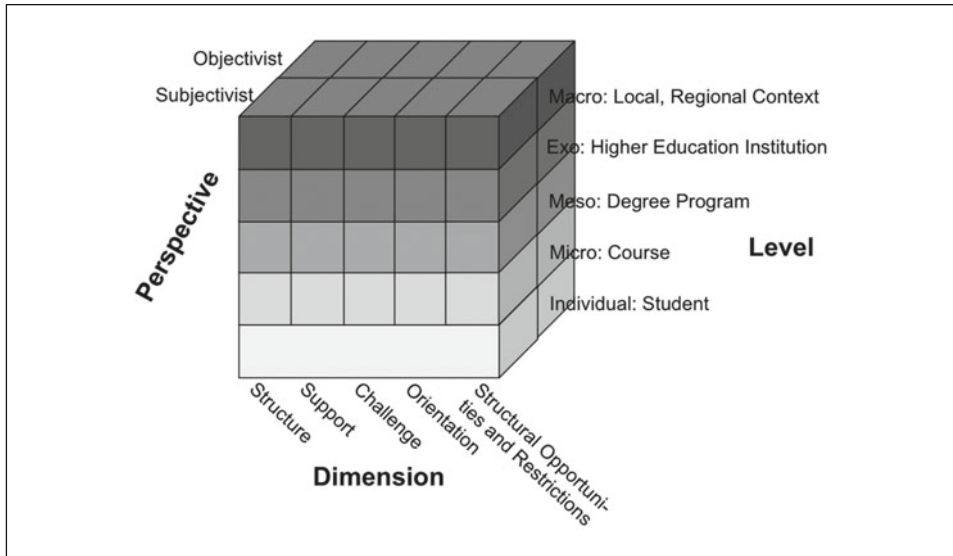
It is well known from instructional research that data on the proximal learning environment in school should be collected for single courses or even smaller instructional units. Such a detailed measurement, however, is difficult to realize in large-scale studies like the NEPS. Even though many items in our questionnaire refer to the classes taken by the students (see Section 3.2), the information gathered provides a generalized description of the courses offered by the degree program in a given period of time. As a consequence, the micro-level is not assessed in the study and, apart from the individual level, the degree program (and not the courses) is the lowest level and smallest unit of analysis.

In addition to the degree program, which is also the focus of our measurement instrument, we consider, as do other researchers, the higher education institution as a separate environmental level. We do not, however, place the educational system, the society, or the culture in the macro-level, but rather the local and regional context. The societal level has to be taken into account when internationally comparative studies or comparisons over time are intended. It can be neglected for the current purpose.

Conceptually, the SSCO-SOR dimensions are relevant for all levels except the individual level. In practice, we place varying emphasis on the different dimensions and measure the broader context and the higher education institution exclusively (or mostly) with regard to structural opportunities and restrictions.

Regarding the perspective, we use the subjectivist as well as the objectivist approach. However, the primary source of information, especially on the study pro-

Figure 1 The multilayered SSCO-SOR model of the formal learning environment in higher education



gram, is the student. The use of subjective evaluations has often been criticized for not providing reliable and valid data. We hold the view that measuring the learning environment as perceived by the students is justified on several grounds: First, it is not possible to carry out analyses of documents describing the broad range of degree programs included in the NEPS (e. g., study and examination regulations) or surveys among lecturers in a time- and cost-effective manner. Second, it follows from several studies that taking the average of the respondents' answers provides a relatively unbiased and valid picture of the learning environment and the quality of teaching and learning (Klieme & Rakoczy, 2003; Teichler et al., 1987). The deviation from the mean can then be interpreted as an individual characteristic of the students. In view of the NEPS's goal of understanding and explaining educational decisions and competence development, the third argument is perhaps the most important one: We assume that the perceived learning environment is at least as important as the "objective" learning environment. Our view is supported by Bronfenbrenner (1979), who asserts that "what matters for behavior and development is the environment as it is *perceived* rather than as it may exist in 'objective' reality" (p. 4). As the famous Thomas theorem puts it: "If men define situations as real, they are real in their consequences" (Thomas & Thomas, 1928, pp. 571–572).

Nonetheless, we also take the objectivist perspective and collect data mainly concerning the structural opportunities and restrictions at the meso-, exo-, and macro-level by analyzing documents and, primarily, statistics.

3 Operationalization

3.1 The Objectivist Approach

In order to operationalize the formal learning environment in higher education on the basis of “objective” indicators, we predominantly used the most recent data collected by the Federal Statistical Office and the Statistical Offices of the *Länder*. In addition, we gathered relevant information published on the internet. Table 1 presents an overview of the levels, dimensions, and subdimensions addressed as well as examples of the generated indicators.

Table 1 Operationalization of the Formal Learning Environment in Higher Education Using Objective Indicators (Examples)

Level	Dimension	Subdimension	Indicator
Local, regional context	SOR	Settlement structure	BK classification of urban regions
		Labor market	Unemployment rate
			Proportion of low/highly educated in the labor force
		Economic structure and situation	GDP per capita
Economic sector of employment			
		Social structure	Graduation rates at tertiary level
Higher education institution	SOR		Type of higher education institution
			Institutional control
	Size	Number of students	
	Orientation		“Excellence Initiative” nominees and winners
			Finalists and winners in the competition “Excellence in Teaching”
Degree program/ subject area	SOR	Size	Number of students
		Social composition	Distribution of students by gender
		Financial resources	Institutional funds
			External funds
		Human resources	Number of academic staff
	Student-teacher ratio		

3.2 The Subjectivist Approach

As mentioned above, the subjectivist approach is our main emphasis, and we chose this perspective exclusively to capture the SSCO dimensions with respect to the degree program. The operationalization of the learning environment at this level was done in several steps. First, we identified relevant subdimensions on the basis of theoretical considerations and results of empirical research. Second, we reviewed existing survey instruments used in higher education research, for example, the questionnaires developed by Dippelhofer-Stiem (1983, 1986) and Wosnitza (2007) as well as the Konstanz Student Survey (Ramm, Multrus, & Bargel, 2011). Third, we assigned the items to the theoretical dimensions and subdimensions, selected the most appropriate ones, and modified them, if necessary. In addition, we constructed new items when a subdimension was not sufficiently represented.

The measurement of the *structural* dimension of the SSCO model focuses on two subdimensions: (1) “structuredness of teaching,” which is represented by the subscales “transparency of performance requirements” (three items) and “structuredness of lectures and classes” (five items), and (2) “structuredness of the study program,” with the subscales “transparency of the structure of the degree program” (three items) and “coordination of courses offered” (three items).

The *support* dimension was operationalized by five subscales, which can be subsumed under the categories “teaching” and “social climate.” Supportive teaching is covered by the subscales “leeway of choice and participation” (six items), “teaching skills and commitment to teaching” (three items), and “motivation” (three items). For measuring the social climate, we have to take two main aspects into account (see also Wosnitza, 2007), which are represented by the subscales “rapport with the lecturers” (six items) and “rapport with fellow students” (seven items).

Seven subscales were developed to measure the dimension *challenge*. Four subscales are based on theoretical conceptualizations of teachers’ approaches to teaching, that is, subjective beliefs about learning and teaching that distinguish between a teacher-focused orientation and a student-centered approach (Kember, 1997; Trigwell & Prosser, 2004; Trigwell, Prosser, & Taylor, 1994). These subscales are labeled “meaning orientation” (four items), “reproduction orientation” (three items), “knowledge construction” (four items), and “knowledge transmission” (three items). The subscales *meaning orientation* and *reproduction orientation* refer to the type of cognitive process stimulated by the learning environment, and the emphasis that is placed on remembering on the one hand and understanding, analyzing, and evaluating on the other (see Anderson & Krathwohl, 2001). The subscales *knowledge construction* and *knowledge transmission* are linked to the lecturers’ beliefs about how learning occurs.

While Trigwell and Prosser (2004) assume that approaches to teaching are a one-dimensional bipolar construct with the teacher-focused approach at one extreme and the student-focused approach at the other, there is now empirical evidence that ap-

proaches to teaching are not a question of either/or, but rather of both/and (Braun & Hannover, 2008; Lübeck, 2009). Correspondingly, we do not necessarily expect the subscales that represent the two approaches to teaching to be highly correlated.

In addition to the aforementioned subscales, the dimension *challenge* also embraces the subscale “pressure to perform” (six items) and subscales that refer to particular instructional (and learning) practices: “collaborative learning” (five items) and “variation” (two items).

In order to measure the dimension *orientation*, we constructed five subscales. The subscales “research orientation” (six items) and “practice orientation” (nine items) refer to two main functions of higher education: the “academic educational function,” that is, preparing future academics, and the “professional educational function” (Teichler & Kehm, 1995), that is, preparing students for professional careers outside academia. Traditionally, the main types of German higher education institutions—universities and universities of applied sciences (*Fachhochschulen*)—used to focus on either the academic educational function (universities) or the professional educational function (universities of applied sciences). The Bologna Process, with its emphasis on employability, however, changed the situation (for a critical appraisal of the employability discourse in German higher education, see Schaeper & Wolter, 2008; Teichler, 2011): Universities, too, are now increasingly expected to pay more attention to professional fields that are not research-related. Conceptually, the educational orientation of a degree program is not considered to be a one-dimensional concept with two poles. Higher education still is (or claims to be) based on scientific knowledge, research, and scientificity (Schaeper & Wolter, 2008). This principle did not become outdated with the Bologna Process. In addition, research may represent a specific type of practice.

The subscale “interdisciplinarity” (six items) captures another educational orientation that is supposed to support the development of students’ personality and generic competencies, particularly critical thinking (Dippelhofer-Stiem, 1986). The subscale “social relevance” (three items) alludes to the ‘social’ educational function of higher education, that is, qualifying for “acting responsibly in a free, democratic, and social state governed by the rule of law” (Hochschulrahmengesetz (*Framework Act for Higher Education*), § 7; our translation) and reflects the social relativity of science and its social and ethical implications (Dippelhofer-Stiem, 1986). Finally, the subscale “internationality” (three items) takes up an issue that became key during the 1990s (Teichler, 2004) and is central to the Bologna Process, namely, promoting international mobility. However, a degree program’s international orientation involves more than studying abroad and also includes a focus on intercultural competencies as well as the internationalization of curricula and the content of teaching (Teichler, 2004).

The first version of the questionnaire for measuring the SSCO dimensions of the formal learning environment in higher education at the program level consisted of 22 scales with a total of 93 items. All items are positively worded. Responses were

given on a fully labeled five-point scale. Depending on the exact wording of the instruction, responses range from “does not apply at all” to “fully applies” and from “do not agree at all” to “completely agree”, for example.

In the survey instrument, additional questions are included that address different dimensions and levels, for example, perceived labor market perspectives (macro-level, SOR dimension) as well as counseling and information services (exo-level, support dimension). As their format mostly differs from that of the SSCO items described above and the questions do not refer to the program level, they are not a subject of the following discussion.

4 Testing, Validation, and Results of the Main Study

The items developed for measuring the SSCO dimensions at the program level served as a pool for selecting the most suitable ones in order to arrive at a parsimonious instrument that meets psychometric standards. To this end, we conducted cognitive interviews, carried out a developmental study, and used the pilot study for further optimizing the questionnaire. As the intent was to administer the final questionnaire online, both the developmental and the pilot study were performed online.

4.1 Cognitive Interviews

In the winter term 2009/2010, we conducted guided interviews with 15 students (six women, nine men) of different ages, semesters, and degree programs from three higher education institutions (university, university of applied sciences, medical school). The objective of the cognitive interviews, which lasted 65 minutes on average, was to identify problems in understanding the questions and to gain insight into the mental processes evoked by the questions. For this purpose, we used the cognitive techniques of probing (comprehension, category selection, information retrieval, general) and paraphrasing (Willis, 2005). Because of the length of the questionnaire, we decided against the think aloud method.

The cognitive testing provided two main results:

- 1) The entire questionnaire included questions addressing different environmental levels, but the intended frame of reference was not always clear to the respondents. As a consequence, we inserted introductory paragraphs before the different sections of the questionnaire that clearly stated the environmental level to which the subsequent questions referred.
- 2) Interviewees who were enrolled in a multiple subject program often had difficulties rating the items referring to the program level, especially when experiencing different educational contexts in their fields of study. In this situation, they tend-

ed to choose the midpoint of the scale, but felt not at ease with it and preferred to answer the questions separately for each subject. As this would have increased the duration of the survey—and thus the burden of the respondents—we introduced the so-called “reference field of study”: In case of multiple majors, respondents were asked to specify the subject area they were referring to. In case of one major and one or two minors, respondents were asked to describe the learning environment in their major.

4.2 Developmental and Pilot Study

A revised version of the questionnaire with all items included was used in a developmental study in the winter term 2009/2010. The aim of this study was to test and validate single items and scales. Furthermore, it was intended to provide a basis for item and scale selection. The questionnaire was administered to a randomly selected subsample of 2,180 students who had agreed to participate in online surveys carried out by the German Centre for Higher Education Research and Science Studies (formerly HIS-Institute for Research on Higher Education) and to give information on current topics of higher education research and policy at regular intervals (“HISBUS online panel”). 614 respondents completed the questionnaires, which corresponds to a response rate of 28.2%. The sample included students from different kinds of higher education institutions, types of degrees, various disciplines, and semesters.

In order to assess the quality of items and subscales, we performed descriptive statistical analyses of item nonresponse and skewness, exploratory item and scale analyses (Kaiser-Meyer-Olkin measure of sampling adequacy, principal factor analysis (PFA), Cronbach's alpha, item discrimination), and confirmatory factor analyses (CFA). The results of the descriptive and exploratory analyses served as a basis for excluding items that proved not to be suitable.

Items that were considered to be acceptable were included in confirmatory factor analyses using a maximum likelihood estimator. Model fit was assessed using the comparative fit index (CFI), the Tucker-Lewis index (TLI), and the root mean square error of approximation (RMSEA).

With some exceptions, the hypothesized factor structure was confirmed empirically, and the internal consistency of the subscales ($\alpha \geq .69$) and item-test correlation ($r_{it} > .70$) were satisfying. Furthermore, the model fit for the different SSCO dimensions was acceptable (RMSEA $\leq .089$) to good (RMSEA $\leq .060$) as was the fit for the complete SSCO model (RMSEA = .053).

On the basis of the developmental study, it was possible to significantly shorten the questionnaire and to confirm the theoretical model with a reduced set of variables. On the other hand, it was necessary to develop additional items in order to improve the quality of some scales. The resulting questionnaire, which was used in the pilot study described below, consisted of 63 items representing 16 subscales.

The pilot study preceded the main study of NEPS Starting Cohort 5—First-Year Students. It served to test data collection procedures and instruments and thus to provide evidence for improving the main study. The sample consisted of first-year students from the winter term 2009/2010 who were enrolled in selected degree programs at three higher education institutions. The learning environment questionnaire was applied in the first online wave, which was carried out in summer 2010. 246 target persons, or 51.4 % of the 479 panel members who were invited to participate in the survey, completed the questionnaire.

By and large, item and scale analyses confirmed the results obtained in the developmental study. Although the findings were acceptable and did not suggest that a fundamental revision of the instrument was required, we had to significantly modify (i. e., shorten) the questionnaire before using it in the main study. The reason for the necessity to reduce the survey instrument lies in a modification of the study design. Initially, the online panel waves were split up into two surveys of 20 minutes in length, which were to be carried out in quick succession. The unsatisfactory response rate achieved in the pilot study, however, led to the decision to combine the two surveys into one. Since a survey length of 40 minutes is not acceptable in online surveys (Bošnjak, 2002), we consequently had to shorten the questionnaire. Instead of eliminating single items, which could have impaired the psychometric quality of the scales, we opted for dropping entire scales (e. g., internationality, social relevance) and for keeping the most essential ones. Analyses completed with the reduced set of 42 items representing 11 subscales revealed acceptable to good psychometric properties. The results are similar to those obtained in the main study (see Section 4.3) and are therefore not reported here.

4.3 The Main Study

Data

The analyses described below use data from Starting Cohort 5 (NEPS SC5, version 4.0.0;² see Blossfeld, Roßbach, & von Maurice, 2011). This substudy longitudinally follows a cohort of new entrants to higher education who enrolled for the first time at a German higher education institution in the winter semester 2010/2011. The sampling procedure can be described as a disproportional one-stage cluster sampling (for details, see Aschinger et al., 2011; Aßmann et al., 2011). Data were collected two or three times a year using different modes of data collection: self-administered questionnaires (only at the beginning of the study), computer-assisted telephone interviewing, online surveys, group-administered tests in classroom settings, and online tests (see Aschinger et al., 2011).

2 Doi:10.5157/NEPS:SC5:4.0.0.

During the winter term 2010/2011 and the summer term 2011, several thousands of first-year students were asked to complete a short questionnaire and to agree to participate in the panel study. Almost 18,000 students who complied with the request and provided valid contact information, belonged to the target population, and participated in the first telephone interview were included in the panel study. The initial questionnaire survey was followed by a telephone interview, which was conducted from winter 2010 to winter 2011 and partly overlapped with the first competence testing in summer 2011. The first online survey was carried out at the beginning of the winter term 2011/2012. A total of 23,809 people were invited to participate in the survey. A portion of them were afterwards excluded from the panel study because they did not belong to the target population or declined to participate in the first telephone interview. 14,606 individuals filled out the questionnaire, which corresponds to a response rate of 61.3 %. The number of valid cases, i. e., cases that were not excluded from the panel, is smaller and amounts to 12,275. The response rate in relation to the invited sample of valid cases is 67.9 %.

As already mentioned, the instrument for measuring the SSCO dimensions at the program level, which was applied in the main study, consists of 11 subscales and 42 items. Table 2 gives an overview of the dimensions, subscales, and variables, as well as a short description of the items.

Results

Item and scale analyses were performed on the basis of the hypothesized model using the methods described above. The values of the item-test correlation coefficients (r_{it}) range between .60 and .92; only three items have values below .70. The results of the confirmatory factor analysis and the analysis of internal consistency are presented separately for each SSCO dimension below.

Regarding the structural dimension, Cronbach's alpha of .65 and .64, which was computed for the two four-item scales displayed in Figure 2, indicates an acceptable internal consistency of the subscales. The results of the confirmatory factor analysis are ambivalent. On the one hand, the overall model fit is good ($RMSEA \leq .05$). The significant chi-square statistic can be neglected because the test is sensitive to sample size and our sample is large. The factor loadings, on the other hand, suggest that the latent constructs are not measured perfectly since some items have low loadings. Because of the low factor loading of variable t243411 in the model with unique loadings and because of the ambiguous character of this item (see discussion in the final section), we additionally estimated a cross-loading of this indicator (see Figure 2). This re-specification resulted in an improvement of the model fit. Because the subscales address different levels—the degree program on one hand and lectures or sessions on the other hand—it is not surprising that the latent constructs correlate only moderately ($r = .47$) and that a second-order factor could not be estimated.

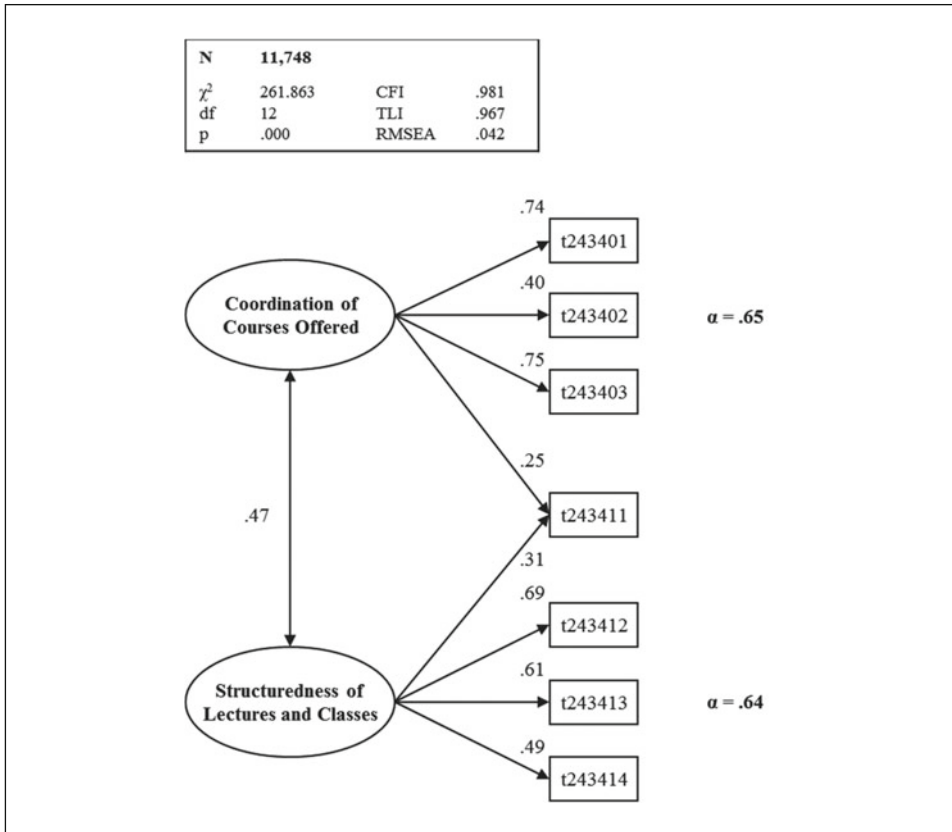
The assessment of the measurement model specified for the support dimension provides empirical evidence for a high degree of reliability and validity (see Figure 3).

Table 2 Measurement Instrument of the SSCO Dimensions at the Program Level Used in the Main Study

Dimension	Subscale	Variable	Description
Structure	Coordination of courses offered	t243401	Structure of courses offered makes it possible to see the connections between them
		t243402	Good coordination of courses in terms of time
		t243403	Good coordination of courses in terms of content
	Structuredness of lectures and classes	t243411	Clearly defined course objectives
		t243412	Lecturers summarize periodically
		t243413	Lecturers establish links between sessions
		t243414	Teaching staff gives introductory overview of session
Support	Rapport with lecturers	t244401	Instructors are responsive to students
		t244402	Teaching staff is cooperative
		t244403	Lecturers give attention to students' problems
	Rapport with fellow students	t244411	Students help each other
		t244412	Students show solidarity
		t244413	Students are working together
	Motivation	t244421	Lecturers present in an interesting way
		t244422	Instructors promote enjoyment of the subject
		t244423	Staff evokes students' interest in the subject
	Challenge	Pressure to perform	t245401
t245402			Enough free time
t245403			Heavy exam load
Meaning orientation		t245411	Emphasis on understanding relationships
		t245412	Promotion of critical reflection
		t245413	High value on critical comparison of theories
		t245414	Emphasis on independent thinking
Reproduction orientation		t245431	Exams mostly require reproduction of what has been learned
		t245432	Good memory is enough to do well
Knowledge construction		t245421	Instructors promote active engagement
		t245422	Teaching staff stimulate thinking
		t245423	Balanced mix of direct instruction and discussion
		t245424	Instructors offer opportunity for discussion
Knowledge transmission		t245441	Teaching is mainly lecturing
		t245442	Lecturers are active, students are passive
		t245443	Mostly teacher-centered teaching
Orientation	Research orientation	t246401	Research-related teaching
		t246402	Lecturers talk about issues of current research
		t246403	Instructors introduce the application of research methods
		t246404	Promotion of ability to do own research
	Practice orientation	t246411	High practice orientation
		t246412	Promotion of professional competencies
		t246413	Close link between theory and practice
	Interdisciplinarity	t246421	Links with other disciplines are established
		t246422	Promotion of cross-curricular knowledge
		t246423	Course topics are addressed from different disciplinary views

Note. The exact wording of the items can be found on the internet (<https://www.neps-data.de/en-us/datacenter/overview/wandassistance/nepsplorer.aspx>).

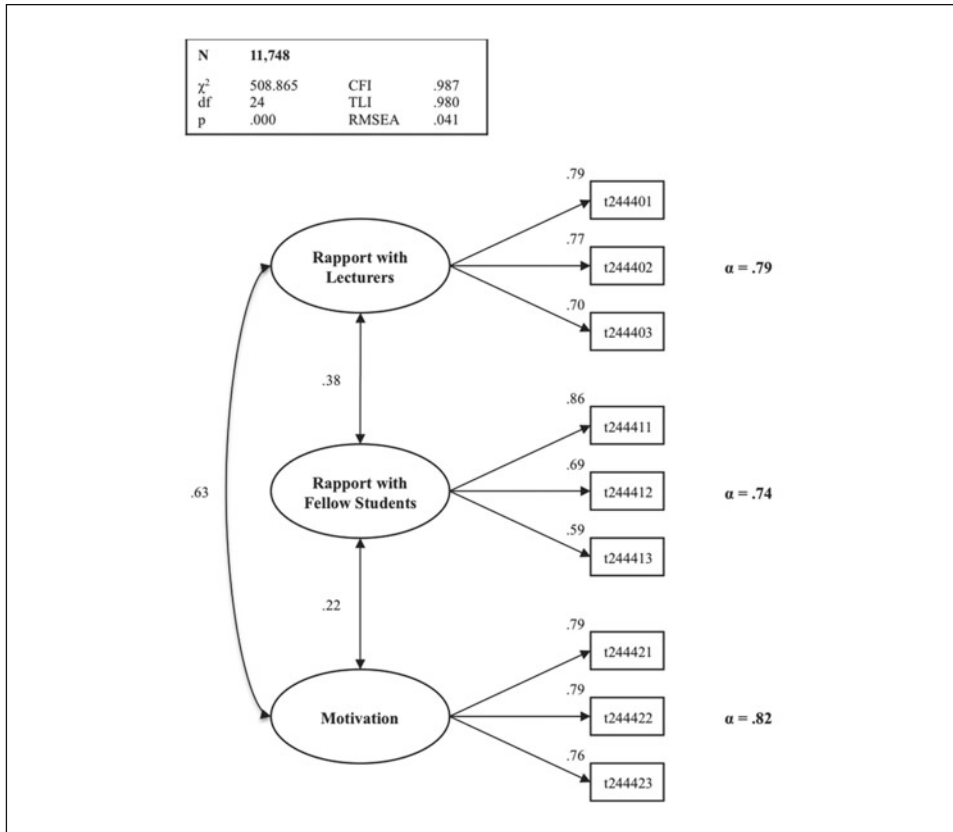
Figure 2 Results of the confirmatory factor analysis of the structural dimension of the SSCO model



In light of the small number of items per subscale, the internal consistency is high ($\alpha \geq .74$), and most factor loadings exceed the value of .70. The low Root Mean Square Error of Approximation (RMSEA) value ($< .05$) indicates that the model fits the data very well. As expected, the subscales that refer to the teaching staff, that is, “rapport with lecturers” and “motivation,” correlate considerably with each other ($r = .63$), while the correlations between the subscale that measures the relations between students (“rapport with fellow students”) and the teacher-related constructs are weak ($r = .38$ and $r = .22$).

As displayed in Figure 4, the subscales for measuring the *challenge* dimension are more or less homogeneous, as well. With one exception, Cronbach’s alpha takes on a minimum value of .70. Since Cronbach’s alpha increases with the number of variables, the low consistency coefficient observed for the subscale “reproduction orientation” ($\alpha = .55$) may be attributed to the fact that we only used two items to represent this

Figure 3 Results of the confirmatory factor analysis of the support dimension of the SSCO model



construct. The factor loadings of the indicators are acceptable ($.54 \leq \lambda \leq .89$), even for the factor just mentioned. With two exceptions, the correlations between the latent constructs are low. The exceptions refer to “knowledge construction,” which is highly negatively correlated with “knowledge transmission” and moderately positively correlated with “meaning orientation.” The RMSEA value of .077 indicates an acceptable fit of the model.

An internal consistency analysis of the subscales representing the orientation dimension yielded satisfactory values ($\alpha \geq .71$; see Figure 5). The size of the factor loadings ($.54 \leq \lambda \leq .91$) suggests that the observed variables are suitable indicators for the latent variables. In addition, a second-order factor was identified, which accounts for 97.4 % of the variance in the latent variable “interdisciplinarity” and for 44.1 % and 25.7 % of the variance in the factors “research orientation” and “practice orientation,” respectively. Model fit as measured by RMSEA (.087) is not good, but still acceptable.

Figure 4 Results of the confirmatory factor analysis of the challenge dimension of the SSCO model

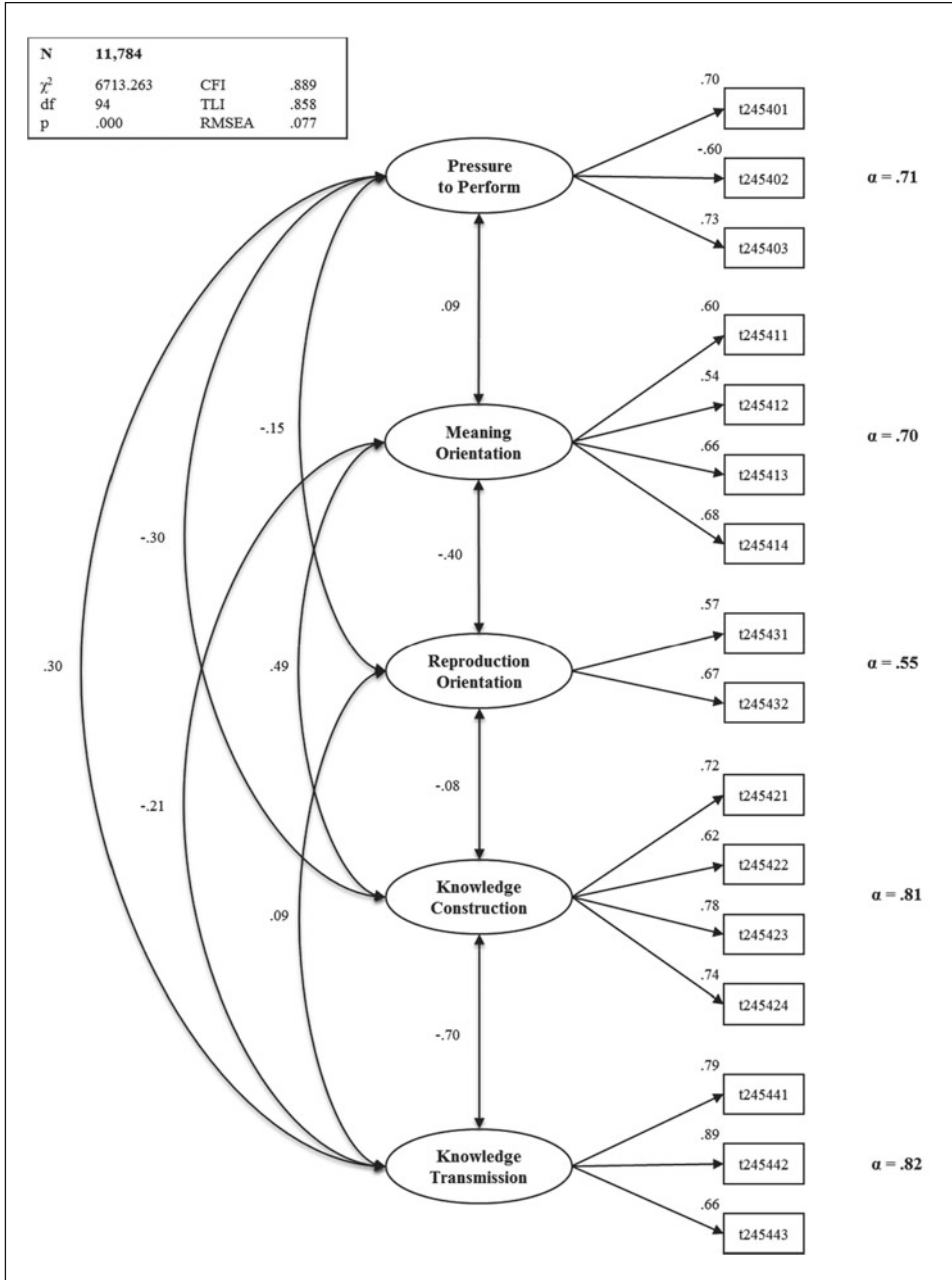


Figure 5 Results of the confirmatory factor analysis of the orientation dimension of the SSCO model



In addition to the analyses conducted separately for each SSCO dimension, we also estimated the complete SSCO model (without cross-loading in the structural dimension) and obtained a satisfactory model fit (RMSEA = .049).

5 Discussion

Our aim was to construct a short, valid, and reliable questionnaire for measuring central dimensions and facets of the formal learning environment in higher education at the program level. By and large, this objective has been achieved, and we consider the quality of the final instrument to be satisfactory. Nonetheless, the findings of the item and scale analyses provide an indication of what can be improved upon and should therefore be discussed.

One result that deserves attention is the cross-loading of variable t243411 on the latent variables “structuredness of lectures and classes” and “coordination of courses offered” (see Figure 2). In our view, there is one main explanation: The other items of the first-mentioned construct explicitly refer to the courses and sessions the respondents attended and to the behavior of the teaching staff in their courses. The other items of the latter scale are clearly directed at the degree program as a whole. The wording of variable t243411 (“The learning objectives of the courses are clearly defined”), however, is ambiguous. It could be related either to teaching behavior in courses and sessions or to the degree program and the question of whether the syllabus, the study guidelines, and module handbooks transparently specify the learning objectives.

The model can be improved upon by removing the variable that loads on the two latent constructs. The fit indices indicate an even better fit (RMSEA = .030), while the internal consistency of the subscale “structuredness of lectures and classes” decreases only slightly ($\alpha = .61$).

Another finding worth mentioning refers to the challenge dimension. According to the outcome of current research on approaches to teaching that reveal that the construct is not unidimensional and bipolar (Braun & Hannover, 2008; Lübeck, 2009), we did not expect high correlations between the corresponding subscales. Since the participants in the NEPS study did not rate teaching behavior, strategies, and practices of individual lecturers, but were instead asked to provide a summarizing and generalized description of all courses attended, the data are not perfectly well suited to decide on the issue of dimensionality. However, the data do not contradict, but rather, support the assumption of unidimensionality for the most part. “Knowledge construction” and “knowledge transmission” are highly negatively correlated, but the correlations between the other latent variables of the challenge dimension are low to modest at best.

The results of our analyses also provide evidence for our assumption that research orientation and practice orientation do not exclude each other, but can go hand in hand. Factor loadings of both constructs on the second-order factor are positive and similar in size. Moreover, in the equivalent model without higher-order factors, the correlation between “research orientation” and “practice orientation” is only moderate ($r = .34$).

Regarding future research, several strands can be considered to be promising. We briefly discuss three of them.

The questionnaire for measuring the SSCO dimensions in higher education focuses on the intermediate level of the degree program. Although we are convinced that this approach opens up the opportunity for answering salient questions in higher education research, the measurement of the learning environment at the level of courses or sessions could reveal more detailed insights into educational processes and their outcomes.

For several pragmatic as well as theoretical reasons, the students themselves are our main source of information on the learning environment. By supplementing this perspective and including the view of the teaching staff and/or observations of classroom practices and interactions, it would be possible to examine the validity of the data sources, to analyze their correlation, to address the question of how teachers' intentions and beliefs are related to observed practices, and to study the links between learning outcomes on the one hand and students' perceptions of the learning environment, teachers' approaches to teaching, and instructional behavior as rated by independent observers on the other hand.

Our questionnaire takes the specific learning environment of students who are enrolled in distance education programs into account insofar as we adapted the wording of items, provided additional explanations, and added items. The instrument, however, does not systematically address new media in higher education that are not only used in distance learning programs, but also increasingly in on-site learning programs. The rise of new media in education (e.g., video teaching, learning platforms, counseling and communication via email) leads to the question of whether they have advantages over conventional forms of teaching and learning as well as whether and under what conditions they impact positively on competence development, motivation, and interest. Case studies have shown that 'heavy' users of virtual learning environments perform better than non-users on the final exam (Stricker, Weibel, & Wissmath, 2011) and that additional e-learning programs succeed in enhancing self-regulated learning, information and communication technology (ICT) competence, working in teams, and subject-specific knowledge (Wagner, Schober, Grading, Reimann, & Spiel, 2010). However, large-scale data on the use and quality of new media in higher education, which include different types of higher education institutions and the entire range of subjects and which make it possible to analyze the effect of new media on competence acquisition and educational decisions, do not exist in Germany.

References

- Anderson, L., & Krathwohl, D. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. New York: Longman.
- Aschinger, F., Epstein, H., Müller, S., Schaeper, H., Vöttiner, A., & Weiß, T. (2011). Higher education and the transition to work. In H.-P. Blossfeld, H.-G. Roßbach, & J. von Maurice (Eds.), *Zeitschrift für Erziehungswissenschaft, 14, Education as a lifelong process: The German National Educational Panel Study (NEPS)* (pp. 267–282). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Aßmann, C., Steinhauer, H. W., Kiesl, H., Koch, S., Schönberger, B., Müller-Kuller, A., ... Blossfeld, H.-P. (2011). Sampling designs of the National Educational Panel Study:

- Challenges and solutions. In H.-P. Blossfeld, H.-G. Roßbach, & J. von Maurice (Eds.), *Zeitschrift für Erziehungswissenschaft*, 14. *Education as a lifelong process: The German National Educational Panel Study (NEPS)* (pp. 51–65). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Bäumer, T., Preis, N., Roßbach, H.-G., Stecher, L., & Klieme, E. (2011). Education processes in life-course-specific learning environments. In H.-P. Blossfeld, H.-G. Roßbach, & J. von Maurice (Eds.), *Zeitschrift für Erziehungswissenschaft*, 14. *Education as a lifelong process: The German National Educational Panel Study (NEPS)* (pp. 67–86). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Blossfeld, H.-P., Roßbach, H.-G., & von Maurice, J. (Eds.). (2011). Education as a lifelong process: The German National Educational Panel Study (NEPS) [Special issue]. *Zeitschrift für Erziehungswissenschaft*, 14. Wiesbaden: VS Verlag für Sozialwissenschaften.
- Bošnjak, M. (2002). *(Non)Response bei Web-Befragungen: Auswahl, Erweiterung und empirische Prüfung eines handlungstheoretischen Modells zur Vorhersage und Erklärung des Partizipationsverhaltens bei Web-basierten Fragebogenuntersuchungen*. Aachen: Shaker.
- Braun, E., & Hannover, B. (2008). Zum Zusammenhang zwischen Lehr-Orientierung und Lehr-Gestaltung von Hochschuldozierenden und subjektivem Kompetenzzuwachs bei Studierenden. In M. A. Meyer, M. Prenzel, & S. Hellekamps (Eds.), *Zeitschrift für Erziehungswissenschaft*, 9. *Perspektiven der Didaktik* (pp. 277–291). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Bronfenbrenner, U. (1979). *The ecology of human development: Experiments by nature and design*. Cambridge, MA: Harvard University Press.
- Bronfenbrenner, U. (1986). Recent advances in research on the ecology of human development. In R. K. Silbereisen, K. Eyferth, & G. Rudinger (Eds.), *Development as action in context: Problem behavior and normal youth development* (pp. 287–309). Berlin: Springer.
- Dippelhofer-Stiem, B. (1983). *Hochschule als Umwelt*. Weinheim: Beltz.
- Dippelhofer-Stiem, B. (1986). How to measure university environment? Methodological implications and some empirical findings. *Higher Education*, 15(5), 475–495.
- Donabedian, A. (1980). *Explorations in quality assessment and monitoring* (Vol. 2). Ann Arbor, MI: Health Administration Press.
- Fend, H. (2008). *Schule gestalten: Systemsteuerung Schulentwicklung und Unterrichtsqualität*. Wiesbaden: VS Verlag für Sozialwissenschaften.
- Hochschulrahmengesetz (HRG 1999) [BGBl. I S. 18]; in der geltenden Fassung von 2007 [BGBl. I S. 506].
- Kember, D. (1997). A reconceptualisation of the research into university academics' conceptions of teaching. *Learning and Instruction*, 7(3), 255–275.
- Klieme, E., Lipowsky, F., Rakoczy, K., & Ratzka, N. (2006): Qualitätsdimensionen und Wirksamkeit von Mathematikunterricht: Theoretische Grundlagen und ausgewählte Ergebnisse des Projekts "Pythagaros". In M. Prenzel, & L. Allolio-Näcke (Eds.), *Unter-*

- suchungen zur Bildungsqualität von Schule: Abschlussbericht des DFG-Schwerpunktprogramms* (pp. 127–146). Münster: Waxmann.
- Klieme, E., & Rakoczy, K. (2003). Unterrichtsqualität aus Schülerperspektive: Kulturspezifische Profile, regionale Unterschiede und Zusammenhänge mit Effekten von Unterricht. In J. Baumert, C. Artelt, Cordula, E. Klieme, M. Neubrand, M. Prenzel, ... M. Weiß (Eds.), *PISA 2000—Ein differenzierter Blick auf die Länder der Bundesrepublik Deutschland* (pp. 333–360). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Klieme, E., & Rakoczy, K. (2008). Empirische Unterrichtsforschung und Fachdidaktik: Outcome-orientierte Messung und Prozessqualität des Unterrichts. *Zeitschrift für Pädagogik*, 64(2), 222–237.
- Lewin, K. (1936). *Principles of topological psychology*. New York: McGraw-Hill.
- Lübeck, D. (2009). *Lehransätze in der Hochschullehre* (Doctoral dissertation). Retrieved from http://www.diss.fu-berlin.de/diss/servlets/MCRFileNodeServlet/FUDISS_derivate_000000005893/01_Dissertationsschrift_DietrunLuebeck.pdf
- Pascarella, E. T., & Terenzini, P. T. (1991). *How college affects students: Findings and insights from twenty years of research*. San Francisco: Jossey-Bass.
- Pascarella, E. T., & Terenzini, P. T. (2005). *How college affects students: A third decade of research* (Vol. 2). San Francisco: Jossey-Bass.
- Radisch, F., Stecher, L., Klieme, E., & Kühnbach, O. (2007). Unterrichts- und Angebotsqualität aus Schülersicht. In H.-G. Holtappels, E. Klieme, T. Rauschenbach, & L. Stecher (Eds.), *Ganztagsschule in Deutschland: Ergebnisse der Ausgangserhebung der "Studie zur Entwicklung von Ganztagsschulen (StEG)"* (pp. 227–260). Weinheim: Juventa.
- Ramm, M., Multrus, F., & Bargel, T. (2011). *Studiensituation und studentische Orientierungen: 11. Studierendensurvey an Universitäten und Fachhochschulen, Langfassung*. Bonn: Bundesministerium für Bildung und Forschung.
- Rousseau, J.-J. (1921). *Emile or education*. London: J. M. Dent & Sons. (Reprinted from *Emile or education*, by J.-J. Rousseau, Ed., 1911, London: J. M. Dent & Sons)
- Schaeper, H., & Wolter, A. (2008). Hochschule und Arbeitsmarkt im Bologna-Prozess: Der Stellenwert von "Employability" und Schlüsselkompetenzen. *Zeitschrift für Erziehungswissenschaft*, 11(4), 607–625.
- Scheerens, J., & Bosker, R. J. (1997). *The foundations of educational effectiveness*. Oxford: Pergamon.
- Stricker, D., Weibel, D., & Wissmath, B. (2011). Efficient learning using a virtual learning environment in a university class. *Computers & Education*, 56(2), 495–504.
- Teichler, U. (2004). The changing debate on internationalisation of higher education. *Higher Education*, 48(1), 5–26.
- Teichler, U. (2011). Der Jargon der Nützlichkeit. Zur Employability-Diskussion im Bologna-Prozess. In B. Hölscher, & J. Suchanek (Eds.), *Wissenschaft und Hochschulbildung im Kontext von Wirtschaft und Medien* (pp. 165–186). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Teichler, U., Buttgerit, M., Baldauf, B., Hermanns, H., Krüger, H., Maiworm, F., ... Winkler, H. (1987). *Hochschule—Studium—Berufsvorstellungen: Eine empirische Untersuchung*

- zur Vielfalt von Hochschulen und deren Auswirkungen (Schriftreihe Studien zu Bildung und Wissenschaft 50). Bonn: Bundesministerium für Bildung und Wissenschaft.
- Teichler, U., & Kehm, B. M. (1995). Towards a new understanding of the relationships between higher education and employment. *European Journal of Education*, 30(2), 115–132.
- Thomas, W.I., & Thomas D.S. (1928). *The child in America: Behavior problems and programs*. New York: Knopf.
- Trigwell, K., & Prosser, M. (2004). Development and use of the approaches to teaching inventory. *Educational Psychology Review*, 16(4), 409–424.
- Trigwell, K., Prosser, M., & Taylor, P. (1994). Qualitative differences in approaches to teaching first year university science. *Higher Education*, 27(1), 75–84.
- Wagner, P., Schober, B., Gradingner, P., Reimann, R., & Spiel, C. (2010). E-Learning unterstützte Förderung von selbstreguliertem Lernen an der Universität. *Zeitschrift für Pädagogische Psychologie*, 24(3-4), 289–303.
- Willis, G. B. (2005). *Cognitive interviewing: A tool for improving questionnaire design*. London: Sage.
- Wosnitza, M. (2007). *Lernumwelt Hochschule und akademisches Lernen*. Landau: Verlag Empirische Pädagogik.

Acknowledgement

This paper uses data from the National Educational Panel Study (NEPS): Starting Cohort 5—First-Year Students, doi:10.5157/NEPS:SC5:4.0.0. From 2008 to 2013, NEPS data were collected as part of the Framework Program for the Promotion of Empirical Educational Research funded by the German Federal Ministry of Education and Research (BMBF). As of 2014, the NEPS survey is carried out by the Leibniz Institute for Educational Trajectories (LIfBi) at the University of Bamberg in cooperation with a nationwide network.

About the authors

H. Schaeper
German Centre for Higher Education Research and Science Studies (DZHW),
Goseriede 9, 30159 Hanover, Germany.
e-mail: schaeper@dzhw.eu

T. Weiß
infas Institut for Applied Social Sciences,
Friedrich-Wilhelm-Str. 18, 53113 Bonn, Germany.

Social Capital, Participation in Adult Education, and Labor Market Success: Constructing a New Instrument

Kerstin Hoenig, Reinhard Pollak, Benjamin Schulz and Volker Stocké

Abstract

The concept of social capital has been extensively used as a predictor for formal educational outcomes and labor market success, whereas its effect on participation in adult education after the end of formal education is a comparatively new application. Different social capital theories assume social networks to be beneficial for individuals in different ways. Within resource-based theories, subjects may profit from information, referrals, and practical support available in networks. In theories based on reference-group processes, significant others are assumed to motivate career mobility and participation in adult education. Most of the available studies on adult education and labor market success are only able to offer data on a single dimension of social capital. Thus, the relative explanatory power of different kinds of social capital is difficult to judge. Furthermore, because of the prevalence of cross-sectional survey designs, the availability of prospective measures of social capital, and consequently, the availability of better opportunities for causal inference is rare. The present chapter describes the development of a social capital instrument for the Starting Cohort 6—Adults of the National Educational Panel Study (NEPS). Our main aim is to provide direct, reliable, and time-efficient measurements for all important dimensions of social capital, combining prospective and retrospective measurement strategies and relying on established and newly developed instruments.

1 Introduction

Social capital has been widely used in educational and labor market sociology to explain competence development, educational attainment, and different returns to human capital (e.g., Brandt, 2006; Coleman, 1988; Croll, 2004; Dufur, Parcel & Trout-

man, 2013; Franzen & Hangartner, 2006; Krug & Rebien, 2012; Moerbeek & Flap, 2008; Morgan & Todd, 2009; Mouw, 2006; Wegener, 1991; Weiss & Klein, 2011; Yakubovich, 2005). Social capital also proves to be an important explanation for ethnic penalties (e.g., Aguilera & Massey, 2003; Kalter & Kogan, 2014; Lancee, 2012) and for disadvantages of women in the labor market (Aguilera, 2008; Smith, 2000). In sharp contrast, the effect of social capital on participation rates in adult education and training remains understudied due to a lack of appropriate data (for an overview, cf. Field, 2005; Strawn, 2003). Social capital is generally understood and conceptualized in very different ways as it mirrors the heterogeneous ideas of Bourdieu (1983), Coleman (1988), Granovetter (1973), Lin (1999), and Sewell, Haller, and Ohlendorf (1970). Furthermore, the field of social capital research is characterized by a lack of consensus about appropriate measures and research methods. There are three main research strategies and respective measures used in social capital research. First, social capital is seen as (potential) access and (potential) utilization of resources in a personal network, typically resulting in a resource generator to capture social capital (van der Gaag & Snijders, 2005). Second, the socioeconomic status composition of the individual's social context is used as a proxy measure for access to social resources, resulting in a position generator strategy (van der Gaag, Snijders & Flap, 2008). Last, social capital might refer to the motivating power of differently ambitious social environments, leading to a focus on normative reference-group influences (Sewell et al., 1970; Singer, 1981).

For the adult sample of the National Educational Panel Study (Allmendinger et al., 2011), researchers of Pillar 3, Pillar 4, and Stage 8 developed a survey instrument whose main objective is to provide data for a better explanation of differential labor market outcomes (both for natives and migrants) and of differential participation in adult education. This instrument integrates the various theoretical approaches and operationalization strategies that exist within the social capital framework. We distinguish clearly between different dimensions of social capital. These alternative but not mutually exclusive mechanisms and operationalizations of social capital are implemented for the same target persons at the same time, providing data that allow for testing these concepts of social capital simultaneously and analyzing their net effect on the different outcomes. Within the framework of the panel design of the National Educational Panel Study (NEPS), we provide prospective measures for the access and usability of different kinds of resources that are beneficial for career mobility and adult learning as well as retrospective measures about the successful use of social capital for reaching these ends.

The instrument was developed in a series of cognitive pretests, in an extensive pre-pilot study, and in a regular pilot study. In the following sections, we focus on the conceptual ideas and on the results of the pre-pilot study. In the final section, we summarize changes in the instrument from the pre-pilot study to the pilot and main study in late 2009.

2 Theoretical Background

Conceptualized in its individual form, social capital may be defined as all possible kinds of resources controlled by social network members that may become available to a focal individual as a result of mutual investments in a shared past that forms the basis of the social relationships within the network (Van der Gaag & Snijders, 2005). Different theories emphasize the importance of social networks for status and educational attainment and implicitly explain the participation in adult education as well. We argue that these can be grouped into two main strands. The first group of theories defines social capital as the *accessibility of different kinds of resources* through social relations (Granovetter, 1973; Lin, 1999). In the case of the social capital theory of Granovetter (1973), information about vacant jobs is regarded as an important resource for labor market success. Weak rather than strong ties are assumed to provide valuable job-related information. Lin (1999) assumes that multiple kinds of resources are important for status attainment.¹ According to this theory, social networks provide information (e. g., about vacant positions), referrals (e. g., formal and informal recommendations), and direct assistance (e. g., support in writing applications).²

Participation in adult education requires a different set of resources, which may be accessible and mobilized through social networks as well. The decision to take courses or to invest in further training can be regarded as a case similar to educational decisions in the school system and is, in principle, subject to the same social capital effects. However, there are important differences as well. First, in adult education, there are no institutionally defined decision points in the life course. Second, the costs and possible benefits of the available decision options vary greatly, from one-day courses during work hours without certification to year-long, privately paid evening classes leading to a formally recognized educational degree. Third, the life situation of adults and school children differ with respect to the control over own resources, resources accessible through social networks, and obligations towards others. Important resources facilitating the participation in adult education include information about the availability, conditions, and utility of different kinds of educational options. Furthermore, the completion of courses often requires substantial financial and time commitments. Being able to mobilize monetary and practical support (e. g., with household or caretaking duties) from others may thus be an important precondition to taking part in adult education.

1 According to Lin, mobilized social capital may provide information, influence, social credentials, and reinforcement (Lin, 2001: 19f.). When developing the survey instruments, we concentrated on information as well as the possibility of influencing relevant actors in the labor market, which is available in social networks.

2 While Granovetter only takes into account determinants of access to resources as a precondition for labor market success, Lin assumes that actors additionally have to motivate their network members to provide the respective resources (Lin, 1999: 473). Unfortunately, the theory remains silent about which factors determine the actors' ability to mobilize accessible resources.

In Lin's resource mobilization theory, inequality in labor market success results from deficits in the endowment with efficient social capital or from unequal returns to this capital (Lin, 2001). In this perspective, labor market disadvantages of women, migrants, and workers of lower social origin are the result of a lower quantity or quality of social capital endowments. Immigrants and women, in particular, are embedded in segregated social networks, which have been proven to be less helpful in the labor market. Thus, an appropriate instrument for measuring social capital should take the gender and ethnic background of social ties to the target person into account.

In a second group of theories, the Wisconsin School being the most prominent, occupational aspirations are assumed to be shaped by *social influence processes* (Sewell et al., 1970). It is argued that significant others' expectations strongly shape an individual's educational and career ambitions. In addition, the aspirations and behavior of social ties may serve as models for the individuals' aspirations. From this perspective, a person possesses more social capital when he or she has relations with significant others who are ambitious themselves and expect ambitious occupational careers from this person. The same argumentation applies for an individual's aspirations to attain adult education. Social capital in this sense consists of social relations that exert normative pressure and shape the target person's motivation to strive for labor market success and participation in adult education.

Aside from the direct effect of significant others' aspirations on the motivation for labor market or educational success, these significant others may also represent sources of emotional support and increase the target person's persistence in achieving these goals (Diewald & Lüdike, 2007). In this case, emotional support is a mediating variable for how the ambitions of the network members influence educational and occupational outcomes. Unfortunately, we were unable to take this perspective or possible negative effects of social capital (Portes, 1998) into account due to constraints in the length of the survey instrument.

3 Development of a Social Capital Instrument

In the literature, a variety of measures have been used to operationalize social capital and analyze its effect on labor market outcomes (for an overview, see Van der Gaag & Webber, 2010). Our instrument is designed to capture the main existing approaches (resource generator, position generator) as well as to add more innovative features (a combination of prospective and retrospective questions in a panel perspective; a combination of motivational and resource perspective). We thus seek to provide a rich dataset for analyzing labor market careers. In the case of participation in adult education, we are not aware of any study using social capital to explain the decision to participate. For the first time, the adult sample of NEPS will provide detailed data on social capital and adult education simultaneously.

3.1 Process of Development

Following an extensive literature review, we conducted a workshop with three experts to establish the main goals and the format of a social-capital and social-networks instrument within the NEPS.³ As a result of this first workshop, several first questionnaire versions underwent intensive cognitive interviewing with selected participants (see Willis, 2005).⁴ Results from these cognitive pretests formed the basis for a second expert workshop, which led to the questionnaire used for the pre-pilot study that is discussed in detail below. Subsequently, we refined and shortened the pre-pilot instrument and tested it again in a pilot study with the entire Starting Cohort 6 questionnaire. Following the pilot study, we performed some additional changes. The final (prospective) instrument was fielded in November 2009, and the retrospective instrument one year later.

3.2 Design of the Pre-Pilot Study

The pre-pilot social capital instrument consisted of three modules. The first module targeted strong ties and reference-group effects. It contained questions about the attitudes and sociodemographics of colleagues, of family members, and—using a Burt generator (Burt, 1984)—of other strong ties. The second module addressed access to resources that facilitate labor market mobility and participation in adult education using a variation of the resource generator, including a detailed account of the resource providers' educational, gender, and ethnic composition. Finally, the third module consisted of a short position generator designed to measure the status and ethnic composition of weak tie networks (Lin, Fu & Hsung, 2001). The pre-pilot study, conducted in April 2009, consisted of a 40-minute Computer Assisted Telephone Interview (CATI) that included the social capital instrument as well as detailed sociodemographic information and questions about participation in the labor market and in adult education. This exclusive focus on social capital sharply distinguishes the pre-pilot from the pilot and main study, which also cover a variety of other topics. Accordingly, the pre-pilot social capital instrument was designed to be much longer than the one we would ultimately be able to include in the main study. Our main goals for the pre-pilot study were:

3 We would like to thank Martin Diewald, Axel Franzen, and Marina Hennig for their extremely valuable comments and suggestions as well as their inspiring discussions at the workshops.

4 With our cognitive interviews, we aimed at identifying problems of question understanding, relevance, and applicability to various subgroups, such as respondents with a migration background, respondents with little formal education, respondents with part time or atypical employment, and self-employed respondents.

- 1) To identify potential problems in the response process, such as question understanding, applicability, and nonresponse.
- 2) To identify technical problems with the instrument, such as filtering and duration.
- 3) To test the quality and reliability of the social capital instrument using descriptive statistics, such as distributional measures, correlations, and item nonresponse as well as reliability and validity measures.⁵

We used a split-ballot design to compare wording options for several items and to test for the optimal number of response categories (i. e., four versus five options) for all reference-group items.⁶ We thereby sought to check whether a middle category, which has the advantage of allowing for more nuanced responses, would attract respondents with a lack of opinion and thus increase measurement error.

All interviewers were asked to keep detailed notes about any problems they recognized during the interview, and several interviews were recorded. We analyzed both of these additional data sources in detail to refine wordings and rule out further problems of comprehension.

Participants were randomly sampled from registers of private household telephone numbers. To compare response patterns of respondents with a migration background and those of native Germans, we oversampled persons whose last names were of Russian or Turkish origin. Our analysis sample consists of 347 subjects (124 male and 223 female) who completed the survey, 36.9 % of whom have a migration background, meaning that they or at least one parent were born abroad.

3.3 Measurement Modules of the Study

Norms, reference groups, and the Burt generator

In the pre-pilot study, we addressed three different reference groups: family members, work contacts (coworkers and supervisors), and other strong ties. The respondents' perceptions about coworkers' and supervisors' attitudes were assessed using four items, three of which concerned the importance of education and training (supervisors expect that respondent attend training; it is common to take courses within firm; training is important to colleagues), and one of which focused on the importance of labor market success (colleagues are ambitious). Additionally, we asked about the composition of the workforce at the respondents' workplace in terms of mi-

5 It is difficult to use construct validity for the assessment of our instrument. In the case of adult education, no reference studies exist. For labor market outcomes, we could only use the intention to achieve success in the career, but here, intentions are highly problematic as indicators for actual labor market success. Thus, we skipped this aspect.

6 In a split-ballot design, the different versions of the question are administered to randomly determined subsamples of the full sample of respondents. Differences in the responses in the subsamples can be attributed to the differences in the wording of the items.

Table 1 Reference-group items

Theoretical construct	Item text	Reference groups	Splits
Definers: labor market success	'How important is it to [person] that you have a career?'/ 'How important is it to [person] that you achieve success professionally? Very important, rather important, [partly], rather unimportant, very unimportant'	Father, mother, sibling, partner, up to three Burt generator names	Split: 4 vs. 5 answer categories Split: question wording
Definers: education and training	'How important is it to [person] that you continually learn new things? Very important, rather important, [partly], rather unimportant, very unimportant'	Father, mother, sibling, partner, up to three Burt generator names	Split: 4 vs. 5 answer categories

gration background, education, and gender. These items performed well with regard to distributions, item nonresponse, and correlations with respondents' own attitudes.

Furthermore, respondents supplied extensive information on their closest family members: father, mother, the sibling they felt closest to, and their current partner. In addition to socio-demographic information, they answered three questions about perceived attitudes towards education and labor market success for each of these family members.⁷ Moreover, we collected respondents' own attitudes on the very same items.

As part of the split-ballot design, the labor market item had two alternative wordings (see Table 1). The 'career' version was closer to our initial theoretical objective, but cognitive interviews raised concerns that the item might not be understood equally across respondents, which is why we introduced the more general 'achieve success' version. A slight trend towards the middle category was discernible for the 'career' item, whereas the 'achieve success' item produced a slightly positively skewed distribution. Other characteristics, such as nonresponse, correlations with respondents' own attitudes, and measures of their labor market success, gave no clear indication as to which wording was preferable. However, interviewer reports and cognitive pretests showed that respondents who worked part-time only or who had low-skilled jobs did not find the first wording option applicable to them as they did not identify with the word 'career.' We did not observe this problem with the phrase 'achieve success,' which we therefore chose for further versions of the instrument.

Regarding the answer category split, we did not find any indication that respondents misused the middle category to report nonattitudes for any of the items except the 'career' wording.

⁷ The third item ('How much would [person] agree with the position that a high level of education is absolutely indispensable?') had a very skewed distribution and is therefore not discussed further.

With the Burt generator, we aimed at collecting detailed information on respondents' close personal ties, especially regarding the nature and strength of their relationships, their education and migration background, and their expectations for the respondents' career mobility as well as education and training. The name generator was part of the split-ballot design. Half of the sample received the original item as introduced by Burt (1984): 'Looking back over the last six months, who are the people with whom you discussed matters important to you?' The other half received a slightly altered version that we suggested in order to be closer to our theoretical objective ('With whom did you discuss future job plans?'). Since we had already collected detailed information on parents, siblings, and partners, these ties were explicitly excluded. Both name-generator versions yielded quite similar results. On average, respondents named 1.8 ties, with 27.4% reporting no names at all. Detailed information was collected on the first three persons named by the respondent. Overwhelmingly, these ties were friends (79.7%), followed by coworkers or supervisors (20.8%) and other family members (7.8%).⁸ As expected, respondents mostly referred to strong ties: 79.9% of the relationships were categorized as 'close' or 'very close.' Respondents in the career-prospects split were slightly more likely to name work contacts (24.1% compared with 17.9%), but they did not differ significantly in terms of tie strength, migration background, education, or their expectations towards ego.

The Burt generator was one of the single most time-consuming parts of the interview, consisting of about three and a half minutes. In light of questionnaire time constraints in the main study, a crucial question was whether information on respondents' close friends might instead be collected in a less time-consuming manner, for example, in a way comparable with the general questions about work contacts as a group. Indeed, respondents' networks were highly homogeneous.⁹ Hence, collecting aggregate information only would presumably save a lot of time without losing much information.

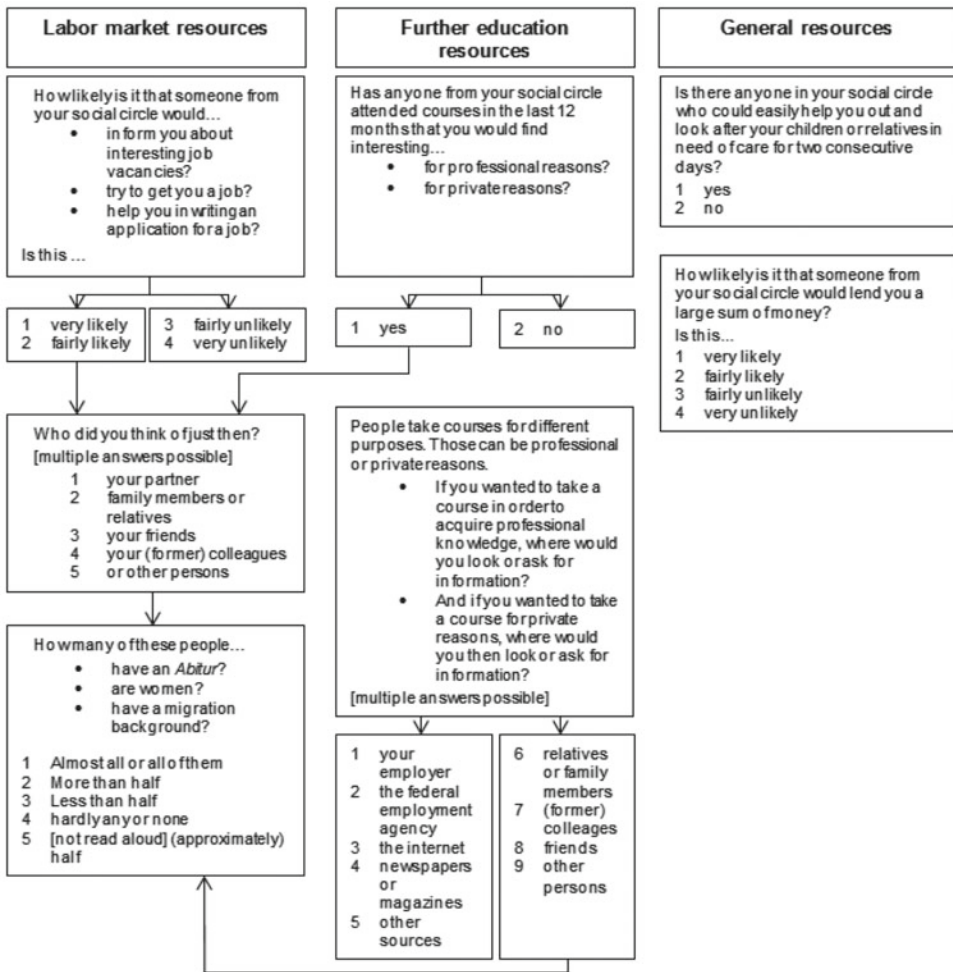
The resource generator

Access to resources that facilitate labor market success and participation in adult education was measured using a variation of the resource generator as proposed by Van der Gaag and Snijders (2005). For labor market success, three resources were included: information about jobs, help with getting a job, and help with job applications. For each of these, respondents were first asked how likely they thought it would be that someone they knew would provide the resource. Respondents who thought this was at least somewhat likely were then asked about the persons about whom they had been thinking and the composition of this group in terms of migration back-

8 Percentages do not add up to 100 because respondents could choose multiple relations for each name.

9 For those respondents who had two or more ties (53.6%), migration background was identical across all ties and across education (Abitur vs. lower) in 80.1% and 69.9% of all cases, respectively. Responses to the first two items in Table 1 were identical in 67.2 and 72.2% of cases, respectively. Correlations between the ties for these items ranged between .55 and .71.

Figure 1 The resource generator



ground, education (Abitur vs. a lower secondary degree), and gender. These follow-up questions provide information about the tie strength, relationship context, and social background of the resource provider, all of which are indicators of the quality of these social resources.

For education and training opportunities, respondents were asked where they would look for information for courses for which they had a) a private or b) a professional interest. The list of sources included both personal contacts, such as friends, colleagues, and family members, and nonpersonal sources, such as the internet, newspapers, their employer, and the federal employment agency. If respondents named at least one personal contact, they were once again asked about the composition of the

group in terms of migration background, education, and gender. In addition, we asked whether participants knew somebody who had participated in a course in which they were interested a) privately or b) professionally within the last twelve months. If they affirmed, they were again asked about the persons about whom they had been thinking and the composition of the group.

In the event that a person named a father, mother, partner, or siblings in the resource-generator questions, respondents became irritated when we asked for these ties' socio-demographic characteristics again. In addition, our pre-pilot study often seemed very repetitive to respondents due to the use of the same questions regarding the composition of a given social environment. To overcome this undesirable situation, we extensively changed the question order and filter structure for the pilot and main study (see Section 4).

Finally, we included two short questions about access to more general resources: the likelihood of somebody's lending a larger sum of money and—for those respondents who were caring for children or older relatives—whether someone could relieve them of this duty for two days.

Using binary logistic regression models, we regressed access to labor market and education resources on migration background, gender, and education (tertiary degree). In line with our expectations, men and respondents with a tertiary degree were significantly more likely to have access to information about jobs and to get a referral. Migrants were as likely as natives to have access to labor market resources, but about a quarter of migrants reported that their network of potential resource providers consisted exclusively or predominantly of other migrants. Migrants were also significantly less likely to know someone who had personal experience with a course taken for professional reasons.

The position generator

To account for the overall status composition of weak-tie networks, we included a short version of the position generator (Lin et al., 2001). Assuming that contacts in higher positions dispose of more resources, the status composition corresponds to the level of resources that an individual can reach through his or her social relations. We therefore expect the position generator to be closely related to the resource generator as described above. However, whereas the resource generator is tailored to provide information on the particular outcomes or job search, career advancement, and participation in adult education, the position generator provides a general measure of clearly defined weak ties. Furthermore, the position generator provides us with an elegant measure of the ethnic composition of these ties. This is very valuable given that our instrument should also be able to differentiate between ethnic and social network composition.

In the position generator, respondents indicated whether they personally knew someone from a list of 13 professions, such as nurses, engineers, sales men, or lawyers, who practice this occupation in Germany. In the pre-pilot study, we did not re-

Table 2 The Position Generator, by Migration Status (probability of endorsement)

	Natives	Migrants
Nurse	0.75	0.66
Engineer	0.74	0.67
Warehouseman	0.42	0.47
Social Worker	0.61	0.55
Salesman/woman	0.65	0.68
Policeman/woman	0.57	0.38
Physician	0.65	0.66
Bank clerk	0.67	0.45
Automotive Mechatronics Technician	0.57	0.52
Lawyer	0.60	0.46
Optician	0.27	0.11
Translator	0.26	0.40
Elementary School Teacher	0.64	0.45
Number of Persons Indicated	7.66	6.90
Mean ISEI	57.23	57.27
Share of Migrants	0.06	0.39
N	219	128

strict the number of persons that respondents could indicate for each profession. For each person in a particular position, we asked for the country of birth of this person.

In the data analysis, we allocated status values to each of the 13 professions using the *International Socio-Economic Index of Occupational Status (ISEI)*. This allows for computing a number of measures describing a respondent's weak-tie network, for example, the mean, maximum, or minimum ISEI of all positions someone indicated. In Table 2, we report descriptive statistics, differentiated according to migrant status, about the probability of knowing somebody from the different professions and the average ISEI as well as the migrant share of all persons.

Overall, migrants tend to know slightly more persons in occupations of lower status, for example, warehousemen or salesmen. When it comes to the mean ISEI, however, these deviations are too small to result in overall differences between migrants and natives. This is because the pattern is not consistently in favor of a higher network composition of natives, who are more likely than migrants to know opticians (27% vs.

11 %), who have a below-average occupational status (ISEI = 48). More often than natives, migrants, by contrast, know translators (40 % vs. 26 %), who have a particularly high occupational status (ISEI = 68). In line with our expectations, we find a much higher share of migrants in migrants' weak-tie networks (39 %) than in those of natives (6 %). According to a t-test ($t = 2.01$, $df = 345$, $p = 0.02$), migrants indicate statistically significantly fewer persons than natives.

Only 18.2 % of all respondents named more than one person in a given profession. Furthermore, excluding multiple nominations per profession does not substantially affect indicators of the total network composition, such as mean ISEI, median ISEI, or its range. Missing values are low throughout the whole position generator. Only seven respondents (2.0 %) did not indicate any person or refused to answer at all. Of those respondents who knew at least one person in any of the 13 positions, 14 respondents (4.1 %) either said that they did not know where this person was born or refused to answer this question. In sum, the position generator seemingly yields an efficient account of the status and the ethnic composition of weak-tie networks.

3.4 Dimensions of Social Capital and Measurement Equivalence

So far, we have analyzed each social capital module separately. Our main argument, however, is that a comprehensive instrument should cover several social capital dimensions. More precisely, we differentiate between i) effects of normative reference groups and ii) effects originating in social resources, such as information or support available in one's social network. Accordingly, we set up a two-dimensional measurement model to test whether or not the reference group and resource modules of our instrument actually represent two distinct latent constructs. To this end, we applied methods of exploratory and confirmatory factor analysis (EFA/CFA; Davidov, Meuleman, Cieciuch, Schmidt & Billiet, 2014; Jöreskog, 1971; Kline, 2011).

We used the former to gain insights into the correlations between variables across modules without specifying any constraints regarding the underlying factorial structure. The results of these exploratory analyses are reported in Tables 3 and 4. Items are ordered by measurement modules, that is, i) social resources, ii) the position generator, and iii) normative reference groups. We furthermore indicate subdimensions for the kind of reference group (job contacts or family) and whether social resources refer to job search or participation in further education.

To estimate the complete factor structure across modules, we included all items in a first EFA. Estimates from principal component analyses with varimax rotation are shown in Table 3. Using standard cut-off criteria (Eigenvalues larger than 1.0) to define the number of factors, the first EFA yields six factors. The first factor contains *family-related reference-group items* as well as one item regarding the expectations of close friends. Further factors cover items on *training related social resources* (Factor 2) and *resources that are relevant for job search* (Factor 3). Three cross loadings between

Table 3 Dimensions of Social Capital, Results from Exploratory Factor Analysis

	<i>F1</i>	<i>F2</i>	<i>F3</i>	<i>F4</i>	<i>F5</i>	<i>F6</i>
<i>Social Resources</i>						
<i>Job Search</i>						
Info vacant positions		0.30	0.62			
Support getting new job			0.81			
Help writing job application		0.56	0.41			
<i>Adult Education</i>						
Info professional courses		0.91				
Info private courses		0.87				
Contacts w/professional courses		0.35	-0.52	-0.47		
Contacts w/private courses			-0.72			
<i>Position Generator</i>						
Status composition (Mean ISEI)				-0.36		0.38
<i>Reference Groups</i>						
<i>Job Contacts</i>						
Colleagues are ambitious					0.79	
Supervisors expect attending training				0.91		
Common to take courses in firm				0.72	0.40	
Training is important for colleagues					0.91	
<i>Family</i>						
Father: importance career	0.86					
Mother: importance career	0.90					
Siblings: importance career	0.75					
Partner: importance career	0.88					
<i>Friends</i>						
Burt: importance career	0.61					0.60
Burt: learn new things						0.88
<i>Eigenvalue</i>	3.44	2.35	2.26	1.87	1.73	1.54
<i>Proportion of Variance Explained</i>	0.19	0.13	0.13	0.10	0.10	0.09
<i>N</i>	53					

Note. NEPS Social Capital Pre-Pilot Study. Factor loadings below .3 are not shown. Results from principal component analysis with varimax rotation; analyses using principal factor- and maximum likelihood methods yield similar results.

these factors indicate that both subdimensions are considerably associated with each other. The four items measuring *normative expectations of job contacts* load on Factors 4 and 5. Whereas we initially assumed that we could capture the achievement climate at the workplace with these four items, respondents seemingly differentiate between colleagues and supervisors. The second and third item of this subdimension, which address supervisors and the firm as a whole, strongly load on Factor 4, while the first and the last items, which refer to colleagues, load on Factor 5. Interestingly, only the fourth and the sixth factor contain items across modules: Both of the Burt generator items load moderately on Factor 6. Besides these items, the mean ISEI generated from the positions generator also loads on this factor, indicating that respondents whose weak tie networks are of higher status also have friends with more pronounced career orientations. The negative cross loading of our status composition measure on Factor 4 suggests that respondents whose weak tie networks are of lower status tend to have supervisors who expect less training and to work in firms in which on-the-job training is less common.

To sum up, the first EFA results support our main argument that we should differentiate between normative reference groups on the one hand and social resources on the other hand because items addressing the former and those measuring the latter consistently load on different factors: While the reference-group items form Factors 1, 4, and 5, the social resource items are attached to Factors 2 and 3.

As several parts of the instrument only apply for certain subgroups—items regarding expectations of colleagues and supervisors, for instance, only apply if respondents are employed—our first EFA was based on a small sample of 53 respondents. As the reference-group module was most often subjected to filtering and splitting, we excluded several parts of it in a second step. The results of these analyses are reported in Table 4. We included all items on social resource but only those reference-group items in which respondents refer to their mother's or their father's expectations. As a result, we were able to use 189 observations (Table 4, Columns 2 to 5). Factors 1 and 2 contain items on job search and training-related *social resources*, while items on parental expectations load on Factor 3. Interestingly, Factor 4 is now in line with our expectation that the position generator gives another account of available social resources and should thus be correlated with the resource module. It shows that respondents whose weak tie networks are of higher status also tend to have more friends who participate in further education. This suggests that the position generator provides us with another measure of resources available through weak ties. Finally, we replaced items on parental expectations by two items regarding the normative expectations of job contacts in a third EFA (Table 4, Columns 6 to 9). This allowed us to analyze 147 cases. Substantively, results are consonant with those of the previous analysis. Most importantly, reference-group items and those capturing social resources again form distinct factors.

To more rigorously test our two-dimensional measurement model, we estimated confirmatory factor analyses using structural equation modeling. As NEPS pays spe-

Table 4 Dimensions of Social Capital, Results from Exploratory Factor Analysis

	<i>EFA 2</i>				<i>EFA 3</i>			
	<i>F1</i>	<i>F2</i>	<i>F3</i>	<i>F4</i>	<i>F1</i>	<i>F2</i>	<i>F3</i>	<i>F4</i>
<i>Social Resources</i>								
<i>Job Search</i>								
Info vacant positions	0.81					0.80		
Support getting new job	0.86					0.87		
Help writing job application	0.52					0.54		
<i>Further Education</i>								
Info professional courses		0.89			0.89			
Info private courses		0.91			0.89			
Contacts w/professional courses				0.69				0.78
Contacts w/private courses				0.75				0.81
<i>Position Generator</i>								
Status composition (Mean ISEI)				0.50	-0.39			0.19
<i>Reference Groups</i>								
<i>Job Contacts</i>								
Colleagues are ambitious								0.84
Training is important for colleagues								0.84
<i>Family</i>								
Father: importance career			0.89					
Mother: importance career			0.88					
<i>Eigenvalue</i>	1.72	1.70	1.57	1.34	1.76	1.75	1.44	1.37
<i>Proportion Of Variance Explained</i>	0.17	0.17	0.16	0.13	0.18	0.17	0.14	0.14
<i>N</i>			189				147	

Note. NEPS Social Capital Pre-Pilot Study. Factor loadings less than 0.3 are not shown. Results from principal component analysis with varimax rotation; analyses using principal factor and maximum likelihood methods yield similar results.

cial attention to respondents of migrant origin and aims at providing data for the analyses of incorporation processes—and as social capital is highly relevant for the socio-economic incorporation of immigrants—another critical aim of our pre-pilot study was to ensure that instruments work for migrants and natives alike. We therefore need to test for intergroup measurement equivalence. As a final step, we accordingly

applied methods of multi-group confirmatory factor analyses (MGCFA), which essentially test whether a certain construct is measured in the same way across different groups (Davidov et al., 2014). If MGCFA also supports our preliminary conclusion that was based on EFA, we would have good reasons to tailor the NEPS social capital instrument in the proposed way.

We first set up a CFA model according to the last EFA model (cf. Table 4), that is, for normative expectations of job contacts (Module 1) and social resources regarding job search (Module 2) as two distinct latent variables. We had to exclude one item on help writing job applications because of 47 missing values. For expectations of job contacts (reference group module), we furthermore excluded the item ‘supervisors expect attending training’ because we found in the EFA that respondents differentiate between colleagues and supervisors. Furthermore, this item has the highest number of missing values. Our final CFA measurement model contains i) social resources regarding the job search (items: information on vacant positions; support getting a new job) and ii) normative expectations of job contacts (items: colleagues are ambitious; in our firm, it is common to take courses; training is important for colleagues).¹⁰

Table 5 shows goodness-of-fit statistics for the overall model (M1) as well as for an MGCFA model differentiating between natives and migrants (M2). For the full sample, that is, ignoring any between-group variance, as well as for M2, a two-factor structure fits the data well. The χ^2 -test shows p-values of about 0.89 (Model 1) and 0.38 (Model 2), meaning that differences between modeled and observed covariances are not statistically significant. Besides exact-fit statistics, approximate-fit indices also suggest that Models 1 and 2 fit the data well (see rows 2–6, Table 5).

Given proper model fits, we could test the equality of factor loadings¹¹ between natives and migrants. We tested the equivalence of factor loadings and intercepts. Table 6 shows standardized coefficients across groups (Model 1) as well as for natives and migrants separately (Model 2). Overall, differences between natives and migrants were small. We designed Model 2 to constrain all coefficients and intercepts to be equal across groups. Only for the item ‘training is important for colleagues’

10 A fundamental assumption of maximum likelihood structural equation modeling is that data are multivariate normally distributed; otherwise, estimates could be biased (see e.g., Byrne & Van de Vijver, 2010: 116). To relax this assumption, we additionally fitted Model 2 using a distribution free method of moments, that is, Stata’s ADF-method. We thereby also checked whether the missing-at-random assumption holds, which we accepted using missing value replacement, that is, Stata’s *mlmv* method. We also cross-checked our results using robust standard errors to account for the survey design of our data, particularly the over-sampling procedures. Both robustness checks (ADF and robust SE) are widely in line with the standard ML model presented. We found no substantial differences.

11 More precisely, we tested for *scalar invariance*, which means that not only factor structures (configural invariance) and item loadings (metric invariance) are invariant between both groups but also that intercepts are virtually the same. Scalar invariance essentially means that mean differences in observed variables translate into respective differences in latent variables. As Davidov et al. (2014) nicely summarized: “scalar equivalence implies that the measurement scales not only have the same intervals but also share origins” (ibid., 64).

Table 5 Goodness-of-Fit Statistics, Based on Confirmatory Factor Analysis

	Model 1	Model 2
χ^2	1.14	14.97
Df	4	14
SRMR	0.01	n/a
RMSEA (Upper Bound)	0.05	0.08
CD	0.99	0.96

Note. Pre-pilot data, own calculations. SRMR = Standardized Root Mean Square Residual, RMSEA = Root mean Square Error of Approximation, CD = Coefficient of Determination. N(M1) = 178, N(M2) = 329. Since the number of migrants (N = 59) would otherwise be too small to calculate stable estimates, missing value replacement applied fitting Model 2.

Table 6 Between-Group Measurement Equivalence, Standardized Coefficients from Confirmatory Factor Analysis

	Model 1	Model 2		
	Across Groups	Natives	Migrants	Δ
<i>Factor Loadings</i>				
<i>Reference Group: Job Contacts (LV)</i>				
Colleagues are ambitious	0.49	0.53	0.49	0.03
Common to take courses in firm	0.46	0.47	0.45	0.02
Training is important for colleagues	0.97	1.00	0.86	0.14
<i>Job Resources (LV)</i>				
Info vacant positions	0.73	0.83	0.78	0.04
Support getting new job	0.79	0.70	0.66	0.03
<i>Intercepts</i>				
Colleagues are ambitious	1.38	1.32	1.21	0.11
Common to take courses in firm	0.92	0.89	0.84	0.05
Training is important for colleagues	1.01	0.98	0.83	0.15
Info vacant positions	2.21	2.09	1.99	0.10
Support getting new job	2.38	2.20	2.10	0.10
N	178	210	119	
Coefficient of Determination	0.98	0.99	0.93	

Note. Pre-pilot data, own calculations. Since the number of migrants would otherwise be too small (N = 59) to calculate stable estimates, missing value replacement applied fitting Model 2.

does this not hold, as revealed by a Lagrange multiplier test ($\chi^2(\text{coefficient}) = 5.09$, $\chi^2(\text{intercept}) = 3.24$, $df = 1$). However, this single difference between natives and immigrants in one item is too small to compromise the overall model.

Altogether, our CFA and MGCFA results clearly support the hypothesis that normative reference groups and social resources concerning the job search form distinct social capital dimensions. We could furthermore establish measurement equivalence between natives and migrants. The proposed two-dimensional social capital measurement model thus applies for both native German respondents and those of immigrant origin alike.

4 Consequences of the Pre-Pilot

While our analyses confirmed our initial theoretical conceptions and general framework, the pre-pilot study also pointed to some weaknesses in the social capital instrument. Consequently, we rigorously shortened and revised the instrument for inclusion in the pilot study. For the reference-group module, the main problem was the interview length. Consequently, we decided not to include the third reference-group item, which had a very positively skewed distribution. Furthermore, we dropped all items regarding siblings due to time constraints and high correlations among different family members. Concerning the response scale split, we opted for the one with five response categories because comparisons of the two splits showed that these yielded more nuanced distributions without strong evidence of a trend towards the middle category.

One item was newly included in the questionnaire: Since some respondents insisted that they did not care about the opinion of certain family members, we asked how important each person's opinion was to the respondent. Finally, we made some changes to the questions about supervisors and coworkers to make them more comparable with other reference groups.

The resource generator was revised concerning filtering, length, and changes in question wording. First, we introduced a very sophisticated filtering system for follow-up questions that avoids collecting redundant information. For instance, respondents who answer that their main resource provider is their partner will not be asked about this person's demographic information, as it is collected in other parts of the interview. Second, we revised the introduction of the generator questions, expanded the number of response options, and made some further changes to the question wording of individual items.

The position generator saw the fewest changes. We improved filters for respondents who know multiple people in an occupation, changed the order of occupations, and extended the list of countries of origin.

Finally, we divided the social capital instrument into thematic submodules. These submodules were spread over the entire 90-minute interview so that social capital

questions were integrated with other questionnaire items covering the same topic. For instance, reference group questions regarding coworkers were asked in the context of other questions about the respondents' occupational history, and resource-generator questions regarding education and training followed respondents' accounts of the training measures in which they had participated. This made the interview more engaging and less repetitive for respondents.

The revised instrument was tested in a pilot study with 197 respondents. The main goal of the pilot study was to test the administration and length of the complete NEPS adult cohort survey instrument for Wave 1. After the pilot study, further cuts had to be made due to time constraints. The Burt generator was severely shortened, and several of the follow-up questions were replaced by generalized items about respondents' close friends as a group since both the pre-pilot and pilot data confirmed that the information for different Burt ties was highly homogeneous. Within the resource generator, information about gender composition was cut for all resources. For help with job applications and information about training, we also had to cut ethnic and educational composition. Moreover, we made a few changes in question wording, the most significant of which was a change to all questions concerning the educational composition of networks, where we switched from asking about the alteri's highest secondary degree to tertiary degrees as many pilot respondents were unsure of their alteri's secondary degrees.

The final instrument of the main study as well as further information about the Scientific Use File of the NEPS adult cohort is available at <https://www.neps-data.de/en-us/datacenter/dataanddocumentation/startingcohortadults.aspx>. The complete questionnaire of the pre-pilot study is available from the authors upon request.

5 Conclusion

The social capital instrument of NEPS Starting Cohort 6 (Adults) aims to explain labor market success and adult education through processes of interpersonal influence, support, and resource transmission. It integrates different theoretical traditions and perspectives as well as different operationalizations, such as a resource generator and a position generator.

The pre-pilot played a crucial role in the development process. Its data were the main basis for cuts, technical refinements, and final changes to question wording and response options. Furthermore, exploratory and confirmatory factor analyses confirm our main theoretical assumption that social capital is not a homogeneous construct but instead works through several theoretical channels. Resources and reference groups form distinct dimensions, whereas the position generator is related to resources and also gives a comprehensive account of networks' ethnic composition. Measurement equivalence exists between natives and migrants.

The prospective social capital instrument discussed in this chapter, which forms

the basis of the 2009 panel wave instrument, is matched by a retrospective instrument in the following year. This allows for a direct comparison between potentially available network resources and the actual use of social capital. In conclusion, the NEPS offers comprehensive, innovative, and reliable data for explicit theory tests in the field of social capital and educational and labor market outcomes.

References

- Aguilera, M. B., & Massey, D. S. (2003). Social capital and the wages of Mexican migrants: New hypotheses and tests. *Social Forces*, 82(2), 671–701. doi:10.1353/sof.2004.0001
- Aguilera, M. B. (2008). Personal networks and the incomes of men and women in the United States: Do personal networks provide higher returns for men or women? *Research in Social Stratification and Mobility*, 26(3), 221–233. doi: 10.1016/j.rssm.2008.05.003
- Allmendinger, J., Kleinert, C., Antoni, M., Christoph, B., Drasch, K., Janik, F., ... Ruland, M. (2011). Adult education and lifelong learning. In H.-P. Blossfeld, H.-G. Roßbach, & J. von Maurice (Eds.), *Zeitschrift für Erziehungswissenschaft*, 14. *Education as a lifelong process: The German National Educational Panel Study (NEPS)* (pp. 283–299). Wiesbaden: VS Verlag für Sozialwissenschaften. doi:10.1007/s11618-011-0197-0
- Bourdieu, P. (1983). Ökonomisches Kapital, kulturelles Kapital, soziales Kapital. In R. Kreckel (Ed.), *Soziale Welt, special issue 2. Soziale Ungleichheiten* (pp. 183–198). Göttingen: Schwartz & Co.
- Brandt, M. (2006). Soziale Kontakte als Weg aus der Erwerbslosigkeit. *Kölner Zeitschrift für Soziologie und Sozialpsychologie*, 58(3), 468–488. doi: 10.1007/s11575-006-0106-6
- Burt, R. S. (1984). Network items and the general social survey. *Social Networks*, 6(4), 293–339. doi: 10.1016/0378-8733(84)90007-8
- Byrne, B. M., & Van de Vijver, F. (2010). Testing for measurement and structural equivalence in large-scale cross-cultural studies: Addressing the issue of nonequivalence. *International Journal of Testing*, 10(2), 107–132. doi:10.1080/15305051003637306
- Coleman, J. S. (1988). Social capital in the creation of human capital. *American Journal of Sociology*, 94, 95–120. doi:10.1086/228943
- Croll, P. (2004). Families, social capital and educational outcomes. *British Journal of Educational Studies*, 52(4), 390–416. doi:10.1111/j.1467-8527.2004.00275.x
- Davidov, E., Meuleman, B., Cieciuch, J., Schmidt, P., & Billiet, J. (2014). Measurement equivalence in cross-national research. *Annual Review of Sociology*, 40, 55–75. doi: 10.1146/annurev-soc-071913-043137
- Diewald, M., & Lüdicke, J. (2007). Akzentuierung oder Kompensation? Zum Zusammenhang von sozialer Ungleichheit, Sozialkapital und subjektiver Lebensqualität. In J. Lüdicke, & M. Diewald (Eds.), *Soziale Netzwerke und soziale Ungleichheit. Zur Rolle von Sozialkapital in modernen Gesellschaften* (pp. 11–52). Wiesbaden: VS Verlag für Sozialwissenschaften.

- Dufur, M. J., Parcel, T. L., & Troutman, K. P. (2013). Does capital at home matter more than capital at school? Social capital effects on academic achievement. *Research in Social Stratification and Mobility*, 31(1), 1–21. doi:10.1016/j.rssm.2012.08.002
- Field, J. (2005). *Social capital and lifelong learning*. Bristol: Policy Press.
- Franzen, A., & Hangartner, D. (2006). Social networks and labour market outcomes: The non-monetary benefits of social capital. *European Sociological Review*, 22(4), 353–368. doi:10.1093/esr/jcl001
- Granovetter, M. S. (1973). The strength of weak ties. *American Journal of Sociology*, 78(6), 1360–1380. doi:10.1086/225469
- Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika*, 36(4), 409–426. doi:10.1007/BF02291366
- Kalter, F., & Kogan, I. (2014). Migrant networks and labor market integration of immigrants from the former Soviet Union in Germany. *Social Forces*, 92(4), 1435–1456. doi:10.1093/sf/sot155
- Kline, R. B. (2011). *Principles and practice of structural equation modeling*. New York City: The Guilford Press.
- Krug, G., & Rebien, M. (2012). Network-based job search: An analysis of monetary and non-monetary labor market outcomes for the low-status unemployed. *Zeitschrift für Soziologie*, 41(4), 316–333.
- Lancee, B. (2012). The economic returns of bonding and bridging social capital for immigrant men in Germany. *Ethnic and Racial Studies*, 34(4), 664–683. doi:10.1080/01419870.2011.591405
- Lin, N. (1999). Social networks and status attainment. *Annual Review of Sociology*, 25, 467–487. doi:10.1146/annurev.soc.25.1.467
- Lin, N. (2001). *Social Capital: A Theory of Social Structure and Action*. Cambridge: Cambridge University Press.
- Lin, N., Fu, Y. C., & Hsung, R. M. (2001). The position generator: Measurement techniques for investigations of social capital. In N. Lin, K. Cook, & R. S. Burt (Eds.), *Social capital: Theory and research* (pp. 57–81). New York: Aldine de Gruyter.
- Moerbeek, H., & Flap, H. (2008). Social resources and their effect on occupational attainment through the life course. In N. Lin, & B. H. Erickson (Eds.), *Social capital: An international research program* (pp. 133–156). Oxford: Oxford University Press.
- Morgan, S. L., & Todd, J. L. (2009). Intergenerational closure and academic achievement in high school: A new evaluation of Coleman's conjecture. *Sociology of Education*, 82(3), 267–286. doi:10.1177/003804070908200304
- Mouw, T. (2006). Estimating the causal effects of social capital: A review of recent research. *Annual Review of Sociology*, 32, 79–102. doi:10.1146/annurev.soc.32.061604.123150
- Portes, A. (1998). Social capital: Its origins and applications in modern sociology. *Annual Review of Sociology*, 24, 1–24. doi:10.1146/annurev.soc.24.1.1
- Sewell, W. H., Haller A. O., & Ohlendorf, G. W. (1970). The educational and early occupational status attainment process: Replication and revision. *American Sociological Review*, 35(6), 1014–1027. doi:10.2307/2093379

- Singer, E. (1981). Reference groups and social evaluations. In M. Rosenberg, & R.H. Turner (Eds.), *Sociological perspectives* (pp. 66–93). New York: Basic Books.
- Smith, S.S. (2000). Mobilizing social resources: Race, ethnic, and gender differences in social capital and persisting wage inequalities. *The Sociological Quarterly*, 41(4), 509–537. doi:10.1111/j.1533-8525.2000.tb00071.x
- Strawn, C. L. (2003). *The influence of social capital on lifelong learning among adults who did not finish high school*. Cambridge: NCSALL.
- Van der Gaag, M., & Snijders, T.A.B. (2005). The resource generator: Social capital quantification with concrete items. *Social Networks*, 27(1), 1–29. doi:10.1016/j.socnet.2004.10.001
- Van der Gaag, M., Snijders, T. A. B., & Flap, H. (2008). Position generator measures and their relationship to other social capital measures. In N. Lin, & B.H. Erickson (Eds.), *Social capital: An international research program* (pp. 27–48). Oxford: Oxford University Press.
- Van der Gaag, M., & Webber, M. (2010). Measurement of individual social capital: Questions, instruments, and measures. In I. Kawachi, S. V. Subramanian, & D. Kim (Eds.), *Social capital and health* (pp. 29–50). New York: Springer.
- Wegener, B. (1991). Job mobility and social ties: Social resources, prior job, and status attainment. *American Sociological Review*, 56(1), 60–71. doi:10.2307/2095673
- Weiss, F., & Klein, M. (2011). Soziale Netzwerke und Jobfindung von Hochschulabsolventen: Die Bedeutung des Netzwerktyps für monetäre Arbeitsmarkterträge und Ausbildungsadäquatheit. *Zeitschrift für Soziologie*, 40(3), 228–245.
- Willis, G.B. (2005). *Cognitive interviewing: A tool for improving questionnaire design*. Thousand Oaks: SAGE Publications.
- Yakubovich, V. (2005). Weak ties, information, and influence: How workers find jobs in a local Russian labor market. *American Sociological Review*, 70(3), 408–421. doi:10.1177/000312240507000303

About the authors

K. Hoenig
Leibniz Institute for Educational Trajectories (LifBi)
e-mail: kerstin.hoenig@lifbi.de

R. Pollak
Berlin Social Science Center (WZB)

B. Schulz
Chair of General Sociology, University of Mannheim
Berlin Social Science Center (WZB)

V. Stocké
Chair of Methods of Empirical Social Research,
University of Kassel

Measuring Students' Social and Academic Integration—Assessment of the Operationalization in the National Educational Panel Study

Gunther Dahm, Oliver Lauterbach and Sophie Hahn

Abstract

Dropping out of higher education is a prevalent phenomenon in Germany—about every fourth college student does not graduate—that affects educational returns to a considerable degree. Therefore, dropouts are a topic of major interest in the higher education stage of the NEPS. With the NEPS data, it is possible to study dropouts from higher education with large-scale, nationwide, representative data from a longitudinal perspective. In order to better understand the mechanisms of dropout, the NEPS provides researchers with the opportunity to analyze the role of social and academic integration (Tinto 1975, 1993)—in addition to rational choice-based measures—in the dropout process. Despite the prevalence of the integration concept in the Anglo-Saxon literature, only a few attempts have been undertaken to operationalize and apply social and academic integration to the German context. NEPS Stage 7 tries to close this gap by reassembling and testing several instruments that are well-established in Germany and can be considered to adequately measure social and academic integration. Analyses of factorial and criterion-related validity show that the NEPS provides a parsimonious measure of relevant aspects of students' integration.

1 Introduction

Decisions to begin or leave an educational program—for example, to attend a particular school or to withdraw from college—structure students' educational and occupational careers and have substantial effects on their educational attainment and thereby on their opportunities in life (Blau & Duncan, 1967; Boudon, 1974; Breen & Jonsson, 2000; Maaz, Hausen, McElvany, & Baumert, 2006; Mare, 1981). Therefore,

one of the main research interests of the German National Educational Panel Study (NEPS) is to explain educational decisions at different stages in the life course (Stocké, Blossfeld, Hoenig, & Sixt, 2011). Regarding Starting Cohort 5—First-Year Students, leaving higher education before graduation can be viewed as one of the most relevant educational decisions in this stage of life (Aschinger et al., 2011). Given the importance of educational attainment for opportunities in life as well as political efforts to increase participation in higher education in Germany (e.g., Powell & Solga, 2011), high dropout rates can be considered a serious problem both for the individual and for society as a whole.

In Germany, 28 % of a cohort of first-year students studying in a bachelor program drop out of higher education (Heublein, Richter, Schmelzer, & Sommer, 2014). However, the “mechanisms of dropout still remain to be studied with nationwide representative longitudinal data” (Aschinger et al., 2011, p. 276). Various theoretical perspectives exist that account for student persistence or dropout. These perspectives derive from different scientific disciplines, such as psychology, economics, educational science, and sociology. Various paradigms and theories can be distinguished further within these broad theoretical approaches (e.g., Heublein & Wolter, 2011; Robbins, Lauver, Davis, Langley, & Carlstrom, 2004; Sarcletti & Müller, 2011). To account for educational decisions, the NEPS-wide focus lies upon rational-choice-based explanations. In addition to these measures, Stage 7 of the NEPS allows researchers to assess the role of social and academic integration in the dropout process. The idea that students’ integration in the social and academic systems of higher-education institutions is relevant to their persistence was first introduced by Spady (1970) and further developed and refined by Tinto (1975, 1993). The concept of integration has been highly prevalent in the Anglo-Saxon debate on dropout in higher education since the 1970s. In the context of Germany’s higher-education system, however, there is still a lack of instruments that parsimoniously operationalize social and academic integration. Stage 7—From Higher Education to the Labor Market—tries to close this gap by reassembling and testing several instruments that are well-established in Germany and can be considered to measure relevant aspects of students’ integration. In this paper, the quality of these measures is discussed in terms of factorial structure and criterion validity.

2 Tinto’s Model of Student Departure

Research on dropout from higher education has been strongly influenced by the theory of Vincent Tinto, which has reached a “near paradigmatic status” (Braxton, Milem, & Sullivan, 2000, p. 569). According to Tinto, dropout from higher education can be regarded as “the outcome of a longitudinal process of interactions between the individual and the [higher education] institution” (1975, p. 103). Tinto’s model of the dropout process is based on Durkheim’s theory of suicide. Durkheim argues that

committing suicide stems from low integration in the moral system and having only few interactions with others (Durkheim, 1961). Likewise, Tinto considers the degree of congruency with the standards, objectives, and values of the community at college and interactions with peers, faculty, and administrative staff to be major determinants of dropout. Both aspects are described as academic and social integration. Academic integration is a bilateral process during which the individual is evaluated by the system (e.g., by earning grades) while the system is simultaneously evaluated by the individual, which results in an adaptation to and identification with norms of the academic system (intellectual development; Tinto, 1975, p. 104). Furthermore, social integration evolves via the congruency between the individual and the social environment and also via interactions with peers, faculty, and administrative personnel within the higher-education institution (Tinto, 1975, p. 107). Social and academic integration are usually interrelated but may also develop independently: Academic discussions with peers at a university may increase academic performance, and attending class regularly improves contact with peers. However, students can also be socially well integrated but perform poorly, and vice versa (Tinto, 1975, p. 92). Finally, social and academic integration shape students' institutional and goal commitments, which in turn affect their persistence or dropout behavior.

Tinto's concept of social and academic integration has been criticized for theoretical ambiguities and missing empirical support for some of its propositions (Braxton, Milem, & Sullivan 2000; Neuville et al., 2007), although the latter may partially result from secondary analyses of data collected for purposes other than validation. Some authors argue that the aspect of academic integration, in particular, needs revision (Braxton & Lien, 2000), and that the integration concept in general neglects several influencing factors, such as college climate and the experiences of social minorities (Baird, 2000; Tierney, 1992). Another criticism is that Tinto's approach is sociological in nature, whereas student departure as an individual decision may be better explained by psychological theories such as motivation theory, attitude-behavior theory, or self-efficacy (Bean & Eaton, 2000; Robbins et al., 2004). Notwithstanding these possible limitations, Tinto's concept "is still the bar by which other models are measured" for the prediction of academic persistence (Tillman, 2002, p. 5).

3 Operationalization of the Tinto Model

In the U.S., Tinto's approach to explain student departure has been prominent for decades (Borglum & Kubala, 2000; Halpin, 1990; Napoli & Wortman, 1998). However, despite the prevalence of the integration concept, "there is not a widely accepted metric for either academic or social integration" (Davidson, Beck, & Milligan, 2009, p. 375). Several instruments exist that measure notions of students' integration in college, such as the Institutional Integration Scale by Pascarella and Terenzini (1980; see also French & Oakes, 2004), the College Persistence Questionnaire by Davidson

et al. (2009), and the Student Adaption to College Questionnaire (SACQ) developed by Baker and Siryk (1984, 1999). In Germany, Tinto's model has only been applied in a few empirical studies, either at a single institution or mostly in small samples of higher-education institutions (Gold, 1988; Henecka & Gesk, 1996; Winteler, 1984). Based on the SACQ, Leichsenring, Sippel, and Hachmeister (2011) developed a questionnaire to measure students' adaption to the demands of academic studies in the context of German higher education. The SACQ also served as a starting point to measure students' academic and social integration for NEPS Stage 7.

3.1 Developmental Study

In search of a measure of social and academic integration, NEPS Stage 7 initially intended to use the SACQ because it "closely parallel[ed] Tinto's model of institutional departure" (Krotseng, 1992, p. 101), had been successfully tested in the context of a European higher-education system (Beyers & Goossens, 2002), and was available as a translated German version (Sippel, 2006). Since the full SACQ was too long for an application in the NEPS, a shortened version containing 21 items and five subscales was tested in a developmental study (Müller & Sarcletti, 2010).

Results based on the data of an online survey of 788 students showed that the factorial structure of the subscales could not be replicated. A five-factorial model consisting of the subscales of "motivation," "effort," "achievement," "general social integration," and "contact to others" indicated an insufficient model fit ($\chi^2 = 927.179$, $df = 142$, CFI = .742, TLI = .689, RMSEA = .089). In principal component analyses, several of the items did not load on the corresponding subdimensions of the SACQ. Furthermore, reliabilities and correlations with other measures, such as academic satisfaction, academic commitment, and dropout intentions, were not consistent. Therefore, alternative instruments that were originally included in the developmental study for validation purposes were selected to measure social and academic integration in the NEPS pilot study and in the main survey.

3.2 Operationalization in the NEPS

Because of the reciprocal relationship of commitment and integration in Tinto's model and the difficulty of disentangling these processes by a parsimonious instrument, aspects of students' academic commitment were chosen to measure both academic integration and the resulting behavioral intentions regarding studies. From an organizational perspective, different facets of organizational commitment are generally used to predict prosocial behavior or employee-turnover in work environments (Cohen, 2003; Meyer & Allen, 1997), but these facets are also suggested to be relevant predictors for academic attainment and dropout (Bean, 1980). From the per-

spective of motivation theory, commitment describes an individual's attachment to a desirable goal of action (Brunstein, 1995) and may therefore also be used to predict academic attainment. Consequently, drawing on aspects of academic commitment as indicators of students' integration may be pragmatic, but nevertheless appropriate, because commitment refers to the adaption of norms and values (identification or normative commitment; O'Reilly & Chatman, 1986) on the one hand and to more specific behavioral intentions (achievement motivation or goal commitment) on the other hand. In this respect, commitment regarding studies appears to correspond to what Tinto describes as intellectual development, the normative facet of academic integration.

Therefore, the degree of a student's integration in the academic sphere was measured by the *Academic Commitment* scale (Grässmann, Schultheiss, & Brunstein, 1998). The original instrument covers five facets: identification with academic studies, determination to complete studies, willingness to invest effort, pursuance of high aspiration levels, and affective involvement. Since the facet "determination" is regarded as a dependent variable and is covered by a separate dropout scale, the respective items were excluded. Grässmann et al. did not distinguish subscales; however, empirical analyses in the developmental study revealed two discriminable factors, namely *affective involvement* and *achievement orientation*.

To cover the aspect of performance as another indicator of academic integration, one could either refer to the relative performance compared with fellow students or to the fulfilment of self-set standards of achievement. Since the main study was scheduled to be administered in the first year of studies, it might have been too difficult for participants to evaluate their performance in comparison with others. As a result, perceived academic performance was measured by the *Fulfilment of Achievement Expectations* scale (Trautwein et al., 2007).

The social aspect of integration was measured by the *Social Integration Scale* by Schiefele, Moschner, and Husstegge (2002). Since there is no established German scale to measure interactions with faculty, four additional items of different origin were assembled in order to cover this aspect. All four items focus not on quantitative, but rather on qualitative aspects of student-faculty interactions, and three place special emphasis on aspects of social climate at the respective higher-education institution. Due to institutional and cultural differences between the German and the U.S. higher-education system, items referring to off-campus interactions with faculty members were left out.

4 Method

Since the scales measuring social and academic integration were not completely administered in the developmental study or in the NEPS pilot study, the following analyses use data from Starting Cohort 5 that originated from the first online wave of the

main study.¹ Because of limited survey time in the online questionnaire, most of the scales had to be shortened based on reliability tests of the original versions in the NEPS pilot study (for the final measures, see Appendix A).

4.1 Measures

The normative aspect of academic integration was measured by six items of the Academic Commitment Scale (Grässmann et al., 1998). Two subdimensions, each consisting of three items, differentiate between *affective involvement* and *achievement orientation* (Cronbach's $\alpha = .84$ and $.72$). The items have five response alternatives² ranging from 1 (*not true at all*) to 5 (*absolutely true*), including a neutral category.

The performance aspect of academic integration was measured by three items from the *Fulfilment of Achievement Expectations* scale (Trautwein et al., 2007) that describe whether students' achievement expectations have been realized (Cronbach's $\alpha = .81$). The items have four response alternatives ranging from 1 (*not true at all*) to 4 (*absolutely true*).

For social integration, three items from Schiefele et al. (2002) were chosen to cover interactions with fellow students (Cronbach's $\alpha = .84$). Interactions with faculty were measured by one adapted item from the SACQ (Baker & Siryk, 1999), one item by Wosnitza (2007), and two items adapted from PISA (Hertel, Hochweber, Steinert, & Klieme, 2010). Together, the four items have an internal consistency of $.75$. All items from the integration scales have four response alternatives.

Criterion validity was assessed by employing two different aspects of college outcomes: *dropout intentions* and *academic success*. *Dropout intentions* were measured by five items from Trautwein et al. (2007). Three items focus on dropping out of university, and two items focus on changing the particular field of study. The items show an internal consistency of $.85$ and have four response alternatives. *Academic success* was measured by two indicators: The first is self-reported average grades received in the students' current field of study, and the second indicator, obtained by using a five-point Likert scale, is the self-estimated progress in relation to the demands required by study regulations. Both measures are negatively correlated with $r = -.28$.

4.2 Participants

The 12,343 participants of the survey were at the beginning of their second year of studies and 22.0 years old on average ($SD = 3.7$). While 37.6% of them were men,

1 Doi:10.5157/NEPS:SC5:3.0.0.

2 To ensure comparability with the original measures, the response format of all instruments was not changed, although this resulted in either 4 or 5 response alternatives for the different measures.

62.4% were women. The most frequent fields of study were linguistic and cultural studies (27.9%), followed by business, law and social sciences (25.4%), mathematics and natural sciences (21.3%), and engineering sciences (13.8%). A 94.4% majority of the participants were born in Germany, and only 5.6% had a migration background.

4.3 Statistical Analyses

The factorial structure of the instrument was tested by confirmatory factor analyses with Mplus 5.21 (Muthén & Muthén, 2009) using maximum likelihood estimation. Missing data were handled by the integrated Full Information Maximum Likelihood procedure. First, a measurement model was tested that comprised the five constructs related to social and academic integration in which the latent factors were allowed to correlate (see Appendix B). In a second step, by assuming second-order factors, we tested whether the constructs could be separated into social and academic integration. Third, partly based on the results of the second model, another model was specified that does not differentiate between social and academic integration and instead only uses a general integration factor. Models were compared using Chi-Square difference tests and by comparisons of the Comparative Fit Index, the Tucker-Lewis Index, and Root Mean Square Error of Approximation. Predictive validity was tested in latent regression models, with *dropout intentions* and *academic success* modeled as dependent variables.

5 Results

Results of the confirmatory factor analyses are presented in Table 1. All models show a significant Chi-Square statistic, which means that the data do not fit any of the models exactly. Given that the Chi-Square statistic is sensitive to sample size (i. e., in large samples, even small differences between the specified model and the data become significant), the models should not be rejected without considering other fit indices. Fit indices, such as the Comparative Fit Index, the Tucker-Lewis Index, and Root Mean Square Error of Approximation point to a good model fit, which is nearly the same for all three models.

The standardized loadings of the first-order factors on the factor indicators are identical for all three models. They range from .70 to .85 for *interactions with fellow students*, from .61 to .73 for *interactions with faculty*, from .73 to .82 for *academic performance*, from .73 to .86 for *affective involvement*, and from .56 to .83 for *achievement orientation* (see Appendix B).

Chi-Square difference tests are significant for all model comparisons: Model 2 fits significantly worse than Model 1 ($\Delta\chi^2 = 92.359$, $\Delta df = 4$, $p < .001$), and Model 3 fits significantly worse than Models 1 and 2 ($\Delta\chi^2 = 109.595$, $\Delta df = 5$, $p < .001$; $\Delta\chi^2 =$

Table 1 Results of Confirmatory Factor Analyses

	χ^2	<i>df</i>	CFI	TLI	RMSEA
Model 1	1534.066***	94	.978	.972	.036
Model 2	1626.425***	98	.977	.972	.037
Model 3	1643.661***	99	.977	.972	.037

Note. $n = 11,696$.

CFI = Comparative Fit Index, TLI = Tucker-Lewis Index, RMSEA = Root Mean Square Error of Approximation.

*** $p < .001$

17.236, $\Delta df = 1$, $p < .001$). As was said already, these tests should be interpreted with caution when used with large sample sizes (Brannick, 1995; Chen, 2007), and in cases in which competing models show comparable fit statistics, the most parsimonious model should be preferred in general.

Since all three models are identical at the measurement level, the main interest lies in the structural relations of the latent constructs. In Model 2, which differentiates between social and academic integration, it is apparent that *interactions with fellow students* does not load equally strongly on the social integration factor as does *interactions with faculty*. Additionally, *achievement orientation* shows a loading of only .34 on the academic integration factor. Remarkably, the social and academic integration factors are correlated to .92, which possibly calls into question the view of social and academic integration as distinguishable constructs. For this reason, Model 3 was tested by assuming only a general integration factor. In this model, the loadings on the integration factor did not change substantially compared with the loadings on the two separate factors in Model 2. *Interactions with faculty*, *academic performance*, and *affective involvement* show higher loadings than *interactions with fellow students* and *achievement orientation* on the general factor.

To test construct validity, a fourth model was specified that included *dropout intentions* and *academic attainment* as validation criteria ($\chi^2 = 8484.538$, $df = 222$, CFI = .923, TLI = .912, RMSEA = .056). The standardized loadings of the dropout-intention subscale range from .67 to .83, and the indicators of *academic success* have loadings of $-.42$ and .68. For *dropout intentions*, a latent correlation of $-.79$ with integration was found, and for *academic success*, a latent correlation of .82 was found.

6 Discussion

The aim of the present study was to evaluate the validity of the operationalization of social and academic integration in NEPS Stage 7. All scales show sufficient internal consistency and factorial validity in the measurement model. Regarding factorial structure, whether social and academic integration can be viewed as two well-balanced aspects of the same construct seems questionable. At the second-order level, the one-factor model has to be preferred over the two-factor-model because of its parsimony. Another relevant finding is the weaker correlation of the subscales of *achievement orientation* and *interactions with fellow students* with the other measures. Apparently, the core concept in this operationalization of students' integration consists of *affective involvement*, *academic performance*, and *interactions with faculty*. However, the general integration factor shows criterion-related validity regarding dropout intentions and academic success, which demonstrates that the aspects of students' integration measured in NEPS Stage 7 are relevant for students' academic success and retention in higher education.

It should be mentioned that these measures of social and academic integration are not intended as a detailed operationalization either of Tinto's model or of all possible facets of academic commitment. Against the background of limited survey time, the number of constructs was limited and the corresponding scales had to be shortened to a minimum length and may thereby have lost some of their content validity. The aim was not to explore the internal structure of students' integration as a theoretical concept, but to provide a parsimonious instrument for measuring relevant aspects of students' integration. Therefore, general conclusions about the nature of students' integration (e.g., social and academic integration not being distinguishable; Beekhoven, Jong, & van Hout, 2002) may not be appropriate. Furthermore, it remains to be investigated whether personal and institutional characteristics, such as differing study patterns of traditional and nontraditional students and cultural differences between academic disciplines, affect the structure of integration or its relevance regarding outcomes.

The NEPS operationalization of students' integration places a special emphasis on parts of the construct that can be regarded as academic rather than social integration. In any event, researchers interested in students' social integration and adjustment can use the *interactions with fellow students* subscale for this purpose. The *achievement orientation* subscale seems to investigate a somewhat different aspect and may not stand in a linear relation with the other aspects of the construct: Students who show very strong effort in their studies may either be highly motivated in the sense of aspiration or may experience difficulties in their studies for which they need to compensate. As Beekhoven et al. (2002) noted, the consideration of interactions with faculty as academic or social integration depends on whether classroom interactions or extracurricular activities are used as indicators. The same holds for *interactions with fellow students*, which may be characterized either as primarily academic (e.g., in learn-

ing groups) or as social in the case of extracurricular activities that are less related to formal academic integration. Thus, the specific operationalization of students' integration may explain that interactions with faculty are more strongly connected to the academic than to the social aspect of integration, while peer-group interactions are less central to students' integration as measured in the NEPS.

Altogether, the NEPS operationalization of students' integration shows sufficient reliability at the measurement level for all aspects as well as criterion-related validity with dropout intentions and academic success. The size of these correlations is partly due to the cross-sectional character of the first online wave of the main study, in which academic success was measured by self-reports and dropout intentions were taken as an approximation of dropout. Future waves of the survey will allow for testing the predictive validity of the integration concept regarding actual dropout behavior and academic success.

References

- Aschinger, F., Epstein, H., Müller, S., Schaeper, H., Vöttiner, A., & Weiß, T. (2011). Higher education and the transition to work. In H.-P. Blossfeld, H.-G. Roßbach, & J. von Maurice (Eds.), *Zeitschrift für Erziehungswissenschaft*, 14. *Education as a lifelong process: The German National Educational Panel Study (NEPS)* (pp. 267–282). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Baird, L. L. (2000). College climate and the Tinto model. In J. M. Braxton (Ed.), *Reworking the student departure puzzle* (pp. 62–80). Nashville, TN: Vanderbilt University Press.
- Baker, R. W., & Siryk, B. (1984). Measuring adjustment to college. *Journal of Counseling Psychology*, 31(2), 179–189.
- Baker, R. W., & Siryk, B. (1999). *Student Adaptation to College Questionnaire (SACQ): Manual*. Los Angeles: Western Psychological Services.
- Bean, J. P. (1980). Dropouts and turnover: The synthesis and test of a causal model of student attrition. *Research in Higher Education*, 12(2), 155–187.
- Bean, J. P. & Eaton, S. B. (2000). A psychological model of college student retention. In J. M. Braxton (Ed.), *Reworking the student departure puzzle* (pp. 48–61). Nashville, TN: Vanderbilt University Press.
- Beekhoven, S., de Jong, U., & van Hout, H. (2002). Explaining academic progress via combining concepts of integration theory and rational choice theory. *Research in Higher Education*, 43(5), 577–600.
- Beyers, W., & Goossens, L. (2002). Concurrent and predictive validity of the Student Adaptation to College Questionnaire in a sample of European freshman students. *Educational and Psychological Measurement*, 62(3), 527–538.
- Blau, P. M., & Duncan, O. D. (1967). *The American occupational structure*. New York: Wiley.

- Borglum, K., & Kubala, T. (2000). Academic and social integration of community college students: A case study. *Community College Journal of Research and Practice*, 24(7), 567–576.
- Boudon, R. (1974). *Education, opportunity, and social inequality. Changing prospects in western society*. New York: Wiley.
- Brannick, M. T. (1995). Critical comments on applying covariance structure modeling. *Journal of Organizational Behavior*, 16(3), 201–213.
- Braxton, J. M., Milem, J. F., & Sullivan, A. S. (2000). The influence of active learning on the college student departure process: Toward a revision of Tinto's theory. *Journal of Higher Education*, 71(5), 569–590.
- Braxton, J. M., & Lien, L. A. (2000). The viability of academic integration as a central construct in Tinto's interactionist theory of student departure. In J. M. Braxton (Ed.), *Reworking the student departure puzzle* (pp. 11–28). Nashville, TN: Vanderbilt University Press.
- Breen, R., & Jonsson, J. O. (2000). Analyzing educational careers: A multinomial transition model. *American Sociological Review*, 65(5), 754–772.
- Brunstein, J. C. (1995). *Motivation nach Misserfolg*. Göttingen: Hogrefe.
- Cohen, A. (2003). *Multiple commitments in the workplace: An integrative approach*. Mahwah, NJ: Erlbaum.
- Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling*, 14(3), 464–504.
- Davidson, W. B., Beck, H. P., & Milligan, M. (2009). The College Persistence Questionnaire: Development and validation of an instrument that predicts student attrition. *Journal of College Student Development*, 50(4), 373–390.
- Durkheim, E. (1961). *Suicide: A study in sociology*. Glencoe: Free Press.
- French, B. F., & Oakes, W. (2004). Reliability and validity evidence for the Institutional Integration Scale. *Educational and Psychological Measurement*, 64(1), 88–98.
- Gold, A. (1988). Studienabbruch, Abbruchneigung und Studienerfolg: Vergleichende Bedingungsanalysen des Studienverlaufs. *Europäische Hochschulschriften, Reihe 6, Psychologie, Bd. 259*. Frankfurt am Main: Lang.
- Grässmann, R., Schultheiss, O. C., & Brunstein, J. C. (1998). Exploring the determinants of students' academic commitment. In P. Nenniger, R. S. Jäger, A. Frey, & S. Wosnitza (Eds.), *Advances in motivation* (pp. 103–109). Landau: Verlag Empirische Pädagogik.
- Halpin, R. L. (1990). An application of the Tinto model to the analysis of freshman persistence in a community college. *Community College Review*, 17(4), 22–32.
- Henecka, H. P., & Gesk, I. (1996). *Studienabbruch bei Pädagogikstudenten: Eine empirische Untersuchung an Pädagogischen Hochschulen in Baden-Württemberg*. Weinheim: Deutscher Studien Verlag.
- Hertel, S., Hochweber, J., Steinert, B., & Klieme, E. (2010). Schulische Rahmenbedingungen und Lerngelegenheiten im Deutschunterricht. In E. Klieme, C. Artelt, J. Hartig, N. Jude, O. Köller, M. Prenzel, ... P. Stanat (Eds.), *Pisa 2009. Bilanz nach einem Jahrzehnt* (pp. 113–151). Münster: Waxmann.

- Heublein, U., Richter, J., Schmelzer, R., & Sommer, D. (2014). *Die Entwicklung der Studienabbruchquoten an den deutschen Hochschulen. Statistische Berechnungen auf der Basis des Absolventenjahrgangs 2012* (Forum Hochschule 4/2014). Hannover: DZHW.
- Heublein, U., & Wolter, A. (2011). Studienabbruch in Deutschland. Definition, Häufigkeit, Ursachen, Maßnahmen. *Zeitschrift für Pädagogik*, 57(2), 214–235.
- Krotseng, M. V. (1992). Predicting persistence from the Student Adaptation to College Questionnaire: Early warning or siren song? *Research in Higher Education*, 33(1), 99–111.
- Leichsenring, H., Sippel, S., & Hachmeister, C.-D. (2011). *CHE-QUEST—Ein Fragebogen zum Adaptionsprozess zwischen Studierenden und Hochschule: Entwicklung und Test des Fragebogens* (Arbeitspapier Nr. 144). Gütersloh: Centrum für Hochschulentwicklung.
- Maaz, K., Hausen, C., McElvany, N., & Baumert, J. (2006). Stichwort: Übergänge im Bildungssystem. Theoretische Konzepte und ihre Anwendung in der empirischen Forschung beim Übergang in die Sekundarstufe. *Zeitschrift für Erziehungswissenschaft*, 9(3), 299–327.
- Mare, R. D. (1981). Change and stability in educational stratification. *American Sociological Review*, 46(1), 72–87.
- Meyer, J. P., & Allen, N. J. (1997). *Commitment in the workplace. Theory, research, and application*. Thousand Oaks: Sage.
- Müller, S., & Sarcletti, A. (2010). *Bericht über die Auswertung der Entwicklungsstudie "Soziale und akademische Integration"*. Unpublished manuscript.
- Muthén, L. K., & Muthén, B. O. (2009). *Mplus user's guide* (5th ed.). Los Angeles, CA: Muthén & Muthén.
- Napoli, A. R., & Wortman, P. M. (1998). Psychosocial factors related to retention and early departure of two-year community college students. *Research in Higher Education*, 39(4), 419–455.
- Neuville, S., Frenay, M., Schmitz, J., Boudrenghien, G., Noël, B., & Wertz, V. (2007). Tinto's theoretical perspective and expectancy-value paradigm: A confrontation to explain freshmen's academic achievement. *Psychologica Belgica*, 47(1), 31–50.
- O'Reilly, C. A., & Chatman, J. (1986). Organizational commitment and psychological attachment: The effects of compliance, identification, and internalization on prosocial behavior. *Journal of Applied Psychology*, 71(3), 492–499.
- Pascarella, E. T., & Terenzini, P. T. (1980). Predicting freshman persistence and voluntary dropout decisions from a theoretical model. *Journal of Higher Education*, 51(1), 60–75.
- Powell, J. J., & Solga, H. (2011). Why are higher education participation rates in Germany so low? Institutional barriers to higher education expansion. *Journal of Education and Work*, 24(1-2), 49–68.
- Robbins, S. B., Lauver, K., Le, H., Davis, D., Langley, R., & Carlstrom, A. (2004). Do psychosocial and study skill factors predict college outcomes? A meta-analysis. *Psychological Bulletin*, 130(2), 261–288.

- Sarcletti, A., & Müller, S. (2011). Zum Stand der Studienabbruchforschung. Theoretische Perspektiven, zentrale Ergebnisse und methodische Anforderungen an künftige Studien. *Zeitschrift für Bildungsforschung*, 1(3), 1–14.
- Schiefele, U., Moschner, B., & Husstegge, R. (2002). *Skalenhandbuch SMILE-Projekt 2002*. Unpublished manuscript, Department of Psychology, University of Bielefeld, Germany.
- Sippel, S. (2006). *Entwicklung, psychometrische Überprüfung und Validierung einer deutschen Fassung des "Student Adaptation to College Questionnaire SACQ" von R. W. Baker und B. Syrik* (Diplomarbeit). München: GRIN Verlag.
- Spady, W.G. (1970). Dropout from higher education: An interdisciplinary review and synthesis. *Interchange*, 1(1), 64–85.
- Stocké, V., Blossfeld, H.-P., Hoenig, K., & Sixt, M. (2011). Social inequality and educational decisions in the life course. In H.-P. Blossfeld, H.-G. Roßbach, & J. von Maurice (Eds.), *Zeitschrift für Erziehungswissenschaft*, 14. *Education as a lifelong process: The German National Educational Panel Study (NEPS)* (pp. 103–119). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Tierney, W.G. (1992). An anthropological analysis of student participation in college. *Journal of Higher Education*, 63(6), 603–618.
- Tillman, C. A. (2002). *Barriers to student persistence in higher education*. Retrieved from: https://nph.dev.longsight.com/sites/default/files/v2n1_Tillman.pdf
- Tinto, V. (1975). Dropout from higher education: A theoretical synthesis of recent research. *Review of Educational Research*, 45(1), 89–125.
- Tinto, V. (1993). *Leaving college: Rethinking the causes and cures of student attrition*. Chicago: University of Chicago Press.
- Trautwein, U., Jonkmann, K., Gresch, C., Lüdtke, O., Neumann, M., Klusmann, U., ... Baumert, J. (2007). *Transformation des Sekundarschulsystems und akademische Karrieren (TOSCA). Dokumentation der eingesetzten Items und Skalen. Welle 3*. Unpublished manuscript, Max-Planck-Institut für Bildungsforschung.
- Winteler, A. (1984). Pfadanalytische Validierung eines konzeptionellen Schemas zum Studienabbruch, *Hochschulausbildung*, 2(4), 193–214.
- Wosnitza, M. (2007). *Lernumwelt Hochschule und akademisches Lernen. Die subjektive Wahrnehmung sozialer, formaler und materiell-physischer Aspekte der Hochschule als Lernumwelt und ihre Bedeutung für das akademische Lernen*. Landau: Verlag Empirische Pädagogik.

Appendix A: Measures³

Affective involvement

I can completely identify with my studies. (Ich kann mich mit meinem Studium voll identifizieren.)

I enjoy my field of studies very much. (Mein Studium bereitet mir sehr viel Freude.)

To be honest, my studies don't thrill me. (Offen gestanden, macht mir mein Studium wenig Spaß.)

Achievement orientation

I invest a great deal of effort in order to be successful in my studies. (Ich investiere sehr viel Energie, um in meinem Studium erfolgreich zu sein.)

I do not dedicate more time to my studies than absolutely necessary. (Ich tue für mein Studium nicht mehr, als unbedingt erforderlich ist.)

I pursue high aspirations concerning my academic performances. (Wenn es um Leistungen in meinem Studium geht, stelle ich an mich selbst höchste Ansprüche.)

Perceived academic performance

My academic achievements (grades) are better than I had originally expected. (Meine Leistungen im Studium sind besser, als ich ursprünglich erwartet hatte.)

I am satisfied with my performance in the degree program. (Mit meiner Studienleistung bin ich zufrieden.)

I have fully met my own expectations for my performance and grades in this degree program. (Meine Leistungserwartungen und -ansprüche haben sich im Studium voll erfüllt.)

Interactions with faculty

I get along well with the instructors in my degree program. (Mit den Lehrenden meines Studiengangs komme ich gut zurecht.)

3 The English translations of the first six items originate from Grässmann et al. (1998).

Most of the instructors treat me fairly. (Die meisten Lehrenden behandeln mich fair.)

I feel accepted by the instructors. (Ich fühle mich von den Lehrenden anerkannt.)

The instructors are interested in what I have to say. (Die Lehrenden interessieren sich für das, was ich zu sagen habe.)

Interactions with fellow students

I have been successful in building contacts with other students during my studies. (Mir ist es während meines bisherigen Studiums gut gelungen, Kontakte zu anderen Studierenden aufzubauen.)

I know a lot of classmates with whom I can exchange ideas about questions in my field of study. (Ich kenne viele Kommiliton(inn)en, mit denen ich mich über fachspezifische Fragen austauschen kann.)

I have many contacts with students in my cohort. (Ich habe viele Kontakte zu Studierenden aus meinem Semester.)

Dropout intentions

I've often thought about dropping out. (Ich habe schon öfter daran gedacht, das Studium abzubrechen.)

I am seriously thinking of completely abandoning the degree program. (Ich denke ernsthaft daran, das Studium ganz aufzugeben.)

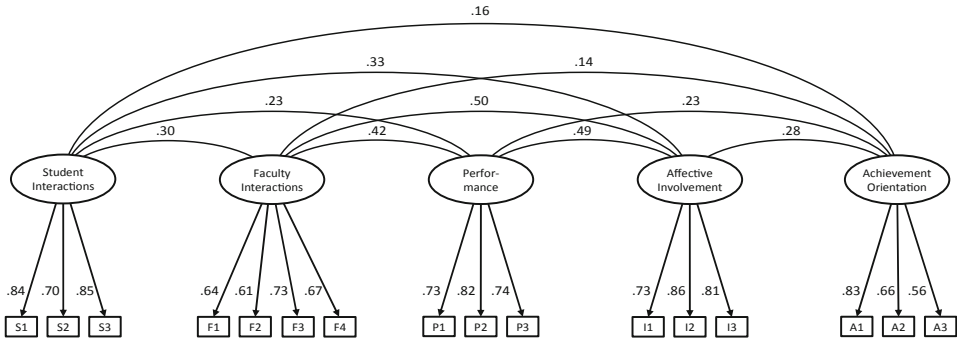
I will complete this degree program no matter what. (Ich werde mein Studium auf jeden Fall bis zum Abschluss weiterführen.)

I am seriously thinking about changing my major field of study. (Ich denke ernsthaft daran, mein Hauptfach zu wechseln.)

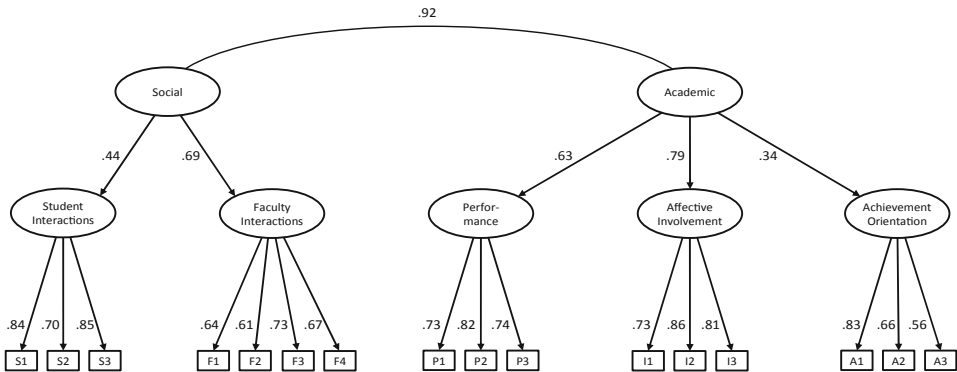
If I could choose again, I would opt for another field of study. (Wenn ich nochmals wählen könnte, würde ich mich für ein anderes Studienfach entscheiden.)

Appendix B: Models of Confirmatory Factor Analyses

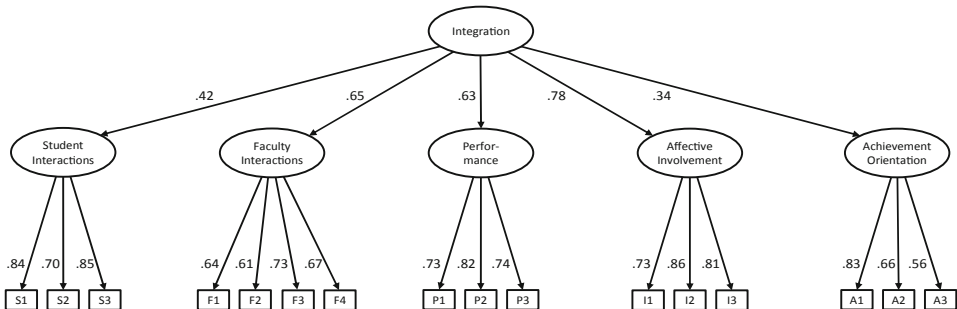
Model 1 Correlated factors



Model 2 Social and academic integration as higher-order factors



Model 3 Only integration as higher-order factor



Acknowledgement

This paper uses data from the National Educational Panel Study (NEPS): Starting Cohort 5—First-Year Students, doi:10.5157/NEPS:SC5:3.0.0. The NEPS data collection is part of the Framework Program for the Promotion of Empirical Educational Research, funded by the German Federal Ministry of Education and Research and supported by the federal states.

About the authors

G. Dahm

German Centre for Higher Education Research and Science Studies (DZHW), Hanover.

e-mail: dahm@dzhw.eu

S. Hahn

German Youth Institute (DJI), Munich.

e-mail: hahn@dji.de

O. Lauterbach

Hessian Teaching Staff Academy, Wiesbaden.

e-mail: oliver.lauterbach@kultus.hessen.de

Why Do We Collect Data on Educational Histories Over the Life Course the Way We Do? Core Questionnaire Design Decisions in Starting Cohort 6—Adults

Katrin Drasch, Corinna Kleinert, Britta Matthes and Michael Ruland

Abstract

Starting Cohort 6—Adults is one of six samples of the German National Educational Panel Study (NEPS) and covers members of the adult population living in Germany from multiple birth cohorts. It aims at collecting data on educational processes and competence development in adult life as well as on learning environments, decision processes, and returns. To achieve these objectives, it is necessary to gather life-course information, particularly in the area of education and employment. In this chapter, we describe our core questionnaire design and justify why we collect life-course data the way we do. We begin by presenting theoretical principles of life course research and discussing their consequences for questionnaire design. Subsequently, we describe how the process of recalling events and their dating is supported by instrument design in order to guarantee that retrospective life course data will be complete and consistent. Finally, we illustrate the analysis potential of the collected life course data.

1 Introduction

Starting Cohort 6—Adults of the German National Educational Panel Study (NEPS) aims at collecting high-quality data on educational processes and competence development in adult life as well as on learning environments, contextual conditions, decision processes, and returns (for a detailed description of the conceptual framework, see Allmendinger et al., 2011). Education in adult age differs in nature from education in childhood and youth: For the most part, learning no longer takes place in institutionalized, age-standardized contexts such as schools, but rather in a variety of shorter courses, self-learning activities, and nonformal/informal learning in the context of work, family, and volunteering. These forms of learning may happen—at least

theoretically—at any time and during any circumstances in adult life. Thus, information on all these different learning activities has to be gathered in NEPS Starting Cohort 6, including data on learning environments and the basic decision processes that lead to participation. Lifelong learning not only takes place in different life-course contexts, but it is also embedded in educational, employment, and family careers. Participation in adult education depends, for example, on previous education, on jobs performed, and on family arrangements. Vice versa, educational outcomes, such as competencies and certificates, may affect further educational activities. As a consequence, we have to collect complete and detailed longitudinal data on education and learning activities, jobs, and family histories (Allmendinger et al., 2011, p. 285).

In order to fulfil these requirements, NEPS Starting Cohort 6—Adults was designed as a large, representative sample of the population living in Germany born between 1944 and 1986, meaning that at the time of the first interview, respondents were between 22 and 65 years old. Similar to the other five NEPS starting cohorts, Starting Cohort 6 was planned as a longitudinal study that observes respondents' individual learning processes and competence development over time. For this purpose, computer-assisted personal and telephone interviews have been being conducted in yearly intervals since 2009 (for details see Allmendinger et al., 2011, pp. 295 ff.).

Against this background, the questionnaire design of Starting Cohort 6—Adults has to meet different challenges simultaneously. On the one hand, in order to gather complete and consistent information on respondents' entire educational histories as well as their contexts, it is necessary to collect *retrospective life-course data*. On the other hand, obtaining information on competence development and subjective information on educational decisions and education-related attitudes that cannot be recalled retrospectively requires a *prospective panel design*. Combining both retrospective life-course and prospective panel designs allows for updating life-course information from panel wave to panel wave and for collecting information on all important events that happened between panel waves. This guarantees that disadvantages of both designs are mutually compensated for: The main shortcoming of panel designs is that events that happened before the first interview or occurred between panel waves remain unknown. This disadvantage is particularly problematic in the field of education because a panel survey would have to start early in the life course and thus sample children at a young age and follow them for many years in order to collect comprehensive data on educational careers. In contrast, purely retrospective surveys have the problem of restricted reliability due to the limitations and biases of autobiographic memory.¹ In order to guarantee that retrospective life-course data are as complete and consistent as possible, the questionnaire design must account for the multi-dimensionality of life courses and support the process of recalling events and

1 It should be noted that disadvantages of both designs in terms of sample selectivity (previously in the case of retrospective surveys, increasing in terms of panel surveys) cannot be eliminated by combining them. Here, other steps have to be taken to minimize these errors.

their dating. In this chapter, we describe and explain how we have implemented these requirements in Starting Cohort 6—Adults.²

In the remainder of this article, we give an overview of how we collect life-course data in the Starting Cohort 6—Adults and how we reached these decisions. First, we discuss basic theoretical considerations and derive conclusions for questionnaire design. Second, we describe how the process of recalling events and their dating is supported to guarantee that retrospective life-course data are complete and consistent. Third, the analysis potential of the life-course data is shown. The article concludes with a summary and considerations for future questionnaire development.

2 Theoretical and Methodological Considerations

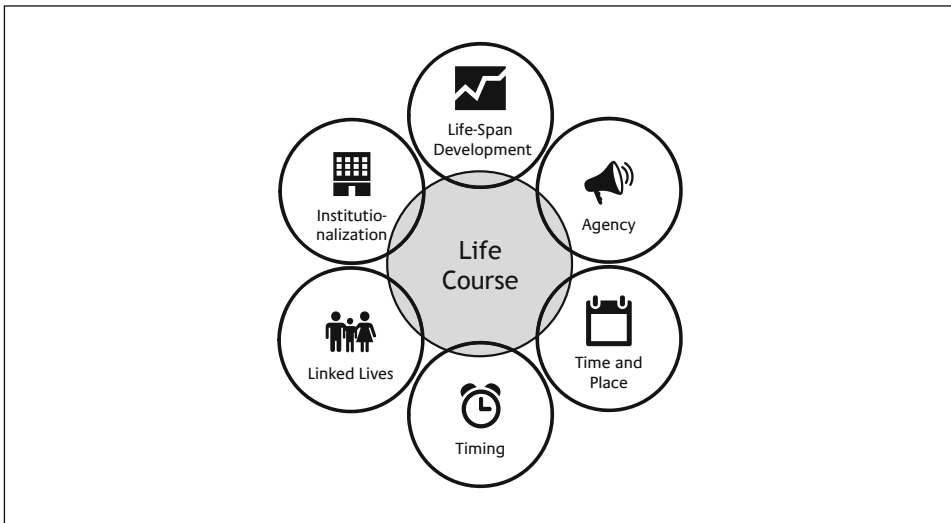
Analyzing life courses as a research strategy has continuously increased in importance over the past decades, mainly in sociological research, but also in psychological, health, and economic studies (e.g., Mayer, 2004; Levy et al., 2005; Ben-Shlomo, Mishra, & Kuh, 2014; Fend, 2014; Layard et al., 2014; Settersten & Hagestad, 2015). This notion brings a dynamic perspective to topics that are often considered “cross-sectional slices of life” (Giele & Elder, 1998, p. VIII). Such data can be used to examine social processes extending over the individual life span or significant parts of it. However, it can also be used to analyze structural processes at the macro level in a common conceptual and empirical frame of reference. Therefore, the collection of life-course data is regarded as the “gold standard of sociological research” (Mayer, 2009, p. 413).

2.1 Principles of life-course research

The concept of the life course refers to a sequence of a person’s decisions among institutionally predefined alternatives in a number of different life domains.³ Life-course events differ from other decisions because they are self-referential and endogenous. They refer to other decisions in the same life domain or in other life domains, and they tend to accumulate over the individual life-span (Meulemann, 1990). Life-course

2 Many of the life-course related design decisions are based on experiences from the German Life History Study (GLHS) (e.g., Brückner & Mayer, 1998; Matthes, Reimer, & Künster, 2007). While the GLHS surveys were confined to single birth cohorts, their design was applied to a representative adult population sample in the survey ‘Working and Learning in a Changing World’ (ALWA) (Kleinert et al., 2011). This survey serves as forerunner study of NEPS Starting Cohort 6—Adults, and its participants were integrated in its sample.

3 The term ‘life course’ applies to an ‘objective’ sequence of events in an individual’s life, whereas the term ‘biography’ is used for the subjective interpretation and processing of these events.

Figure 1 Principles of Life-Course Research

events can be described by six principles that are consequential for data collection and analysis (Figure 1):⁴

- 1) The principle of *life-span development* means that learning is a lifelong activity, and educational processes thus take place over the entire life course.
- 2) The principle of *agency* emphasizes that individuals are not passively influenced by structural constraints and opportunities. Instead, they make planned choices and compromises based on perceived alternatives.
- 3) The principle of *time and place* points out that transitions and life events are shaped by historical conditions.
- 4) The principle of *timing* means that the same event may affect individuals differently depending on when it occurs in their lives.
- 5) The principle of *linked lives* indicates that individuals are influenced by interpersonal relations with relevant others, such as family members, friends, schoolmates, and colleagues. Thus, their life courses are mutually dependent.
- 6) The principle of *institutionalization* refers to the fact that life courses are socially structured. Individuals participate in a society in segmented roles that vary over their life courses, and their decisions about life-course events refer to institutionally predefined patterns. Consequently, the social structure of a society can be described by aggregating these decisions (Kohli, 1985).

4 Principles (1) to (5) were formulated by Elder (2003). Principle (6) was emphasized prominently by German researchers, for example by Mayer and Schoepflin (1989, p. 196).

2.2 Conclusions for questionnaire development

In the following section, the consequences of these six principles in term of data requirements, suitable survey designs, and question formats are discussed for a longitudinal survey of the adult population that is devoted to education.


✓ The principle of *life-span development* requires covering educational processes over the respondents' entire life course as completely and consistently as possible. This includes formal education; shorter, non-certified training courses; and self-learning that takes place in the context of other activities such as working, parenting, or volunteering. For respondents in adult age, this means collecting information on educational attainment retrospectively as well as prospectively. Thus, in NEPS Starting Cohort 6—Adults, data on respondents' previous educational histories were collected in the first panel wave. In subsequent waves, educational activities are continuously updated for the time since the last interview (for details see Trahms, Ruland, & Matthes in this volume).


📢 From the principle of *agency*, we can conclude that in order to map educational decision processes, information on attitudes, expectations, and aspirations regarding education and learning has to be collected continuously over the life course as well. However, asking for these subjective dimensions in retrospect does not mean collecting reliable measures of respondents' attitudes at the time when the decision occurred, but rather gathering re-interpretations that are heavily influenced by past and present experiences (Schnell, 2012, p. 42). Consequently, data gathered this way is plagued by measurement error. Thus, in NEPS Starting Cohort 6—Adults, information on attitudes, expectations, and aspirations is only collected prospectively, for example, by asking in regular intervals for educational decisions planned or proposed for the future (for details see Hoenig, Pollak, Schulz, & Stocké in this volume).

📅 The principle of *time and place* calls for localizing life-course data in terms of information on when and where events took place. In order to assess the impact of historical conditions, life-course events in our questionnaire are dated and enriched with questions on localities, such as respondents' residences, locations of schools, and jobs. Furthermore, if one plans to compare the effects of different historical conditions on life-course events, it is advisable to include a broad range of birth cohorts in the sampling frame, as was done in NEPS Starting Cohort 6—Adults.

🕒 From the principle of *timing*, we can infer that life courses should not be measured 'roughly' on a single time axis or subjectively in the sense of biographies, but all events should be dated as exactly as possible to be able to reconstruct their temporal structure. The principle of timing thus requires collecting event-history data. Here, not only the type of event and its position in the life course are recorded, but

also the precise times of transitions in and out of various states, and thus the timing, duration, and sequence of events (Auriat, 1991). A complete event history ideally covers all relevant domains of the life course and thus all interconnections of different events and transitions from one state to another within and between life domains. Experiences from previous surveys have shown that collecting this type of data works well for reconstructing formal educational careers as well as residence, employment, and family histories (Auriat, 1993). In contrast, shorter training courses, which are typical of adult learning, are hard to remember if they took place some years ago, and self-learning activities often have no clear temporal structure and thus cannot be dated (Dürnberger, Drasch, & Matthes, 2011). Hence, these two types of education are measured differently in Starting Cohort 6—Adults (for details, see Janik, Wölfel, & Eisermann in this volume).

 The principle of *linked lives* means that information on relevant others has to be collected in a way that enables researchers to analyze these links systematically. Thus, we do not only need life-course data from the respondents themselves, but also information of other persons that are or have been relevant for the respondents at certain points in their lives—first and foremost parents, former and current partners, and children (Moen, 2003). Since partners and children may influence respondents constantly in adult age, we decided to collect event-history data on these groups as well. In contrast, information on parents is only collected for single time points in the respondents' youth because we assume that parents affect respondents' educational pathways more prominently in childhood.

 Finally, the *principle of institutionalization* has two consequences for questionnaire development. First, we have to gather information on institutional and organizational contexts of life-course events. Second, the main institutional alternatives have to be known by the questionnaire developers beforehand for all the historical periods and regions relevant for the NEPS adult population. Then, they have to be translated into meaningful questions on available institutional alternatives, and they have to be updated continuously when the social context changes. For example, when asking our respondents about the type of secondary schools, we had to incorporate all existing school types in the Federal Republic of Germany as well as in the GDR before the German re-unification.

Two more general questions have not yet been answered. Since a considerable part of the Starting Cohort 6—Adults respondents' lives already lies behind them, how do we manage to guarantee in practical terms that they report everything that has happened so far in their lives, particularly concerning their educational activities? Moreover, given that the reliability of retrospective reports is limited, how can we support respondents in remembering and dating the events in their lives correctly? These questions are answered in the next section.

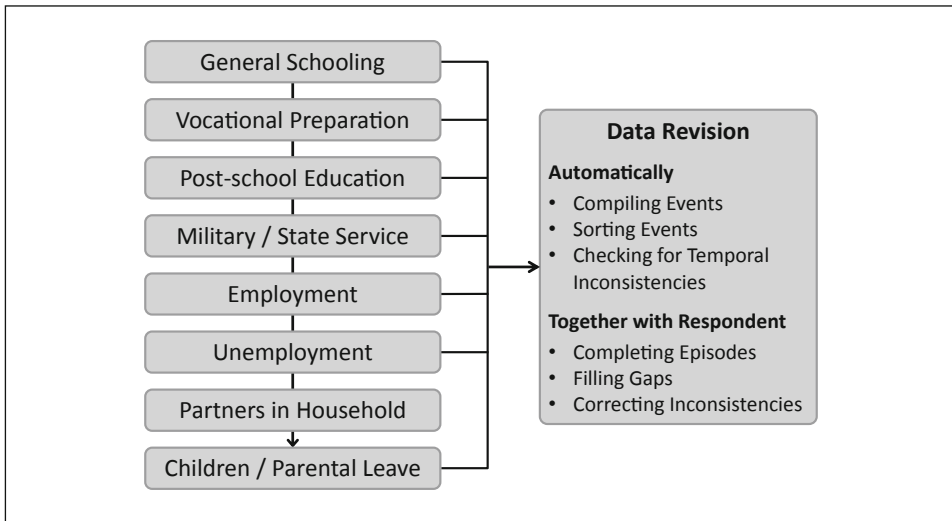
3 Supporting Retrospective Recall

There is wide agreement in research that there are certain problems involved in letting respondents look back in time. On the one hand, they may not remember events that took place in the past or they may have forgotten when they took place. On the other hand, respondents are not only affected by the conditions of the point in time when the event was happening, but they are also influenced by the changes since then, which may lead to misinterpreting or misdating past events. A particular research strand is devoted to identifying the best method of reducing these retrospective recall errors (Drasch, & Matthes, 2013; Reimer, & Matthes, 2007; Belli, Lee, Stafford, & Chou, 2004; Bluck, 2003; Stone et al., 2000; Dex, 1995). As insights from cognitive psychology have shown, the memory process of recalling events and dating them can be supported by questionnaire design (Conway, Rubin, & Rubin, 1996). In the following section, four core design features that may fulfil this function are presented and discussed.

3.1 Modularized Life-Course Reports and Data Revision

One fundamental decision in retrospective surveys is to either collect event-history data by going along a single time line or to separate the life course into various domains and gather all events within each domain along their own time lines, which is referred to as modularization. Modularizing the life course has some obvious advantages (see also Ruland, Drasch, Künster, Matthes, & Steinwede in this volume): Respondents are asked about chronological progressions throughout their lives more than once (Reimer, 2005), every episode is cued specifically (Reimer & Matthes, 2007), and the significant underreporting of shorter, parallel, seemingly irrelevant, and socially undesirable episodes is avoided (Auriat, 1991; Glasner, van der Vaart, & Belli, 2012). For example, research has shown that collecting unemployment episodes in a separate module results in a more precise and complete acquisition than collecting the entire history of employment states within one module (Drasch & Matthes, 2013). Thus, in Starting Cohort 6—Adults, life-course data are collected via modularized self-reports, and the respondent's life course is split up into several thematic domains.

If a modularized design is chosen, the next question is how to order the modules within the questionnaire. Since the optimal sequence of life-course modules has not been examined by empirical research yet, our decisions are based on the principle of *institutionalization* (■). According to Kohli (1985), the labor market profoundly structures everyday life in industrialized societies via a common temporal ordering of life courses into three sequential life periods: preparation, activity, and retirement. We assume that these institutional structures are reflected in the representation of the life course in the respondent's memory so that cues such as schooling, vocational training, and employment should correspond to a certain stored lifetime period, and

Figure 2 Modularized Life-Course Report in Starting Cohort 6—Adults

recall is stimulated by sequencing the interview in this way. In Starting Cohort 6—Adults, the first four domains are devoted to ‘preparation’ (Figure 2). The instrument starts with the earliest life domain that respondents are able to remember from their own experience, namely school education. It proceeds with vocational preparation and post-school education, such as vocational training, tertiary educational, and further training, followed by a short module on military, civilian, and voluntary service. The ‘activity’ phase in the life course is operationalized with four more modules: First, a module on employment episodes in which detailed information on all jobs, primary and secondary, as well as on further training on-the-job is collected. Unemployment information is gathered in a separate module. Finally, two modules collect information on family histories—first on partners in the household and second on children and parental leave.⁵

By modularizing the questionnaire into various life domains and by following a single timeline within each module, the biographical context sometimes gets lost. Hence, gaps and inconsistencies in life-course reports often do not become apparent to both interviewers and respondents. For this reason, a *data-revision module* was implemented in the interview (for details, see Ruland, Drasch, Künster, Matthes, & Steinwede in this volume). In this module, all episodes collected in the different longitudinal modules that are part of the education and employment history are com-

5 Due to the age range of the respondents, the retirement phase was not implemented in the core life-course questionnaire of the first panel wave. However, a separate module was added in later waves.

plied and sorted automatically.⁶ Temporal inconsistencies, such as overlaps or gaps, are visualized on the interviewer's screen, and tools are provided to fill in these gaps (by adding episodes and dates) and to solve inconsistencies (by either accepting overlaps or correcting dates) together with the respondent. The data-revision module is also used to identify times of inactivity that have not been captured by the other modules, such as being a housewife, having a longer illness, or being in early retirement.

3.2 Contextualization

In Starting Cohort 6—Adults, a great amount of data is recorded for each of the various life domains: For example, school history is not characterized merely by the number of schools attended and the accompanying starting and ending dates. Instead, additional information on every single school episode is asked for, such as the type of school, its location, how it ended, certificates, and grades that were earned. In contrast to asking ad hoc for some details about a school career, by using *contextualization*, the respondent's memory is carried back to a specific episode, and the respondent's recall of detailed aspects is supported. *Contextualization* can be justified with the organization of autobiographic memory in certain themes, such as life domains. When first stimulating these themes, more detailed memories of events within the themes can be retrieved (Conway, Rubin, & Rubin, 1996). Another way to contextualize events is to relate starting and ending dates from one episode to another, for example, in the data-revision module. Contextualization also plays an important role when collecting life-course information for the time between interviews. Proactive dependent interviewing—explicitly reminding the respondent of answers in the previous survey wave before asking what happened next—is another suitable way to contextualize retrospective reports. We use this technique in successive panel waves to update respondents' life courses (for details, see Trahms, Matthes & Ruland in this volume). In combination, the different techniques of contextualization ensure that better information is collected on long-forgone events and on details that are difficult to remember retrospectively when asked for without context.

6 Apart from parental leave, information on partners and children is not included here because events in the history of others do not systematically have an impact on respondents' educational and employment careers. We decided not to implement a second data-revision module for the family history in order to save interview time and because these events are of minor importance in the context of the NEPS survey.

3.3 Personalization

Collecting data on third persons creates a special case of recall problems because it can be assumed that recalling events in other persons' lives is more difficult than recalling one's own experiences. If a respondent is asked to provide information on partners, children, or co-workers, recall works via self-related information of the occasion in which he/she first learned about the respective fact. Only if the respondent experienced an event in the life course of the third person and this event is linked to emotional, visual, and other impressions long-term storage and retrieval in the respondent's memory take place, and this event can be reliably asked about later on (Kuiper & Rogers, 1979; Larsen & Plunkett, 1987). In order to stimulate a respondent's recall of the third person and to link data on this person to questions in further panel waves, the respondent needs unique cues to (re-)identify this person. Since one important and unchangeable cue is the first name of the person, this is the first piece of information that is asked about partners and children in the first interview and that is given in subsequent panel waves when new questions about them are asked.

3.4 Computer-Assisted Interviewer Tools

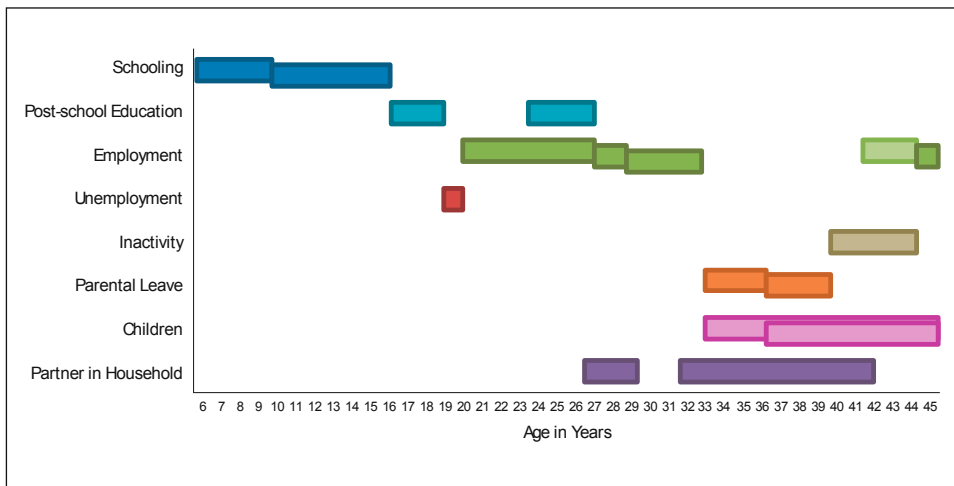
Gathering retrospective life-course data is a highly complex task not only for respondents, but also for interviewers. Besides asking questions and recording responses, interviewers sometimes have to clarify concepts in case of ambiguities to detect and point out misunderstandings or data problems and to collaborate with the respondent in order to correct incomplete biographical constellations. To help interviewers master these demanding tasks and to account for the high inter-individual variance of adult lives, we use a computer-based questionnaire with complex screening and filtering as well as personalized question insertions that draw on earlier responses. For example, when asking for detailed information on a certain employment episode, the starting date as well as the occupation is repeated. Additionally, the interviewer has the possibility to open a list of all episodes that have been reported so far during the interview in order to clarify dates jointly with the respondent. The interviewer is supported when addressing respondents' life courses, particularly in the data-revision module (see Section 3.1), because this module identifies potential problems automatically and prompts the interviewer to resolve them. These tools help the interviewer to concentrate on recording the data as precisely and accurately as possible and on assisting respondents' recall by referring to information provided earlier during the interview.

4 Analysis Potentials

The richness and multidimensionality of life-course data collected in the form of event-history data can be shown best by looking at a fictive life-course pattern. Even though life domains, such as education, employment, and family, are highly interconnected in reality, for analysis purposes, it is extremely useful to consider them as separate spheres. Every respondent’s event history consists of several spells that represent the single episodes in different life domains. As a result, episodes can be visualized by plotting them along a continuous time axis that depicts the respondent’s age. By devoting the vertical axis to the different life domains, it is possible to produce a two-dimensional illustration of a single respondent’s life course (Figure 3).

The respondent shown in Figure 3 went to two different types of schools between age 6 and 16, followed by a three-year episode of post-school education. Afterwards, she had been unemployed for a short time before entering employment at the age of 20. In her mid-twenties, the respondent took up education again, this time in addition to continuing to work. After two job changes and one short partnership, a longer phase of cohabitation began. Shortly after, two children were born, for whom the respondent took three years of parental leave each. In her early forties, she was not active in the labor market (presumably being a full-time caregiver). After some years, she took up secondary employment, and in the last two years before the interview, she was employed again while living alone with her two children. While Figure 3 only shows the basic states of the life course, NEPS Starting Cohort 6 data provide much more information on each of these episodes. Thus, we know, for example, that the respondent first had visited elementary and then lower secondary school (Realschule)

Figure 3 Example of a Life-Course Pattern in Starting Cohort 6—Adults



before she was trained as a nurse. Some years later, she specialized in surgery nursing, which led to a better-paid and more prestigious position in another hospital.

Applying the principles of life-course research (cf. Section 2), Starting Cohort 6—Adults' life-course data may be analyzed with a broad range of foci. First, researchers may examine *transitions and durations* to estimate the timing of certain life events. Thus, the research questions addressed either concentrate on reasons why some people leave or enter a particular state (e. g., leaving school, attending university) or they focus on factors that affect the length of waiting times until persons enter a certain activity (such as waiting for an apprenticeship) or the duration of a certain state (such as the length of unemployment). In these micro-level research questions, the causes influencing transitions and durations and their temporal and causal order are the main point of interest. For example, certain steps that may lead to upward mobility in adult life, such as investment in further training, can be analyzed in detail. Empirical studies may detect not only their general effect, but also consequences of their timing in the life course and short-term as well as long-term impact. This analysis potential is not restricted to educational or labor market transitions, for any transition that is captured by the longitudinal modules can be explored. This is ideally done by techniques of event-history modeling (Blossfeld, Golsch, & Rohwer, 2007).

A second focus is to not concentrate on single transitions, but to explore *trajectories* over a longer time in the life course, for example, educational pathways. Here, researchers usually aim at identifying certain patterns of trajectories and cluster individuals based on these patterns, usually by applying sequence analysis techniques, such as the Optimal Matching Algorithm. Sequence analysis is becoming increasingly popular nowadays in social science research due to new methodological developments (Aisenbrey & Fasang, 2010). The combination of analyzing transitions and durations with respect to the timing of certain life events and describing patterns of trajectories can then be used to explore endogenous path dependencies of life events (Mayer, 1987).

Third, since data of the Starting Cohort 6—Adults is not restricted to members of specific birth cohorts, it is possible to investigate how individuals' life courses are influenced by *historical and spatial contexts*. Comparing several historical contexts also means investigating the influence of institutional change on individual behavior. Disentangling the effects of time and place becomes possible by examining regional or temporal variations of structural conditions on events and transitions in respondents' life courses. In consequence, NEPS Starting Cohort 6 data are useful for arriving at empirical results regarding the distinction of age, period, and cohort effects (e. g. Mayer & Huinink, 1990).⁷

7 However, there are limitations to this objective even with the NEPS adult data, as discussed in the heated controversy on methods to disentangle age, cohort, and period effects. A good starting point to become familiar with this discussion can be found, for example, in Winship and Harding (2008) and Yang et al. (2008).

Fourth, the broader social structure of the family can be incorporated in the research framework as proposed by the *linked lives* principle, which accounts for the fact that individual life courses are embedded in broader structures (Moen, 2003). Most importantly, life courses of adults depend on their family context, for example, on parents, partners, and children. Thus, some proxy information of former and current partners as well as children, for example, the highest educational degree or current employment status, is available in the data.

5 Summary and Outlook

NEPS Starting Cohort 6—Adults aims at collecting data on educational processes and competence development in adult life as well as on learning environments, contextual conditions, decision processes, and returns. Six basic theoretical principles of life-course research—life-span development, agency, time and place, linked lives, and institutionalization—provide important guidelines how to reach this aim.

The principle of *life-span development* requires covering educational processes over respondents' entire life courses as completely and consistently as possible. For respondents in adult age, this means collecting information on educational attainment retrospectively and updating it in regular terms. While this works well for formal education, past nonformal/informal learning activities cannot be remembered well or dated exactly. Hence, the data collection has to combine retrospective as well as prospective survey designs. Furthermore, the principle of *agency* means that attitudes, expectations, and aspirations regarding education have to be collected over the life course as well. Since retrospective data on these subjective evaluations is plagued by measurement error, this kind of information can also only be collected prospectively. The principle of *time and place* calls for localizing life-course data in terms of information on when and where events took place. Similarly, from the principle of *timing*, we can conclude that all events in the life course should be dated as exactly as possible in order to be able to reconstruct their temporal structure. Implementing the principle of *linked lives* means that the NEPS adult survey should collect life-course data not only on respondents, but also on relevant others, particularly on partners and children. The *principle of institutionalization* calls for gathering information on institutional and organizational contexts of life-course events and for designing instruments that ask for the whole range of available institutional alternatives.

A central requirement for collecting valid and reliable life-course data is that every respondent reports all relevant episodes in his/her life course and dates them correctly. To make sure that retrospective reports are valid and reliable and that survey estimates measure what they are meant to questionnaires have to be designed in a way that they prevent recall errors and biases as far as possible. In Starting Cohort 6—Adults, this method has been attempted by four core design features: by splitting the life course into different thematic modules (modularization), by stimulating broader

life-course themes and relating starting and ending dates from different episodes in order to retrieve more detailed memories (contextualization), by personalizing life-course questions on third persons (such as spouses and children), and by providing computer-assisted tools for the interviewers. As a result, the questionnaires used in the Starting Cohort 6—Adults' survey are, in fact, time-consuming, complex computer programs that only work with the employment of well-trained and active interviewers who keep up respondents' motivation.

However, three trends make social research more attentive towards collecting data via the World Wide Web: the rapid growth of internet coverage, the increasing problems of coverage and selection bias in telephone surveys, and the increasing need to be more cost-efficient, which restricts the usage of face-to-face interviews in large-scale surveys. To our knowledge, hardly any attempts have yet been made to collect event-history data with adaptive online instruments, which optimally stimulate retrospective recall. On the one hand, online interviewing would allow respondents to give information about their lives very flexibly and in shorter units of interview time. On the other hand, online interviews would pose a particular challenge to retrospective life-course research due to the lack of an interviewer, who acts as an informed agent, more or less knows what the researcher is looking for, and controls data collection. Thus, developing life-course instruments suited for web-based surveys and utilizing these surveys' particular advantages represent an important challenge for future survey research.

References

- Aisenbrey, S., & Fasang, A. E. (2010). New life for old ideas: The 'second wave' of sequence analysis bringing the 'course' back into the life course. *Sociological Methods & Research*, 38, 420–462. doi: 10.1177/0049124109357532
- Allmendinger, J., Kleinert, C., Antoni, M., Christoph, B., Drasch, K., Janik, F., Leuze, K., Matthes, B., Pollak, R., & Ruland, M. (2011). Adult education and lifelong learning. In H.-P. Blossfeld, H.-G. Roßbach, & J. von Maurice (Eds.), *Zeitschrift für Erziehungswissenschaft, 14. Education as a lifelong process: The German National Educational Panel Study (NEPS)* (pp. 283–299). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Auriat, N. (1991). Who forgets? An analysis of memory effects in a retrospective survey on migration history. *European Journal of Population*, 7, 311–342. doi: 10.1007/BF01796872
- Auriat, N., (1993). My wife knows best: A comparison of event dating accuracy between the wife, the husband, the couple and the Belgium population register. *Public Opinion Quarterly*, 57(2), 165–190.
- Belli, R. F., Lee, E. H., Stafford, F. P., & Chou, C.-H. (2004). Calendar and question-list survey methods: Association between interviewer behaviors and data quality. *Journal of Official Statistics*, 20(2), 185–218.

- Ben-Shlomo, Y., Mishra, G., & Kuh, D. (2014). Life course epidemiology. In W. Ahrens, & I. Pigeot (Eds.), *Handbook of epidemiology* (pp. 1521–1549). New York: Springer.
- Blossfeld, H.-P., & von Maurice, J. (2011). Education as a lifelong process. In H.-P. Blossfeld, H.-G. Roßbach, & J. von Maurice (Eds.), *Zeitschrift für Erziehungswissenschaft, 14. Education as a Lifelong Process: The German National Educational Panel Study (NEPS)* (pp. 19–34). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Blossfeld, H.-P., Golsch, K., & Rohwer, G. (2007). *Event history analyses with Stata*. Mahwah, New Jersey: Lawrence Erlbaum Association.
- Bluck, S. (2003). Autobiographical memory: Exploring its functions in everyday life. *Memory, 11*, 113–123. doi: 10.1080/741938206
- Brückner, E., & Mayer, K. U. (1998). Collecting life history data: Experiences from the German life history study. In J. Z. Giele, & G. H. Elder (Eds.), *Methods of life course research: Qualitative and quantitative approaches* (pp. 152–181). Thousand Oaks: Sage.
- Conway, M. A., Rubin, D. C., & Rubin, D. C. (Eds.) (1996). *Autobiographical knowledge and autobiographical memories*. Cambridge, MA: Cambridge University Press.
- Dex, S. (1995). The reliability of recall data: A literature review. *Bulletin de Méthodologie Sociologique, 49*, 58–89. doi: 10.1177/075910639504900105
- Drasch, K., & Matthes, B. (2013). Improving retrospective life course data by combining modularized self-reports and event history calendars. Experiences from a large scale survey. *Quality & Quantity. International Journal of Methodology, 47*, 817–838. doi: 10.1007/s11135-011-9568-0
- Dürnberger, A., Drasch, K., & Matthes, B. (2011). Kontextgestützte Abfrage in Retrospektiverhebungen. Ein kognitiver Pretest zu Erinnerungsprozessen bei Weiterbildungsergebnissen. *Methoden, Daten, Analysen. Zeitschrift für empirische Sozialforschung, 5*(1), 3–35.
- Elder, G. H., Jr. (2003). The life course in time and place. In W. R. Heinz, & V. W. Marshall (Eds.), *Social dynamics of the life course. Transitions, institutions, and interrelations* (pp. 57–71). New York: Aldine de Gruyter.
- Fend, H. (2014). Bildungslaufbahnen von Generationen: Befunde der Life-Studie zur Interaktion von Elternhaus und Schule. In K. Maaz, J. Baumert, & M. Neumann (Eds.), *Herkunft und Bildungserfolg von der frühen Kindheit bis ins Erwachsenenalter* (pp. 37–72). Wiesbaden: Springer.
- Giele, J. Z., & Elder, G. H., Jr. (Eds.) (1998). *Methods of life course research. Qualitative and quantitative approaches*. London: Sage.
- Glasner, T., van der Vaart, W., & Belli, R. F. (2012). Calendar interviewing and the use of landmark events—Implications for cross-cultural surveys. *Bulletin de Méthodologie Sociologique, 115*, 45–52. doi: 10.1177/0759106312445701
- Kleinert, C., Matthes, B., Antoni, M., Drasch, K., Ruland, M., & Trahms, A. (2011). ALWA—new life course data for Germany. *Schmollers Jahrbuch. Zeitschrift für Wirtschafts- und Sozialwissenschaften, 131*(4), 625–634.

- Kohli, M. (1985). Die Institutionalisierung des Lebenslaufs. Historische Befunde und theoretische Argumente. *Kölner Zeitschrift für Soziologie und Sozialpsychologie*, 37(1), 1–29.
- Kuiper, N. A., & Rogers, T. B. (1979). Encoding of personal information: Self-other differences. *Journal of Personality and Social Psychology*, 37, 499–514. doi: 10.1037/0022-3514.37.4.499
- Larsen, S. F., & Plunkett, K. (1987). Remembering experienced and reported events. *Applied Cognitive Psychology*, 1, 15–26. doi: 10.1002/acp.2350010104
- Layard, R., Clark, A. E., Cornaglia, F., Powdthavee, N., & Vernoit, J. (2014). What predicts a successful life? A life-course model of well-being. *The Economic Journal*, 124, F720–F738. doi: 10.1111/eoj.12170
- Levy, R., Ghisletta, P., Le Goff, J.-M., Spini, D., & Widmer, E. E. (2005). *Towards an interdisciplinary perspective on the life course (Advances in life course research, Vol. 10)*. Oxford: Elsevier.
- Matthes, B., Reimer, M., & Künster, R. (2007). Techniken und Werkzeuge zur Unterstützung der Erinnerungsarbeit bei der computergestützten Erhebung retrospektiver Längsschnittdaten. *Methoden, Daten und Analysen. Zeitschrift für empirische Sozialforschung*, 1(1), 69–92.
- Mayer, K. U. (1987). Lebenslaufforschung. In W. Voges (Ed.), *Methoden der Biographie- und Lebenslaufforschung* (pp. 51–73). Opladen: Leske & Budrich.
- Mayer, K. U. (2004). Whose lives? How history, societies, and institutions define and shape life courses. *Research in Human Development*, 1, 161–187. doi: 10.1207/s15427617rhd0103_3
- Mayer, K. U. (2009). New directions in life course research. *Annual Review of Sociology*, 35, 413–433. doi: 10.1146/annurev.soc.34.040507.134619
- Mayer, K. U., & Huinink, J. (1990). Age, period, and cohort in the study of the life course: A comparison of classical APC-analysis with event history analysis, or farewell to Lexis? In D. Magnusson, & L. R. Bergman (Eds.), *Data quality in longitudinal research* (pp. 211–232). Cambridge University Press.
- Mayer, K. U., & Schoepflin, U. (1989). The state and the life course. *Annual Review of Sociology*, 15, 187–209.
- Meulemann, H. (1990). Schullaufbahnen, Ausbildungskarrieren und die Folgen im Lebensverlauf. Der Beitrag der Lebenslaufforschung zur Bildungssoziologie. In K. U. Mayer (Ed.), *Lebensverläufe und sozialer Wandel* (pp. 89–117). Opladen: Westdeutscher Verlag.
- Moen, P. (2003). Linked lives. Dual careers, gender and the contingent life course. In W. R. Heinz, & V. W. Marshall (Eds.), *Social dynamics of the life course. Transitions, institutions, and interrelations* (pp. 237–258). New York: Aldine de Gruyter.
- Reimer, M. (2005). *Autobiografisches Gedächtnis und retrospektive Datenerhebung: Die Rekonstruktion und Validität von Lebensverläufen* (Studien und Berichte des Max-Planck-Instituts für Bildungsforschung No. 70). Retrieved from <http://edoc.mpg.de/237582>

- Reimer, M., & Matthes, B. (2007). Collecting event histories with TrueTales. Techniques to improve autobiographical recall problems in standardized interviews. *Quality & Quantity*, *41*, 711–735. doi: 10.1007/s11135-006-9021-y
- Schnell, R. (2012). *Survey-Interviews. Standardisierte Befragungen in den Sozialwissenschaften*. Wiesbaden: VS-Verlag.
- Settersten, J. R. A., & Hagestad, G. O. (2015). Subjective aging and new complexities of the life course. *Annual Review of Gerontology and Geriatrics*, *35*, 29–53. doi: 10.1891/0198-8794.35.29
- Stone, A. A., Turkkan, J. S., Bachrach, C. A., Jobe, J. B., Kurtzman, H. S., & Cain, V. S. (Eds.). (2000). *The science of self-report: Implications for research and practice*. Mahwah, NJ: Erlbaum.
- Winship, C., & Harding, D. J. (2008). A mechanism-based approach to the identification of age-period-cohort models. *Sociological Methods & Research*, *36*, 362–401. doi: 10.1177/0049124107310635
- Yang, Y., Schulhofer-Wohl, S., Fu, W. J., & Land, K. C. (2008). The intrinsic estimator for age-period-cohort analysis: What it is and how to use it. *American Journal of Sociology*, *113*, 1697–1736. doi: 10.1086/587154

About the authors

K. Drasch
Friedrich-Alexander University Erlangen-Nürnberg, Erlangen.
e-mail: katrin.drasch@fau.de

C. Kleinert
Leibniz Institute for Educational Trajectories (LifBi)
and Otto Friedrich University Bamberg.
Institute for Employment Research (IAB), Nuremberg.

B. Matthes
Institute for Employment Research (IAB), Nuremberg.

M. Ruland
Institute for Applied Social Sciences (infas), Bonn.

Collecting Life-Course Data in a Panel Design: Why and How We Use Proactive Dependent Interviewing

Annette Trahms, Britta Matthes and Michael Ruland

Abstract

The National Educational Panel Study (NEPS) has to combine the retrospective collection of life-course data with repeated competence measurements in a panel design by updating life-course information on an ongoing basis. The greatest challenge to updating life courses in a panel study is ensuring the overall consistency and completeness of the life course across multiple waves and preventing seam effects. These effects occur in the transitions between different states of interest from one panel wave to the next, and their number is much higher when the data for each period come from two different interviews than when the reports come from the same interview. To minimize this effect and to ensure that episodes collected in different panel waves are connected with each other, NEPS researchers use dependent interviewing techniques that draw on information collected in previous panel waves in order to phrase questions and direct respondents through the questionnaire. Proactive Dependent Interviewing—whereby information from the previous interview (named preload) is used to stimulate the memory as part of the questioning process—is particularly widely used because of its potential to lower respondent burden, increase efficiency, and reduce measurement errors, such as seam effects. Against the background of findings from cognitive psychology, we describe how we implemented this technique in the NEPS Starting Cohort 6-Adults. We then evaluate the quality of this kind of “anchoring” by empirically analyzing the conditions under which respondents disagree with preloaded data.

1 Introduction

In order to answer the research questions formulated in the National Educational Panel Study (NEPS), it is essential to not only interview respondents about their past lives, but also to repeatedly collect information about the ways in which these lives keep evolving (Blossfeld, & von Maurice, 2011). To do so, the approach applied in the NEPS is to repeatedly measure the competencies of one and the same individual on the one hand and to collect other information that changes over time and may not be measured reliably in retrospect, such as attitudes, subjective assessments, or expectations, on the other hand (Allmendinger et al., 2011). For this approach to work, however, it is essential to ensure that the life courses of those individuals be continued from one panel wave to the next because only then is it possible to perform a causal analysis of participation in education and educational outcomes (Blossfeld, Golsch, & Rohwer, 2007). From a methodological point of view, these requirements can only be met by combining retrospective life-course surveys with a prospective panel survey in which the life courses are continued (e.g., see Drasch et al. this volume).

In addition to measuring competencies across the entire life course, the major challenges of such an approach are making sure that the life stories collected in the panel are complete and consistent. When collecting retrospective life courses across multiple panel waves, it is especially important to use suitable methodological tools to avoid the so-called *seam effect*, a typical flaw of many panel studies (Hill, 1987; Lemaitre, 1992; Jäckle, & Lynn, 2007). The term seam effect refers to the fact that the number of changes reported to have occurred at the transition of two successive panel waves is systematically higher than it would have been had only a single interview been conducted. This is probably due to the position effect (Murdock, 1962), which is well-documented in laboratory experiments in cognitive psychology. Applied to the context of recalling life events as part of a panel study, this means that events that occurred at the beginning (*primacy effect*) or at the end (*recency effect*) of a reference period are more likely to be recalled because the position of these events serves as an anchor for memory: The oldest events (in a panel interview, those that took place at the point of the last interview) are remembered better because the interviewer's memory cue refers to that point in time, encouraging the respondent to reconstruct past happenings starting from there. Likewise, the most recent events are remembered better either because they are continuing or because only a short period of time has elapsed between event and report (retrospective interval). The seam effect, therefore, may be interpreted as the result of an underreporting of changes occurring during a reference period and an overreporting of changes occurring at the time of the interview (Rips, Conrad, & Fricker, 2003). Frequently, however, the seam effect also results from the fact that, after the panel interview, researchers are unable to decide whether a measured change does, in fact, correspond to an actual change or whether it is merely the result of the respondent's different description of the same situation. Overmeasure-

ment and overreporting both lead to a higher number of changes reported to have occurred at the transitions between panel waves.

To avoid such overmeasurement or overreporting in panel studies that repeatedly collect data from the same individuals, dependent interviewing techniques are most widely used. In Dependent Interviewing, respondents are confronted with their answers from a previous interview either because the interviewer wants to ask questions about information that does not match the original response (Reactive Dependent Interviewing, RDI) or because the interviewer wants to actively stimulate the respondents' memory (Proactive Dependent Interviewing, PDI). Both methods are capable of substantially reducing the seam effect (Brown, Hale, & Michaud, 1998; Jäckle, 2008). PDI has been shown to be the most suitable method for connecting episodes between two panel waves in life-course surveys (Hoogendoorn, 2004).

As a result, NEPS researchers implement PDI to continue life-course episodes not completed at the time of the previous interview into the subsequent panel wave. PDI is also used, however, to remind respondents of key status information provided in a previous interview and then to update this information, if necessary. We refer to the respondents' information from a previous interview as an anchor preload¹ because it is intended to trigger the respondent's memory during the interview and has to become a memory anchor. To gain a better understanding of how and under what conditions these anchor preloads could work, we report findings from cognitive psychology regarding the requirements that anchor preloads have to satisfy as well as the most effective ways of using them. Afterwards, we show how anchor preloads were used in the NEPS Starting Cohort 6-Adults. Since the NEPS survey gives respondents the opportunity to disagree with a given anchor preload, we can use these instances of disagreement to analyze the conditions to "anchor" more or less effectively. In a final summary, we conclude how successful the use of PDI techniques has turned out to be.

2 Findings from Cognitive Psychology

From a cognitive psychology point of view, answering a question about autobiographical content from the past is a constructive achievement (Sudman, Bradburn, & Schwarz, 1996). As various studies have shown, remembering an event and dating that event are two processes that are performed independently of one another

1 Additionally, respondent information from the previous interview is needed to (re-)identify the "right" person (tracking preload) and to keep each interview as short and as nonrepetitive as possible by customizing the questionnaire for each respondent in order to be able to ask questions about certain groups of persons or to formulate questions in a target-group-specific way (control preload). However, since the correct functioning of tracking and control preloads only depends on a sufficient testing of the technical processes, the focus of this chapter is on the conditions for anchor preloads to function correctly.

(Conway, 1996).² Cognitive psychologists regard memory as a process consisting of three components: encoding,³ storage, and recall. There is considerable variation in the extent to which memory content may be encoded and stored. Therefore, what can be remembered is a permanent representation stored in long-term memory reconstructed by means of selection and interpretation from mental representations of the original experience stored in working memory. The recall process is initiated by a stimulus, such as a question. The stimulus triggers a strategic search for the corresponding stored information. Recall may take place implicitly and quasi automatically (“knowing,” “coming to mind”) or may be performed explicitly (“thinking”). If no corresponding representation is found (after an appropriate amount of time), the person remembering either abandons the search or starts generating new stimuli of his or her own (see Figure 1).

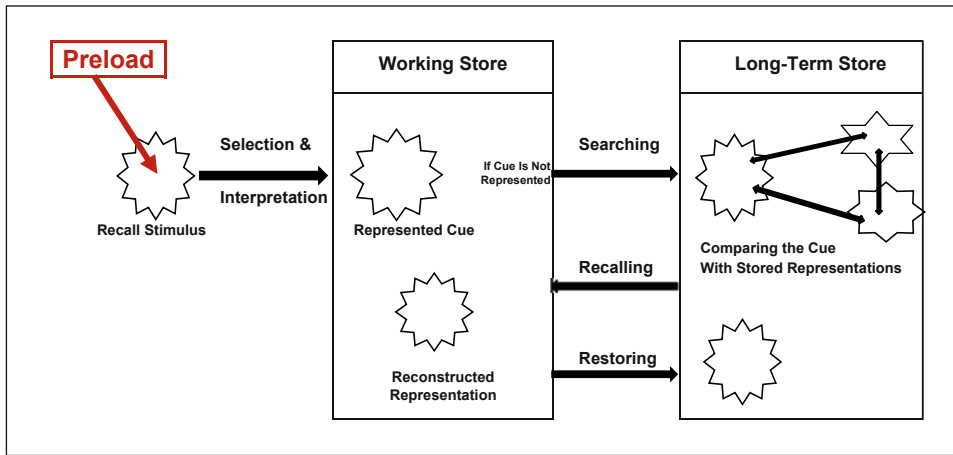
Even though it is not possible to influence what respondents encode and store or how they do so, providing a suitably designed stimulus can assist respondents with their memory process in a way that makes their memory search as likely to succeed as possible.⁴ As a result, stimuli are turned into anchors, ideally to an equal degree among all respondents. The challenge is to formulate a question in such a way that the respondent will not only understand it but also interpret it in line with the researcher’s intention (Lessler, & Forsyth, 1996). Respondents are more likely to interpret the question correctly if the things or events to be remembered are named as specifically as possible (reference content) and if the time period of interest (reference period) is demarcated as precisely as possible. Preloads should be included in the stimulus using the exact wording of the respondent’s previous answer wherever possible (open-text format). By contrast, if the wording is changed, for instance, by introducing categories, that is, identical formulations of the stimulus for a selected target group (fixed format), the preload should be less likely to anchor in the respondent’s memory.

Second, the respondent—equipped with this interpreted stimulus (cue)—has to embark on a strategic search for matching representations in his or her memory storage. Whether or not such a search for representations will be effective depends on a variety of factors, including the retrospective interval, the number and density of the things or events to be remembered, and also the time the respondent is allowed for recall or that he or she is willing to invest in the search (e. g., Loftus, & Marburger, 1983; Means et al., 1989). Since surveys take place annually in the NEPS, the retrospective interval is not that much of an issue. However, preloads should be more difficult to

2 With regard to autobiographical memory, it is rare for respondents to make up additional events. In contrast, the major problem is that respondents tend to report far too few events rather than far too many (underreporting).

3 This is necessary in order not to overburden the memory’s limited capacity for processing and storing information.

4 This is why preloads should not be edited under any circumstances since editing would make it more difficult for respondents to recognize them.

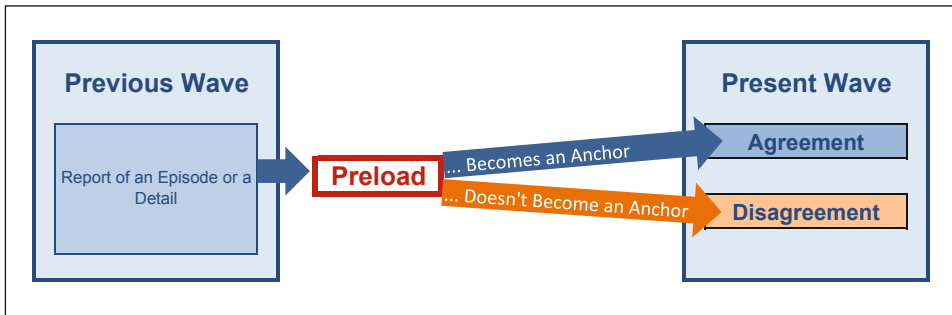
Figure 1 Reconstruction process when recalling autobiographical information

Source: Reimer, 2001, p. 26, translated and complemented

anchor if there is a high number for an event to be remembered, especially if the respective information is similar.

Third, the respondent has to assess the quality of the retrieved contents to see whether they match the specified reference content and period. The contents found in the memory storage are assessed in terms of whether they match the (interpreted) reference content and period on the one hand and in terms of the degree to which the respondent is sure of having remembered them correctly on the other hand. This assessment is made subjectively, primarily by looking at whether the retrieved content is consistent with other memories and whether it was easy or difficult to recall, as well as by assessing the amount of detail and the vividness of the memory (Sudman, Bradburn, & Schwarz, 1996). The easiest case is when the respondent is relatively quick to retrieve a memory that he or she considers to be sufficient and secure and that unequivocally falls into the requested reference period. The more the cue resembles the memory in question, the easier and faster the recall process will be. This is why preloads should not be categorized or edited under any circumstance if doing so would make it less likely for the respondent to recognize the cue. If only less-perfect and less-convincing memories are found, respondents have to decide whether they want to search their memory a second time, whether it might be possible to formulate an answer to the cue based on the imperfect memory,⁵ or whether they would rather abandon the search by disagreeing with the preload (see Figure 2).

⁵ This is typically done by applying a more "generous" interpretation of the reference content or by rough estimates.

Figure 2 Disagreement with a preload

Searching for a second time requires time and energy, but by agreeing to participate in the survey respondents have made a certain commitment to answer to the best of their ability. They can therefore be expected to try and make a certain effort to come up with at least some sort of utterable response (Schwarz, & Sudman, 1994). This has also led to the suspicion that PDI produces a certain agreement bias, meaning that respondents say that nothing has changed even though changes have, in fact, occurred. As of yet, however, this suspicion has not been able to be confirmed empirically (Jenkins, Lynn, Jäckle, & Sala, 2006). Therefore, if panel surveys provide respondents with the possibility of disagreeing with a given stimulus, this problem does not seem to be relevant.

What matters is that the process of recalling this autobiographical content is based on what respondents have experienced themselves. This self-related information has a special status in the memory-storage process and is also linked to emotional, visual, and other impressions that favor its long-term storage and retrieval (recollective memory). By contrast, if respondents are asked to provide information on third persons—partners, children, or co-workers—the recall process takes place via the recollective memory of the occasion in which the respondent first learned about the respective fact, or via having personally witnessed events in the life of the third person (Reimer, 2001).

During the interview, the interviewer must first stimulate the respondent's memory of the person about whom information is to be collected. In other words, the respondent must first be able to recognize which third person the following questions refer to. The preloads best suited to accomplish this task are those that contain as a cue the precise name or title that the respondent used in the previous interview to identify this person. This may be the person's proper name or also a term of endearment that the respondent has used to refer to the person in question. Only after this has been ensured may information about these third persons be updated. When doing so, we can expect these pieces of information about third persons to be more reliable so that respondents can connect them to their own experiences. State-

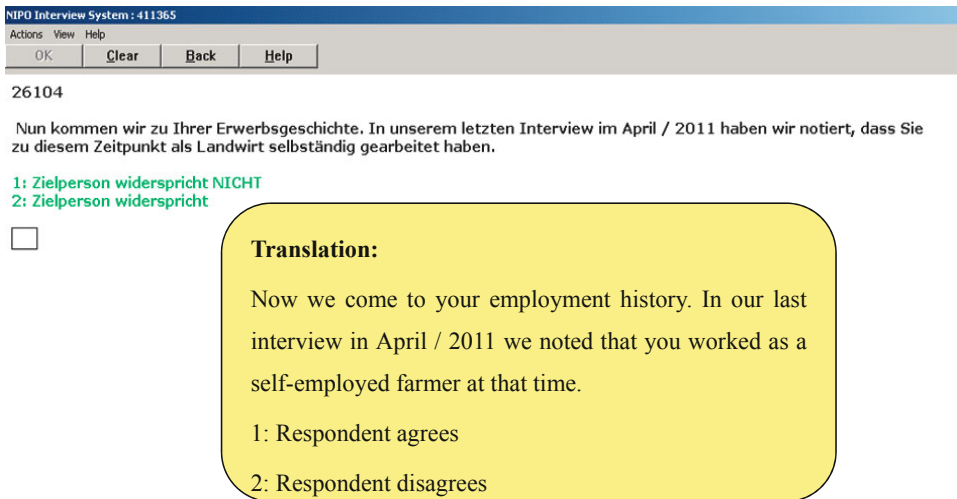
ments about the highest level of educational attainment of the respondent's partner or spouse, for example, should be most precise if the respondent was already living with that person at the time the educational credential was obtained.

3 Anchor Preloads in NEPS Starting Cohort 6-Adults

Anchor preloads are used in the NEPS to remind respondents of information they provided in a previous interview, of events going on at that point, and of relevant third persons or information about these third persons. The following description is limited to a number of anchor preloads used in NEPS Starting Cohort 6-Adults, but we will go back to the key differences in remembering that were described in the theory section and illustrate them with the help of a few typical examples. As we saw in the theory section, it is generally fair to assume that memories of one's own personal experiences are more reliable than remembered information about third persons. This is why we distinguish between these two dimensions and present the preloading approach for each in a separate section. In order to remain in keeping with the theoretical argument, it would be best to distinguish between anchor preloads implemented in an *open-text format* and those implemented in a *fixed format*. Open-text format means that the interviewer presents the respondent with a verbatim reproduction of the answer collected in the previous interview. For example, the exact wording of the occupational title given by the respondent in last interview is used to cue an employment episode. In contrast, fixed format means that the wording of the question varies depending on the respondent's membership in a certain target group. For example, in the case of temporary workers, the following wording is used: "In our last interview in <date of the last interview>, we noted that you were working as a <temporary worker> at that time." Sometimes, both formats are *combined*. For example, when interviewing self-employed, the occupational title is shown as an open text, with the additional fixed comment that this activity was recorded as self-employment at the last interview (see Figure 3).

However, these formats were used not only with regard to the recall process,⁶ but also depending on the availability of suitable open-text information, the effort required to provide the preloads, and the intelligibility of the questions. This is why we have decided to introduce an additional distinction between the key dimensions (from the methodological perspective of questionnaire design) of continuing episodes or persons from one wave to the next and the process of updating longitudinal information.

6 As verbatim repetitions of the open-ended answers from the previous interview have proven to be the best possible anchor, it would have been best to *always* collect open-ended answers and display them in the questions of the next panel wave. However, doing so would not only have required too much technical effort, but it would also have endangered the survey's level of standardization. As a result, preloads should be used sparingly and reasonably.

Figure 3 Screenshot, anchoring in employment module

Source: Detail of the questionnaire for NEPS Starting Cohort 6-Adults, wave 2009/2010, programmed version, infas 2009*

*We would like to thank infas for giving us permission to publish this screenshot.

3.1 Anchor Preloads for Updating Information on the Respondent

In NEPS Starting Cohort 6-Adults, anchor preloads are used primarily as cues for continuing one or more episodes from one panel wave to the next. In all cases, the preloads provided for the purpose of continuation refer to both the reported event itself and the date of the last interview. In the introductory question of the *unemployment* module, for example, respondents who said in the last interview that they were currently unemployed are reminded of having said they were unemployed at the time of previous interview. They are shown the fixed format wording of being unemployed as well as the date of the last interview.⁷

In the longitudinal modules, it is generally possible to also report concurrent events. When continuing into the next panel wave, respondents are reminded of these concurrent events, one after the other, in the order in which the events were previously reported. For example, if a respondent said in the last interview that he

⁷ The specific question here is: "Now we are interested in the times during which you were unemployed, regardless of whether you were officially registered as unemployed or not. Please tell us about all unemployment periods, even if they only lasted one month. In our last interview in <date of the last interview>, we noted that you were unemployed at that point."

worked as a self-employed farmer and as a packer, he is first cued about his work as a farmer and then about his work as a packer.

For the purpose of continuing episodes from one panel wave to the next, we use anchor preloads in open-text format, fixed format, or both formats combined. The decision of which of these three formats to use depends on the type of episode. Unemployment episodes, for example, are only anchored in a fixed format because this type of episode does not require a detailed, individualized description. In the employment module, the format varies depending on the type of employment and the occupational status. When interviewing self-employed individuals or freelancers, both formats are combined.

3.2 Anchor Preloads for Updating Information on Third Persons

In addition to the continuation of life-course episodes from one wave to the next and the updating of information on respondents themselves, the NEPS also collects ongoing information about respondents' living together with partners and children, as well as on their relationship status. To stimulate respondents' recollection of persons mentioned in the previous interview, the questions about partners and children also include a verbatim repetition of the name the respondent used to identify each person (mostly the person's name, or a term of endearment). In addition to this open text, questions concerning the respondent's partner or spouse also include the respective type of relationship at the point of the previous interview.⁸ Depending on the specific type of relationship, the interviewer goes on to ask questions about possible changes in this status. Married respondents, for example, are asked whether they still live together with their spouses. If multiple relationships existed at the point of the previous interview, for instance because a respondent was married but lived apart from his or her spouse in a non-marital relationship with another person, all of these partners are used as preloads in the order in which they were mentioned in the previous interview.

Likewise, all of the respondent's own children, as well as those of their partners (if the children live in the same household with the respondent), are implemented as anchor preloads in subsequent panel waves in the order in which they were mentioned in the previous interview (usually according to age, starting with the oldest and moving on to the youngest).⁹ The child mentioned first is anchored using the following question: "Now I would like to ask you a few questions about your children. In our last interview in <date of last interview>, we noted that you have a child named <name of child>." Then the interviewer goes on to ask for some details to update the

8 For example, the introductory question in the "partner" module is worded as follows: "Now I am moving on to your family. In our last interview in August 2009, we noted that you were married to and living with <name of partner> at that point."

9 Once the interviewers know that a child has died, they do not make any further references to that child either in survey questions or in anchor preloads.

information on that child. If there were other children, the following question is used for all other children: “In our last interview in <date of interview>, we noted that you also have a child named <name of child>.” What matters here is that the names of the partners and the children are included in the question in an open-text format.

4 Effectiveness of Anchor Preloads

To assess the effectiveness of anchor preloads, we analyze the extent to which respondents disagreed with the anchor preloads presented to them. As a general rule, questions in the interview are always worded in a way for the anchor preload to focus directly on the updating of an event or a piece of information. During the interview, therefore, respondents are not asked whether a certain anchor preload is correct according to their memory; rather, the anchor preload is explicitly assumed to be correct. However, respondents always have the option of disagreeing with the statement formulated in the question.¹⁰ We assume that a disagreement occurs if respondents are unable to make a connection between the anchor preload and their memories, that is to say, if the anchor preload is ineffective.

As explained in the theoretical section, anchoring should work more-or-less effectively depending on the importance and relevance of the episode to be recalled and on the format used for the anchor preloads. To examine this hypothesis, we start with a descriptive analysis of how frequently respondents disagree with anchor preloads in the various life domains (modules). Second, we look what effect the format of the anchor preload has on the frequency of disagreement. And third, to answer which format is more suitable for facilitating respondents’ recall process (the open-text format, the fixed format, or a combination of both), a multivariate analysis is conducted.

The analysis of disagreement is based on data from NEPS Starting Cohort 6-Adults, collected in the 2009–10 main survey. Even in this first wave of the survey, there were 6,495 panel cases since part of the sample had already been surveyed prior to the foundation of the NEPS as part of the “Arbeiten und Lernen im Wandel” (ALWA)¹¹ study, which was subsequently integrated into the NEPS. At least one anchor preload was used with 6,440 respondents in the first panel wave. Overall, a total of 20,079 pieces of information were used as anchor preloads. It is fair to say that using anchor

10 For pragmatic reasons, it is assumed in this case that the information or episode in question was collected correctly in the previous interview and that a subsequent change occurred. After all, when reporting current events, the retrospective interval is zero, and it is fair to assume that the current memory of past events is more flawed than is the reporting of events while they are current. With regard to collecting episodes, this means that we take the respondent’s disagreement to mean that the event in question ended at the last interview date.

11 The “Arbeiten und Lernen im Wandel” (ALWA) survey was carried out at the Institut für Arbeitsmarkt- und Berufsforschung (Institute for Employment Research) in 2007–2008 (see Antoni et al., 2011). ALWA participants who expressed their willingness to become part of the panel continue to be surveyed as part of NEPS Starting Cohort 6-Adults.

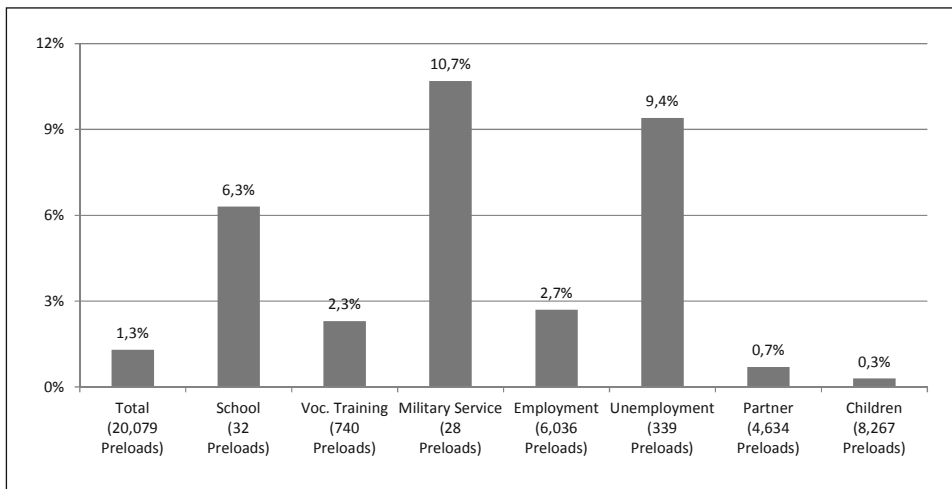
preloads turned out to be highly successful in this wave as respondents disagreed with only 1.3 % of the anchor preloads. A closer look, however, reveals very interesting differences, which are discussed below.

4.1 Disagreement by life domain (module)

Figure 4 shows the frequency of disagreement with an anchor preload, shown separately for each of the life domains in which anchor preloads were used. Disagreement with anchor preloads was especially pronounced (at about 10 %) in episodes of unemployment compared with much lower levels of disagreement in the other modules.¹² For example, the frequency of disagreement in the employment and vocational training modules, at below 3 % in each case, is considerably lower. The lowest level of disagreement is found in the modules on partners and children. Compared with the other modules, this is where anchor preloads seem to work best.

According to our theoretical assumptions, the more importance respondents attach to an episode to be recalled, the less likely they should be to disagree with the

Figure 4 Disagreement by life domain (module)



Source: SC6 w1 original data, NEPS Starting Cohort 6-Adults, own calculations

12 For the sake of completeness, Figure 3 also shows the frequency of disagreement for the military service and school modules. These should not be used as a basis for interpretation, however, since the number of anchor preloads in these modules is very small. As a consequence, they are not taken into account in the following analyses.

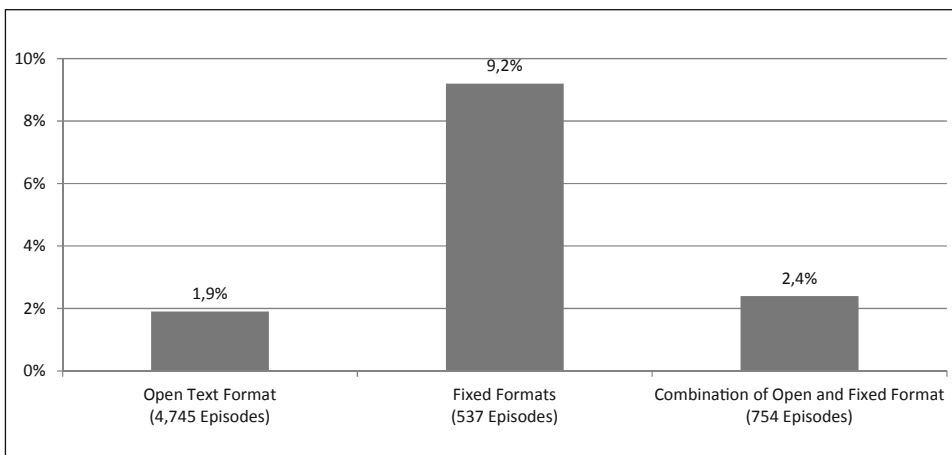
corresponding anchor preloads. The module on children is a particularly good example of this. People tend to have very strong memories of their children. After all, there is hardly any other event that has such a profound and lasting impact on our life course as the birth of our own children. Moreover, children—especially those still living in our household—usually remind us of their existence simply by being around us on a daily basis (Reimer, 2001).

4.2 Disagreement by preload format

Another possible explanation for the differences in the frequency of disagreement is that this might be due to the specific format of the anchor preloads used in the interview. Because all three formats were used by updating the employment episodes, resulting differences between the three different preload formats can be interpreted as signals for diverging effectiveness.

Figure 5 shows that respondents were much less likely to disagree with open-text anchor preloads than they were with fixed-format anchor preloads. Hardly any difference in terms of disagreement can be observed between purely open-text anchor preloads and those containing both open-text and fixed-format elements. This finding confirms our assumption that using individualized and verbatim stimuli in the anchor preloads supports respondents in making a connection to certain events in the past and thus helps reduce the likelihood of these respondents' disagreeing with a given preload.

Figure 5 Disagreement by format of anchor preload in employment episodes



Source: SC6 w1 original data, NEPS Starting Cohort 6—Adults, own calculations

4.3 The likelihood of respondents' disagreeing with anchor preloads

In sum, these descriptive outcomes confirm findings by cognitive psychologists who say that the importance of episodes or life domains is essential for them to be remembered well. However, we were also able to show that the format of an anchor preload has an influence on how well respondents can connect stimuli with their own memory. But since a specific question format is used for cueing a specific type of episode, multivariate analysis is needed to find out under what conditions respondents disagree with an anchor preload. The best way to do that is to prior test the suitability of the various types of anchor preloads for cueing different types of episodes or information in a methodological experiment. However, we didn't have enough time to do this. Therefore, in the following section, we employ a multivariate analysis. With the help of logistic regression, we investigate the extent to which the abovementioned factors of life domain and preload format have an influence on the likelihood of disagreement (Table 1).

Model 1 shows quite clearly that the likelihood of disagreement is significantly lower in the case of episodes involving partners and children than it is with episodes of employment, and if anchor preloads are used in a fixed format instead of an open-text format, respondents' odds of disagreeing are four times higher. Furthermore, it does not make any difference whether open-text formats are used alone or in combination with fixed formats.

In order to check the robustness of these results, we include additional predictors measuring importance of the episode to be recalled (if it is a main activity¹³), their relevance (duration of the episode) and the complexity of the recalled situation in Model 2. Most importantly, the impact of involving partners and children and of the preload format didn't change even if additional predictors were included. As explained above as part of the theoretical assumptions, Model 2 confirms that the more important and more relevant episodes are less likely to disagree with a given cue. Hence, if the episode to be continued is an additional rather than a main activity, the updated episode lasts longer if there is a higher total number of anchored episodes; however, if there is a higher number of anchored episodes in the same module, respondents are significantly more likely to disagree with a given preload.

What is most interesting about Model 2 is that by including these additional predictors, the likelihood of disagreement with unemployment episodes becomes significant, as well. This means that even if we take into account the fact that cues may refer to a registered or non-registered period of unemployment and that unemployment

13 In the interviews, the following episodes are defined as main activities : training episodes that respondents said were their main activity at the time (i. e., that were not undertaken in addition to a different activity), employment episodes that comprised more than 15 hours per week, and episodes during which respondents were registered as unemployed. In addition, episodes that respondents themselves afterwards classified as main activities in the examination and supplementary module were also counted as main activities.

Table 1 Logistic Regression of Disagreement with an Anchor Preload

	Model 1		Model 2	
	Odds Ratios	AME (dy/dx)	Odds Ratios	AME (dy/dx)
<i>Life domain (Ref. employment)</i>				
School	0.72	-0.0025	1.07	0.0005
Vocational training	0.78	-0.0018	1.1	0.0007
Military service	1.35	0.0022	1.96	0.0047
Unemployment	0.99	0.0001	1.60*	0.0033*
Partner	0.36***	-0.0075***	0.37***	-0.0070***
Children	0.14***	-0.0144***	0.17***	-0.0125***
<i>Format (Ref. open-text format)</i>				
Fixed format	4.03***	0.0127***	4.06***	0.0098***
Combination of open and fixed format	1.28	0.0031	1.22	0.0014
<i>Importance of activity (Ref. secondary activity)</i>				
Primary activity			0.58***	-0.0038***
<i>Duration of episode (Ref. Less than 6 months)</i>				
Between 7–12 months			1.06	-0.0004
Between 13–24 months			0.97	-0.0002
Between 25–60 months			0.85	-0.0011
More than 60 months			1.78**	0.0040**
<i>Complexity of recalled situation</i>				
Number of anchored episodes in the same module			1.23*	0.0015*
Total number of anchored episodes			1.09*	0.0006*
<i>Constant</i>				
Log likelihood	0.02***		0.01***	
		-1255		-1232
		20,079		20,079

Significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Source: SC6 w1 original data, NEPS Starting Cohort 6-Adults, own calculations; controlled for age and sex (not shown).

was anchored exclusively in a fixed format, respondents are significantly more likely to disagree with cues about unemployment episodes than with cues about employment episodes. This finding suggests that respondents generally tend to remember unemployment periods less well. One reason for this is certainly the fact that respondents do not like recalling an episode that is less accepted socially and thus unpleasant. Another reason might be, however, that respondents' subjective view of their life has changed from that at the point of the previous interview. For example, a subjective self-classification as unemployed may look more like a transition period (vacation, break) in retrospect. Another possibility might be that the respondent was, in fact, unemployed at the time of the last interview but found a new job in the same month, which is why he or she would disagree with the anchor preload. However, the analysis of what precisely causes the higher likelihood of disagreement with unemployment episodes in general must be left to further research.

5 Summary and Conclusions

The idea of the NEPS is not to have respondents report on their lives only once, but rather to collect such information on an ongoing basis, continuing life courses in each panel wave. The resulting challenge in terms of questionnaire design, therefore, is how to combine a retrospective life-course survey with a prospective panel study that allows for consistently updating all of the episodes going on at the point of the last interview. However, we know from other surveys that panel studies typically feature a seam effect. Using PDI techniques has become the most widely used approach to minimize this effect. Against this backdrop, the decision was made to use PDI to continue life-course episodes going on at the point of the last interview into the next panel wave and to update essential status information. The question is, however, what information (anchor preloads) to use from the previous interview to support respondents' recall process and how to present this information as a stimulus for respondents. This is why we considered insights from cognitive psychology in order to be better able to understand how and under what conditions such anchor preloads are most likely to become an anchor in respondents' memory rather than just a stimulus as part of the question. Looking at the extent to which respondents disagreed with the anchor preloads in the first NEPS panel wave of the Starting Cohort 6-Adults provides the opportunity to assess the preloads' effectiveness at enhancing the quality of respondents' recall. The overall finding is that the anchor preloads work quite well since respondents disagreed with only a very small fraction of the cues provided. Most importantly, using open-text preloads emerged as the best way of stimulating respondents' memory. Even though the likelihood of disagreement also decreases along with the relevance of the episodes to be continued, presenting stimuli in an open-text format was the most effective way of successfully supporting respondents with recalling past episodes. This finding, however, should not be interpreted to mean

that we should always use open-text preloads to continue episodes into subsequent panel waves. When designing questionnaires, researchers have to keep in mind that using open-text preloads is complex and thus costly. While the introduction of computer-assisted surveys has made the extensive use of PDI techniques possible, these techniques make questionnaire programming highly complex and should therefore always be used sparingly.

In this paper, we looked at the likelihood of disagreement with an anchor preload as an indicator of the effectiveness of that preload. Clearly, this approach is based on the assumption that all anchor preloads were collected correctly or at least do not vary systematically with regard to the aspects considered here. However, it is not possible to generate empirical evidence for this assumption with the data currently available. In order to make empirically informed decisions on how anchor preloads may be used even more effectively for continuing life-course episodes into subsequent panel waves, researchers should collect and analyze suitable data in the experimental design.

References

- Allmendinger, J., Kleinert, C., Antoni, M., Christoph, B., Drasch, K., Janik, F., ... Ruland, M. (2011). Adult education and lifelong learning. In H.-P. Blossfeld, H.-G. Roßbach, & J. von Maurice (Eds.), *Zeitschrift für Erziehungswissenschaft, 14. Education as a lifelong process: The German National Educational Panel Study (NEPS)* (pp. 283–299). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Antoni, M., Drasch, K., Kleinert, C., Matthes, B., Ruland, M., & Trahms, A. (2011). ALWA—New life course data for Germany. *Schmollers Jahrbuch, 131*(4), 625–634.
- Blossfeld, H.-P., & von Maurice, J. (2011). Education as a lifelong process. In H.-P. Blossfeld, H.-G. Roßbach, & J. von Maurice (Eds.), *Zeitschrift für Erziehungswissenschaft, 14. Education as a lifelong process. The German National Educational Panel Study (NEPS)* (pp. 19–34). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Blossfeld, H.-P., Golsch, K., & Rohwer, G. (2007). *Event history analyses with Stata*. Mahwah, NJ: Lawrence Erlbaum Association.
- Brown, A., Hale, A., & Michaud, S. (1998). Use of computer assisted interviewing in longitudinal surveys. In M. P. Couper, R. B. Baker, J. Bethlehem, C. Z. F. Clark, J. Martin, W. L. Nicholls II, & J. M. O'Reilly (Eds.), *Computer assisted survey information collection* (pp. 185–200). New York: Wiley.
- Conway, M. A. (1996). Autobiographical knowledge of autobiographical memory. In D. C. Rubin (Ed.), *Autobiographical knowledge and autobiographical memories* (pp. 67–93). Cambridge, MA: Cambridge University Press.
- Hill, D. H. (1987). Response errors around the seam: Analysis of change in a panel with overlapping reference periods. In American Statistical Association (Ed.), *Proceedings of the survey research methods section* (pp. 210–216). Washington, DC: American Statistical Association.

- Hoogendoorn, A. W. (2004). A questionnaire design for dependent interviewing that addresses the problem of cognitive satisficing. *Journal of Official Statistics*, 20(2), 219–232.
- Jäckle, A. (2008). Dependent interviewing: Effects on respondent burden and efficiency of data collection. *Journal of Official Statistics*, 24(3), 411–430.
- Jäckle, A., & Lynn, P. (2007). Dependent interviewing and seam effects in work history data. *Journal of Official Statistics*, 23(4), 529–551.
- Jenkins, S. P., Lynn, P., Jäckle, A., & Sala, E. (2006). The effects of dependent interviewing on responses to questions on income sources. *Journal of Official Statistics*, 22(3), 357–384.
- Lemaitre, G. (1992). *Dealing with the seam problem for the survey of labour and income dynamics. SLID* (Research Paper No. 92-05). Ottawa: Statistics Canada.
- Lessler, J. T., & Forsyth, B. H. (1996). A codier system for appraising questionnaires. In S. Sudman, N. M. Bradburn, & N. Schwarz (Eds.), *Thinking about answers: The application of cognitive processes to survey methodology* (pp. 259–291). San Francisco, CA: Jossey-Bass.
- Loftus, E. F., & Marburger, W. (1983). Since the eruption of Mt. St. Helens, has anyone beaten you up? Improving the accuracy of retrospective reports with landmark events. *Memory and Cognition*, 11(2), 114–120.
- Means, B., Nigam, A., Zarrow, M., Loftus, E. F., & Donaldson, M. (1989). *Autobiographical memory for health-related events* (Vol. 2). Hyattsville: US Department of Health and Human Services.
- Murdock, B. B., Jr. (1962). The serial position effect of free recall. *Journal of Experimental Psychology*, 64(5), 482–488.
- Reimer, M. (2001). *Die Zuverlässigkeit des autobiographischen Gedächtnisses und die Validität retrospektiv erhobener Lebensverlaufsdaten: Kognitive und erhebungspragmatische Aspekte* (Materialien aus der Bildungsforschung No. 71). Berlin: Max-Planck-Institut für Bildungsforschung.
- Rips, L. J., Conrad, F. G., & Fricker, S. S. (2003). Straightening the seam effect in panel surveys. *Public Opinion Quarterly*, 67(4), 522–554.
- Schwarz, N., & Sudman, S. (1994). *Autobiographical memory and the validity of retrospective reports*. New York: Springer.
- Sudman, S., Bradburn, N. M., & Schwarz, N. (1996). *Thinking about answers: The application of cognitive processes to survey methodology*. San Francisco, CA: Jossey-Bass.

About the authors

B. Matthes
Institute for Employment Research (IAB), Nuremberg.

M. Ruland
Institute for Applied Social Sciences (infas), Bonn.

A. Trahms
Institute for Employment Research (IAB), Nuremberg.
e-mail: annette.trahms@iab.de

Data-Revision Module—A Beneficial Tool to Support Autobiographical Memory in Life-Course Studies

Michael Ruland, Katrin Drasch, Ralf Künster, Britta Matthes and Angelika Steinwede

Abstract

The key objective of the National Educational Panel Study (NEPS) is to enable analyses of the development of competencies, educational processes, educational decisions, and returns to education throughout the life span. These analyses are only possible by collecting complete and consistent educational and employment histories as well as the relevant contexts in which these histories are embedded, in other words, life-course data. In this respect, the most important challenge is remembering life histories retrospectively. To support both the cognitive memory capacity and the temporal integration of reported episodes in the life course, we decided to use a modular technique for collecting life-course data retrospectively. However, modularization makes it more difficult for respondents to recall the temporal integration of the episodes reported in the different life domains. To compensate for this disadvantage, we implemented a data-revision module that integrates all reported episodes from the different life domains immediately after collecting all relevant life-course data. In the data-revision module, the interviewer can edit all existing temporal inconsistencies in the life course in collaboration with the respondent by correcting the time span of episodes, by deleting and inserting episodes, and by clarifying overlaps of episodes. The module also pays attention to episodes with incomplete or missing calendar dates that can—by using estimates—be included in the life-course data and be edited directly during the interview in collaboration with the respondent. The result is a marked improvement in the data quality and validity of the recorded life histories.

1 Introduction

Event-history data can provide valuable information for the analysis of educational processes, the development of competencies, returns to education throughout the life course, and long-term social change (Blossfeld, & von Maurice, 2011). One major challenge of collecting event-history data retrospectively is that researchers have to rely on respondents' autobiographical memories. However, autobiographical memory has often been shown to be inaccurate (e. g. Auriat, 1993). That is why post-interview data editing was necessary in past retrospective life-course surveys, such as the German Life History Study (GLHS), to check for inconsistencies related to memory problems (Brückner, Hoffmeyer-Zlotnik, & Tölke, 1983; Mayer, Papastefanou, & Tölke, 1989). This editing was lengthy and costly because each single case had to be reviewed after the interview by trained editors who used standardized consistency and completeness checks (cf. Tölke, 1989). Furthermore, it was not always possible to derive comprehensive and unambiguous editing rules without contacting the respondent again (cf. Hillmert, 2002). To improve data quality and to guarantee the cost efficiency of the survey, therefore, one challenge was to more accurately support a respondent's autobiographical memory with the help of a specific questionnaire design (cf. Matthes, Reimer, & Künster, 2007).

One major question that arises is how to design such a questionnaire. To answer this question, we begin with some brief insights into the functioning of human autobiographical memory. Second, we argue that combining modularized self-reports and event-history calendars helps optimize the use of several retrieval strategies from autobiographical memory organization to arrive at "better" data. Third, we show the design of the data-revision module in the NEPS Starting Cohort 6—Adults as an example. In this module, the interviewer is able to edit all existing temporal inconsistencies in the life course in collaboration with the respondent. Fourth, we empirically assess what type of data problems benefit most from the data-revision module and investigate whether this module really improves completeness and dating accuracy in the study. Finally, a summary and some practical conclusions and suggestions for further research are given.

2 Theoretical Background: Autobiographical Human Memory¹

The design of the NEPS questionnaire is based mainly on five central insights from cognitive psychology. First, autobiographical memory is organized as a network of mental representations residing in the long-term memory of an individual. When a

1 For more detailed insights into cognitive processes in retrospective self-reports, see Tourangeau (2000); for remembering and dating events, see Bradburn (2000); for an overview of recall problems in retrospective surveys, we recommend Reimer and Matthes (2007).

survey question is asked as a recall stimulus, the required information is reconstructed by searching this network of mental representations for information that matches this stimulus to a sufficient degree (e.g., Conway, 1996).

Second, representations in autobiographical memory are grouped top-down (Conway, & Pleydell-Pearce, 2000). At the top level, “lifetime periods,” meaning episodes within a thematic domain (such as “my time in firm X”), are stored. The intermediate level consists of information on shorter sub-episodes (such as “initially I worked part-time”) or special events (such as a job interview) and recurring episodes or events (“coffee breaks at company X”). The lowest level contains detailed, event-specific knowledge.

Third, representations share information and are interconnected by pathways: Hierarchical pathways connect respondents’ broad general memory to more detailed memory. Sequential pathways use the chronological order of events to relate autobiographical memory. Parallel pathways use the fact that events can occur at the same historical time and can thus be connected (Barsalou, 1988; Conway, 1996). Collecting valid retrospective information on autobiographical events requires a questionnaire that stimulates the retrieval of events by using all three types of pathways.

Fourth, a central principle of autobiographical memory organization is the temporal order of events (Conway, 1996). Quite often, events are not time-tagged because knowledge about events is usually not stored together with information about the timing of the event (Huttenlocher, Hedges, & Bradburn, 1990). In contrast to the events themselves, dates are not so much explicitly recalled but rather inferred from an event’s biographical context by relating a reconstructed event to one or more private landmark events, meaning dates the respondent is aware of, such as birthdays, or dates the respondent has already reconstructed (Friedmann, 1993; Larsen, Thompson, & Hansen, 1996).

Fifth, all represented and reconstructed episodes and transitions are subjective constructions (Neisser, 1988). In order to provide a sense of identity and biographical meaning, representations are organized into a life story within a framework of normative expectations about biographies. Respondents tend to adjust their life course in such a way that it becomes more conventional and more consistent with the individual’s self-perception at the time of recall (e.g. Barsalou, 1988). Essentially, this means that individuals idealize and smooth over their life courses, leaving out episodes or re-organizing the order of episodes (Conway, & Pleydell-Pearce, 2000; Reimer, & Matthes, 2007).

In sum, retrospective reports about respondents’ life courses are a reconstructive cognitive process that is based on memory representations stored in human memory and guided by recall stimuli that specify what and how information has to be reconstructed. Furthermore, retrospective reports have to be consistent with previously stored information and individual or normative notions of their biography. Completeness and accuracy, therefore, are severely at risk when collecting life-course data retrospectively. Effective interview techniques and tools must be designed to prevent

errors and biases and to make sure that survey reports are as complete and correct as possible.

3 Collecting Life-Course Data by Combining Modularized Self-reports and Data Revision

When collecting life-course data, the central unit is the respondent's event history. It is characterized by the episodes it contains (e. g., schooling or employment episodes) and these episodes' respective dating. A large amount of temporal information is derived from this basic structure, such as the frequency, incidence, timing, pacing, and duration of life events. As mentioned above, three retrieval pathways exist to gather detailed information on event histories: hierarchical, sequential, and parallel pathways. However, parallel retrieval is difficult to standardize because the underlying memory processes are rather individual. Therefore, the two other pathways are frequently used to collect life-course data retrospectively.

Studies that primarily stimulate sequential memory pathways (e. g., SHARELIFE, the life history calendar of the Survey of Health, Ageing and Retirement in Europe) have to reconstruct an individual's entire life course by repeatedly asking "What happened next?" and then recording the details of this episode. This approach brings a respondent's whole biographical context to mind, supporting the chronological retrieval of events along the historical timeline. However, there are several problems that limit the completeness and the level of standardization of these kinds of reports (cf. Barsalou, 1988). Reconstructing the life course along a single timeline delegates the responsibility of deciding what specific episode to remember next to the respondents. Because the conceptualization of a specific episode type is often ambiguous, this decision is likely to be different in repeated interviews or when interviewing different respondents in the same survey. For example, all respondents who receive the recall stimulus "schooling" are very likely to have the same idea of what is meant by this term. What is meant by "training," however, is much less evident because some types of training, such as an apprenticeship in a firm, may also be interpreted as employment. Respondents might have different things in mind when the stimulus "training" is given, and the interviewer has almost no possibility to control the respondent's interpretation. Moreover, specifying the criteria that define the start and the end of an episode is problematic in sequential questioning. This approach makes it difficult for researchers to communicate their episode concept. The respondent often does not know when a new episode should be reported because often only key-words are used to define different types of episodes. In sum, the level of standardization is low when life-course data are collected sequentially.

Additionally, referring only to respondents' sequential retrieval strategy produces fewer reported episodes than actually occurred (Belli, 1998). When going through a respondent's life sequentially, short or seemingly unimportant episodes are more like-

ly to be left out, and minor changes, such as promotions, are less likely to be reported (Reimer, Matthes, 2007). More importantly, it becomes easier for the respondent to omit distressing or unpleasant time periods (e. g. Drasch, & Matthes, 2013). The completeness of the retrospective self-reports is thereby threatened, especially if an episode that only comprises a few hours per week (e. g., a language course) can close an unfilled time period even though the respondent was unemployed at the same time. Considering the increasing complexity of individual life courses, this problem has become more important in recent years (e. g. Brückner, & Mayer, 2005). Temporal overlap with previously reported episodes or two concurrent episodes starting at the same time make it increasingly difficult for respondents to identify the next episode to be reported. As a result, after reporting the details of a specific episode, the respondent often does not recall the other, concurrent episode, which then remains unreported.

To overcome these disadvantages, we decided to rely primarily on hierarchical memory pathways and to combine them, in a second stage, with sequential retrieval strategies. More precisely, we split up the life course into different life domains (that are of interest to the researcher, such as schooling, employment, or partnership). First, we begin each module with a short explanation regarding the types of events that are the central scope of the module and those that are not to be collected in the module. We also explain how start and end dates of the episodes are defined in the specific module and which time period we are interested in (in the first panel wave: the period between school enrollment and the interview date; in the second and later panel waves: the period between the previous interview and the current one). In each module, we start with the question “Have you ever been ...?” in the first loop and continue with “Were you ... a subsequent time?” in the following loops. Because interviewing techniques should contextualize recall and encourage multiple retrieval strategies, we stimulate parallel pathways in some of the more complex modules by addressing several subdomains of a module. In the employment module, for example, we first collect information about regular jobs and then explicitly address secondary jobs. When the respondent cannot recall additional episodes in a life domain, the interviewer continues to the next life domain. This procedure is called modularized self-reporting and largely avoids the aforementioned “smoothing” of life courses. Moreover, by giving the possibility to report more than one episode per time unit, the approach avoids omitting parallel or overlapping sequences (Reimer, & Matthes, 2007).

However, modularization has certain drawbacks, as well. For example, this approach does not stimulate parallel recall pathways connecting different life domains (such as remembering employment interruptions by recalling the date of giving birth to a child) or sequential recall pathways linking events in and between life domains (such as starting an apprenticeship after finishing school). Furthermore, modularized life-course questionnaires leave some time periods unfilled because some types of episodes (such as illness or housekeeping, so-called gap episodes) are not modularized and therefore not collected in the first part of the interview. As a consequence, the interviewers are not aware of missing periods because they lose track of the entire life

course while collecting respondents' life events in different modules. Furthermore, because episodes are often not time-tagged, time periods containing contradictory temporal information may occur. After finishing the modularized self-reports, it is unclear whether the collected episodes indeed form a consistent, plausible, and complete life course.

These problems can be solved by implementing an additional module to sort and analyze all reported episodes sequentially, giving visualized feedback to the interviewer about missing periods and temporal inconsistencies in the collected event histories and providing tools to complete or correct them. Therefore, the modularized self-reports are combined with the idea of event-history calendars (EHC). Traditionally, EHC interviewing has helped gather life-course data sequentially (Freedman et al., 1988) by supporting the use of parallel memory pathways (cf. Belli, 1998). However, the EHC is hardly suitable for large-scale Computer Assisted Telephone Interviews (CATI) because when speaking on the phone, the calendar only serves as a reference point for the interviewer and not for the respondent; it does not use standardized question formulation; and its application would require experienced, skilled, and therefore expensive interviewers. As a consequence, we do not use EHC interviewing techniques in the first stage of data collection. We do use them in the second stage, however, as a data-revision module. In the following chapter, we describe how we aim to support the application of all three retrieval strategies while simultaneously keeping a high level of standardization by implementing this module.

4 The NEPS Data-Revision Module

The data-revision module is based on developmental work in the framework of the German Life History Study (GLHS) at the Max-Planck-Institute for Human Development (Matthes, Reimer, & Künster, 2007; Reimer, & Matthes, 2007). At the Institute for Applied Social Sciences (infas), the basic principles of the data-revision module of the survey "Working and Learning in a Changing World" (Antoni et al., 2011) have been implemented as an adapted application. This application has been transferred to all NEPS stages collecting life-course data (Hess, Steinwede, & Schneider, 2012). In the NEPS, it was first used in the survey questionnaire of Starting Cohort 6—Adults (Allmendinger et al., 2011).

The central purpose of the data-revision module is to check life-course data for completeness and temporary consistency. To perform these checks, reported episodes are merged and displayed in a historical timeline. In addition to showing overlapping episodes, the data-revision module identifies unfilled time periods between reported episodes and between interview dates and reported episodes. Afterwards, interviewers are guided by scripted questions prescribing the identified data problem and unfilled time periods, and overlaps can be clarified based on the respondents' answers by adding missing episodes or correcting dates. To ensure standardized re-

porting, dependent interviewing strategies (see Trahms et al. in this volume) are used. Additionally, data revision is facilitated by visualization. Below, the way in which the data-revision module works is documented in detail.

4.1 First Step: Establishing a Chronological Order for Modularized Collected Episodes

First, the data-revision module orders the episodes that are collected in the modules chronologically. Technically speaking, all episodes are sorted according to their starting (and, if necessary, to their ending) dates along the historical timeline. If data on the starting month and year (as well as on the ending month and year) are non-missing and exactly determined on a monthly basis for all episodes, everything works well. However, previous research has shown that respondents in retrospective interviews often have problems remembering exact dates (Reimer, 2005). Therefore, in the interview, vague dating of the month in which the episode started or ended was allowed, for example, in terms of seasons, with “beginning of the year” or “end of the year.” To include these vague dates in a chronological order, they have to be replaced with appropriate date estimates prior to the sort sequence.²

Nevertheless, allowing respondents to report vague month information still fails to avoid missing values in the dating of episodes. In this case, the data-revision module performed a specific sorting algorithm: Even if episodes did not contain any information on the year in which the episode started or ended and the year for the start date or end date of an episode is thereby missing, it is possible to estimate the episode’s dating on the basis of complementary dates in the respondent’s life. Only if both start and end dates are missing is it impossible for them to be placed on the timeline without additional information. In such cases, before starting the chronological sorting, the respondent is asked to place the episode on the timeline of his or her life course in relation to other episodes (by asking if the episode started before or after the beginning of another episode in the life course). If the respondent is not able to place the episode on the timeline, it is excluded from the chronological sorting and therefore from the test of the life course.

² The information “beginning of the year” was replaced by January, “end of the year” by December, “spring” by April, “summer” by July, “autumn” by October, and “winter” by January.

4.2 Second Step: Revision of Completeness and Temporal Consistency

After chronological sorting, the data-revision module checks the consistency and completeness of the event histories by analyzing each transition from one episode to the next, starting with the earliest episode.³ The data-revision module distinguishes four test results:

- Test Result 1: The two successive episodes are connected without an unfilled time period or overlap.
- Test Result 2: There is an unfilled time period between the two episodes.
- Test Result 3: The successive episodes overlap, or the following episode is temporally embedded in the previous episode.
- Test Result 4: Because of missing dates, whether or not the successive episodes overlap, whether or not they connect perfectly, or whether or not there is an unfilled time period between them cannot be determined.

If there is neither an unfilled time period nor an overlap between two successive episodes (Test Result 1), it is assumed that the dates of this transition have been collected correctly. Thus, it is not necessary to question the respondent for any corrections with respect to these episodes.

If an unfilled time period has been detected between successive episodes (Test Result 2), the data-revision module generates a scripted question containing information about the unfilled time period, the type and end date of the previous episode, and the type and start date of the following episode. In collaboration with the respondent, the unfilled time period can be closed by changing the dates of the involved episodes or by collecting one or more new episodes. For changing dates, the interviewer can either alter the end date of the previous episode or the start date of the following episode, or both. When collecting a new episode, the type of episode can be selected from the predefined list of the questioned longitudinal modules, and full information about the new episode can then be collected in the same way as it was collected in the original module. Alternatively, the new episode can be an episode that is not defined in one of the longitudinal modules (gap episodes). In this case, information about start and end dates and the type of activity are collected. Afterwards, the data-revision module adds the new episode to the chronological list of the respondent's life course and tests again for temporal consistency and completeness.

If the data-revision module discovers a temporal overlap of episodes (Test Result 3), a scripted question is presented to verify the correctness of the overlap. The overlap can be confirmed by the respondent and registered as approved and valid. If the respondent declares that the overlap of the episodes is not correct, the temporal over-

3 The earliest time point at which the NEPS data-revision module takes place is the first month after the person's ninth birthday.

lap has to be eliminated by correcting the dating of the overlapping episodes. After the dates have been corrected, the life course is tested for temporal consistency again.

If it is impossible to decide whether there is a direct temporal connection, an overlap, or an unfilled time period between two successive episodes due to missing date information (Test Result 4), the respondent is made aware of the problem. The respondent is then asked if these episodes directly succeeded each other chronologically. If the respondent confirms this, the missing date is replaced with the exact date of the complementary episode. However, if the temporal succession is not confirmed, the respondent is asked if there was an unfilled time period between the two episodes. If the respondent confirms this, the same procedure as the one described above with regard to filling unfilled time periods is performed (see Test Result 2). If the respondent disagrees, the two episodes are assumed to overlap, and a corresponding estimation of the dates is made.

In order to support respondents' recall as naturally as possible, we also allow a less standardized form of correcting life-course data in the data-revision module. Respondents often spontaneously recall episodes or dates that they had temporarily forgotten during the modular collection of episodes. As a result, the data-revision module also contains flexible options to directly correct the life courses beyond the strictly pre-set order of scripted testing questions. If respondents' recall requires this, the interviewer is allowed to collect new episodes, delete or reject already-collected episodes, and change the dating of the episodes independently of the standardized routines. These changes are also included recursively in the data-revision process.

The data revision is carried out until all problems have been addressed and solved in collaboration with the respondent. Thereafter, the data-revision module signals that the life-course collection and revision has been completed.

5 Effectiveness of the Data-Revision Module

To describe the effectiveness of the data-revision module, estimations, corrections, and completions (henceforth referred to as data modification) in the NEPS Starting Cohort 6—Adults are analyzed. Based on the methodological data file of the data-revision module, which includes information on whether an episode has been deleted or added and whether dates have been estimated and corrected, the following descriptions begin with a short overview of the reported episodes and the proportion of overall data modifications. In the NEPS Starting Cohort 6—Adults, 11,649 respondents reported 69,278 episodes. Table 1 reveals the total number of reported episodes along with the number and proportion of data modifications.

Reviewing the proportion of data modifications with regard to all reported episodes, it is clear that 28% of episodes have been modified. If considering the number of reported episodes for women and men separately, only small differences emerge even though a slightly higher proportion of data modifications can be reported for

Table 1 Number of Data Modifications

		Total number of episodes	All data modifications		Data modifications (with- out gap episodes)	
			Number	Percent	Number	Percent
Men	Age: 23–35	5,453	1,087	19.9%	840	15.4%
	Age: 36–50	8,908	1,881	21.1%	1,495	16.8%
	Age: 51–65	20,021	6,341	31.7%	4,937	24.7%
	Total	34,382	9,309	27.1%	7,272	21.2%
Women	Age: 23–35	5,001	1,165	23.3%	783	15.7%
	Age: 36–50	10,373	2,487	24.0%	1,580	15.2%
	Age: 51–65	19,522	6,598	33.8%	4,157	21.3%
	Total	34,896	10,250	29.4%	6,520	18.7%
Total		69,278	19,559	28.2%	13,792	19.9%

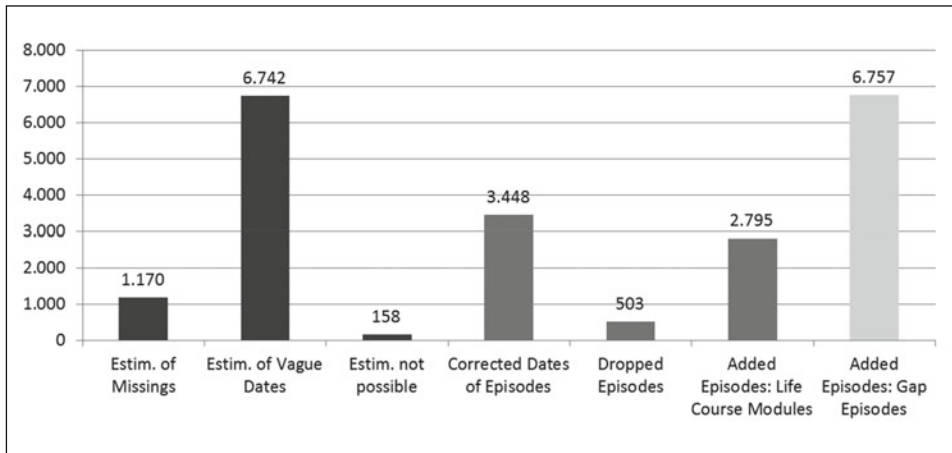
Note. Source: NEPS Starting Cohort 6—Adults, methodological data file of the data-revision module, own calculations.

women. However, since the data-revision module is also used to add gap episodes (which cannot be captured through the predefined longitudinal modules), the number of data modifications that can be ascribed to the implementation of the data-revision module should be calculated without gap episodes. By excluding gap episodes, 20 % of all episodes have been modified. Here, interestingly, the gender difference is reversed. Excluding gap episodes led to a slightly higher proportion of data modifications in men's event histories. Looking at differences between the age groups confirms previous research (e.g. Peters, 1988). Data modifications were primarily caused by the time difference between the interview and the retrospectively recalled past events. In sum, given the assumption that these changes improve data quality, the data-revision module is necessary to collect high-quality life-course data.

5.1 What Kinds of Errors Will Be Corrected in the Data-Revision Module?

To gain an insight into the different kinds of data corrections, the following description focuses on the different types of data modifications that arise during the data-revision module (see Figure 1).

The first step in the data-revision module is the estimation of dates. Accordingly, the first three bars of Figure 1 show how many dates for the start and end times of episodes were estimated. More than 1,000 start or end dates, initially reported as missing

Figure 1 Types of estimations and corrections

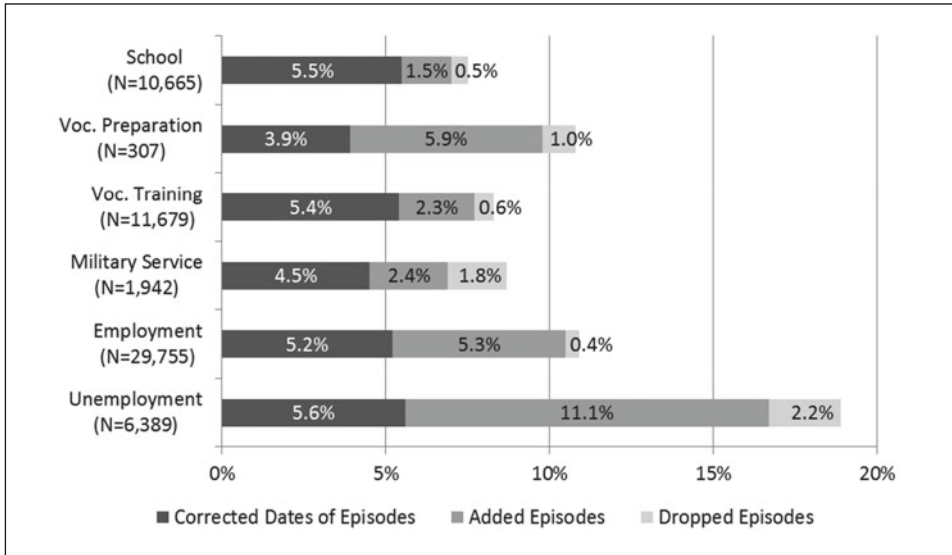
Note. Source: NEPS Starting Cohort 6—Adults, methodological data file of the data-revision module, own calculations

in the longitudinal modules, were estimated with the help of the data-revision module. Additionally, both the estimation of vague dates in more than 6,700 cases and the fact that only 158 start or end dates could not be automatically estimated give an impression of the great advantage of using the data-revision module.

Based on their original or estimated dates, episodes were ordered chronologically and checked for temporal completeness and consistency during the second step of the data revision. Corresponding to the arising test results, life-course data were modified (see the fourth to sixth bars of Figure 1). The start or end dates of nearly 3,500 episodes were corrected due to overlaps or unfilled time periods in the life courses. About 500 episodes were subsequently dropped because the respondent stated that the episode had been incorrectly reported. Additionally, by identifying unfilled time periods in the life course, almost 2,800 subsequent activities were captured by one of the pre-defined longitudinal modules. For the sake of completeness, Figure 1 additionally reports that nearly 6,200 gap episodes were added by using the data-revision module to fill time periods in the life course not pre-defined by one of the longitudinal modules.

5.2 What Episodes Benefit Most From Using the Data-Revision Module?

During the data-revision process, various data modifications took place that were induced either by the module itself (automatic estimations) or by the respondents' answers during the data revision (manual corrections). Now, we can take a closer look at the different types of modified episodes. For this purpose, we exclude the automati-

Figure 2 Corrections within type of episode

Note. Source: NEPS Starting Cohort 6—Adults, methodological data file of the data-revision module, own calculations

cally conducted estimations and focus on the three data-revision procedures: correcting dates and adding or dropping episodes. In the following two figures, we compare the data revisions made within and between the different longitudinal modules. This is important for getting an idea of what episodes are affected by and therefore benefit from the data-revision process most. Figure 2 shows the respective proportions of the three types of data revisions and their relative frequency compared with all episodes reported in each module. For instance, the dating was corrected in more than 5% of the unemployment episodes, 11% of the unemployment episodes were added, and 2% were dropped.

As a result, correcting the dating seems not to be a function of module affiliation because of the dates' random distribution over the different module types. Moreover, with respect to dropped episodes, it seems that the episode type is not as important as expected. Even if the proportion of dropped unemployment episodes is higher than that of other types of episodes, the number of dropped episodes is too small to allow for drawing valid consequences.

However, considering the quantity of episodes added, episodes of unemployment are primarily more likely to be added to the data-revision module. In line with previous research, unemployment episodes have by far the largest relative probability of being omitted or forgotten in the modularized questionnaire and of being added later on (e.g. Drasch, & Matthes, 2013). Collecting unemployment episodes seems to benefit most from the data-revision module because these modules are now considerably

more complete than they would have been if they had not been made use of. For employment episodes and episodes of vocational preparation, the probability of making additions is also quite large at approximately 5 % or 6 % but amounts to only half of the likelihood of unemployment episodes. Other types of episodes, such as school, vocational training, and military service, are less likely to be forgotten. Short and less-important episodes are particularly difficult to remember retrospectively, which has been confirmed by previous research with respect to unemployment episodes (e.g. Dex, & McCulloch, 1998). This indicates that the data-revision module seems to be important for episode types that are less bound to institutions or are less salient to the respondent.

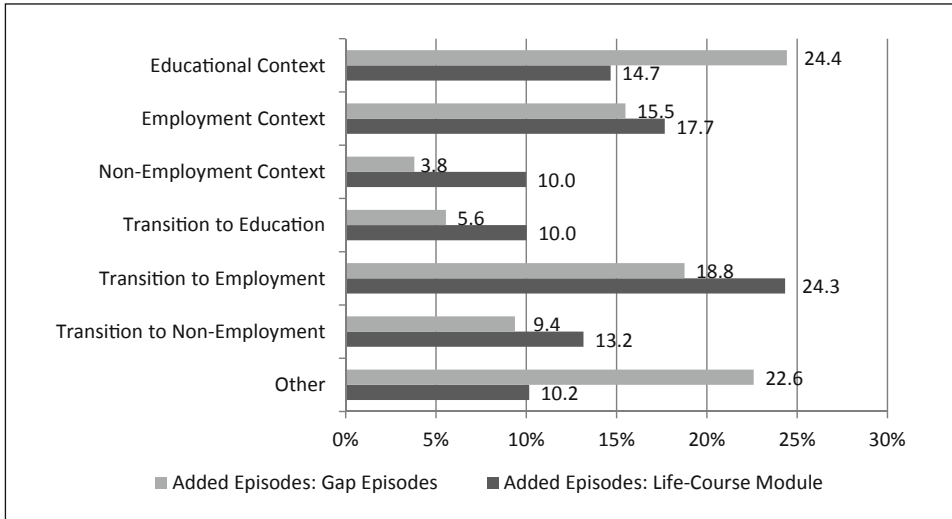
5.3 What Type of Data Problems Will Be Prevented by Using the Data-Revision Module?

Building on the previous results, we want to take a closer look at the life course and answer the question of which situation in life is most error prone due to initially omitting and forgetting episodes and why it is important to collect these episodes. To do so, we consider respondents' whole life courses and analyze the specific situations in which episodes were added. We distinguish between several contexts: Educational context refers to the period an educational episode was reported by the respondent before and after an unfilled time. Employment context (as well as non-employment context) is defined in the same way, meaning that employment (or non-employment) occurred before and after the unfilled time period. Transitions to education refers to unfilled time periods that ended by the start of an educational episode, and transition to employment (and transition to non-employment) is defined as an unfilled time period that ends in employment (or non-employment), both regardless of which episode was reported before the unfilled time period began.⁴

The results of Figure 3 show that 24 % of the added gap episodes and nearly 15 % of the added episodes that were initially forgotten or omitted (added life-course module episodes) occurred in educational contexts. Since adding gap episodes is not supposed to be reported in a life-course module (like periods between different educational steps, e.g., graduating from school and entering university, vacations), it is not surprising that adding gap episodes is especially relevant in the educational context. By using the data-revision module, the underreporting of such episodes, which is crucial to know for analytical purposes, is prevented.

Added episodes, which are supposed to be reported in a life-course module, suggest recall problems. Figure 3 strongly indicates that episodes at transitions to em-

4 In addition to these categories, some added episodes are found at the beginning or end of the time period checked in the data-revision module, and these episodes are added to the subsequent category.

Figure 3 Context before and after added episodes

Note. Source: NEPS Starting Cohort 6—Adults, methodological data file of the data-revision module, own calculations

ployment, in particular, are often initially forgotten or omitted by the respondent. Using the data-revision module stimulates the parallel recall pathway by giving respondents the opportunity to remember what had happened immediately before a specific event (e. g., taking up a new job), regardless of which life-course module the event took place in. Thus, the contextualized cue regarding the episodes that took place before and after the unfilled time period improves autobiographical recall.

6 Summary and Practical Conclusions

Event-history data as collected in the NEPS are an important data source for analyzing educational histories and their embedding in a social structure. However, to provide reliable results, event-history data have to be complete and consistent. Insights from cognitive psychology suggest that retrospective data collection should be organized in a way that provides memory cues and stimulates different memory pathways for recalling retrospective information. Three different pathways can be distinguished: parallel, sequential, and hierarchical pathways. However, due to several memory problems, such as respondents' intentionally or unintentionally adjusting their life courses to what is considered a normal biography, completeness and consistency are severely at risk.

In order to avoid the aforementioned problems, the NEPS life-course questionnaires are designed in a special manner. Modularizing is used to aim at collecting

complete and consistent life-course data by avoiding omitting parallel or overlapping sequences. Therefore, the life course is spilt up into several thematic domains, whereby the questionnaire within each domain begins with the first episode of its kind, for example. For first school or first job, the questionnaire asks for the start date, end date, and a number of detail variables, and it then progresses through all episodes of this kind. By providing cues about the definition of episodes focused on in this module and specifying episode types that are typical of this module, hierarchical recall pathways are stimulated. However, in doing so, neither sequential recall across and within life domains nor parallel recall across life domains is stimulated. By implementing a data-revision module, an attempt is made to use sequential and parallel retrieval strategies while simultaneously maintaining a high level of standardization.

The effectiveness of implementing data-revision modules is analyzed by looking at data modifications in the data-revision module in NEPS Starting Cohort 6—Adults. As a first result, it can be shown that nearly all vague or missing dates could be estimated in the first step of the data-revision module. Second, above all, the data-revision module is beneficial for unemployment episodes (which are highly prone to being omitted or forgotten and being added by using a data-revision module). However, third, the data-revision module also seems to be beneficial to episode types that are less bound to institutions or are less salient to the respondent. Fourth, since gap episodes (which are not pre-defined by one of the longitudinal modules) are not supposed to be reported before data revision, the frequent addition of such gap episodes to data-revision modules is not surprising. However, using the data-revision module prevents the underreporting of such gap episodes. Finally, episodes at transitions to employment, in particular, which are supposed to be reported in a life-course module, are often initially forgotten or omitted by the respondent. Using the data-revision module stimulates the parallel recall pathway and thereby improves autobiographical recall. In summary, by implementing a data-revision module in a life-course interview, more consistent and more complete retrospective life-course data can be obtained.

References

- Allmendinger, J., Kleinert, C., Antoni, M., Christoph, B., Drasch, K., Janik, F., ... Ruland, M. (2011). Adult education and lifelong learning. In H.-P. Blossfeld, H.-G. Roßbach, & J. von Maurice (Eds.), *Zeitschrift für Erziehungswissenschaft*, 14. *Education as a Lifelong Process. The German National Educational Panel Study (NEPS)* (pp. 283–299). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Antoni, M., Drasch, K., Kleinert, C., Matthes, B., Ruland, M., & Trahms, A. (2011). ALWA—New life course data for Germany. *Schmollers Jahrbuch: Zeitschrift für Wirtschafts- und Sozialwissenschaften*, 131(4), 625–634.

- Auriat, N. (1993). "My wife knows best": A comparison of event dating accuracy between the wife, the husband, the couple, and the Belgium population register. *Public Opinion Quarterly*, 57(2), 165–190.
- Barsalou, L. W. (1988). The content and organization of autobiographical memories. In U. Neisser, & E. Winograd (Eds.), *Remembering reconsidered: Ecological and traditional approaches to the study of memory* (pp. 193–243). Cambridge, MA: Cambridge University Press.
- Belli, R. F. (1998). The structure of autobiographical memory and the event history calendar: Potential improvements in the quality of retrospective reports in surveys. *Memory*, 6(4), 383–406. doi: 10.1080/741942610
- Blossfeld, H.-P., & von Maurice, J. (2011). Education as a lifelong process. In H.-P. Blossfeld, H.-G. Roßbach, & J. von Maurice (Eds.), *Zeitschrift für Erziehungswissenschaft*, 14. *Education as a Lifelong Process. The German National Educational Panel Study (NEPS)* (pp. 19–34). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Bradburn, N. M. (2000). Temporal representation and event dating. In A. A. Stone, J. S. Turkkan, C. A. Bachrach, J. B. Jobe, H. S. Kurtzman, & V. S. Cain (Eds.), *The science of self-report* (pp. 49–61). Mahwah, NJ: Erlbaum.
- Brückner, E., Hoffmeyer-Zlotnik, J., & Tölke, A. (1983). Die Daten-Edition als notwendige Ergänzung der Datenerhebung bei retrospektiven Langzeitstudien. *ZUMA-Nachrichten*, 7(13), 73–83.
- Brückner, H., & Mayer, K. U. (2005). De-Standardization of the life course: What it might mean? And if it means anything, whether it actually took place? *Advances in Life Course Research*, 9, 27–53. doi: [http://dx.doi.org/10.1016/S1040-2608\(04\)09002-1](http://dx.doi.org/10.1016/S1040-2608(04)09002-1)
- Conway, M. A. (1996). Autobiographical knowledge of autobiographical memory. In D. C. Rubin (Ed.), *Autobiographical knowledge and autobiographical memories* (pp. 67–93). Cambridge, MA: Cambridge University Press.
- Conway, M. A., & Pleydell-Pearce, C. W. (2000). The construction of autobiographical memories in the self-memory system. *Psychological Review*, 107(2), 261–288.
- Dex, S., & McCulloch, A. (1998). The reliability of retrospective unemployment history data. *Work, Employment & Society*, 12(3), 497–509. doi: 10.1177/0950017098123005
- Drasch, K., & Matthes, B. (2013). Improving retrospective life course data by combining modularized self-reports and event history calendars: Experiences from a large scale survey. *Quality & Quantity*, 47(2), 817–838.
- Freedman, D., Thornton, A., Camburn, D., Alwin, D., & Young-DeMarco, L. (1988). The life history calendar: A technique for collecting retrospective data. *Sociological Methodology*, 18, 37–68.
- Friedman, W. J. (1993). Memory for the time of past events. *Psychological Bulletin*, 113(1), 44–66.
- Hess, D., Steinwede, A., & Schneider, B. (2012). *Erhebung von retrospektiven Längsschnittdaten—Prüfmodul* (Dokumentation 4/2012). Bonn: Institut für angewandte Sozialwissenschaft.

- Hillmert, S. (2002). Edition von Lebensverlaufsdaten: Zur Relevanz einer systematischen Einzelfallbearbeitung bei standardisierten Befragungen. *ZUMA-Nachrichten*, 26(51), 120–140.
- Huttenlocher, J., Hedges, L. V., & Bradburn, N. M. (1990). Reports of elapsed time: Bounding and rounding processes in estimation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 16(2), 196–213. doi: 10.1037/0278-7393.16.2.196
- Larsen, S. F., Thompson, C. P., & Hansen, T. (1996). Time in autobiographical memory. In D. C. Rubin (Ed.), *Remembering our past: Studies in autobiographical memory* (pp. 129–156). Cambridge: Cambridge University Press.
- Matthes, B., Reimer, M., & Künster, R. (2007). Techniken und Werkzeuge zur Unterstützung der Erinnerungsarbeit bei der computergestützten Erhebung retrospektiver Längsschnittdaten: Methoden, Daten und Analysen. *Zeitschrift für empirische Sozialforschung*, 1(1), 69–92.
- Mayer, K. U., Papastefanou, G., & Tölke, A. (1989). Editionsregeln: Anweisungen zur Durchsicht und Korrektur einer retrospektiven Lebensverlaufsbefragung. In K. U. Mayer, & E. Brückner (Eds.), *Konzeption, Design und Methodik der Erhebung von Lebensverläufen der Geburtsjahrgänge 1929–1931, 1939–1941, 1949–1951*. (Vol. 2, pp. 232–264). Berlin: Max-Planck-Institut für Bildungsforschung.
- Neisser, U. (1988). Nested structure in autobiographical memory. In D. C. Rubin (Ed.), *Autobiographical memory* (pp. 71–82). Cambridge, MA: Cambridge University Press.
- Peters, H. E. (1988). Retrospective versus panel data in analyzing lifecycle events. *The Journal of Human Resources*, 23(4), 488–513.
- Reimer, M. (2005). *Autobiografisches Gedächtnis und retrospektive Datenerhebung: Die Rekonstruktion und Validität von Lebensverläufen* (Studien und Berichte No. 70). Berlin: Max-Planck-Institut für Bildungsforschung.
- Reimer, M., & Matthes, B. (2007). Collecting event histories with TrueTales: Techniques to improve autobiographical recall problems in standardized interviews. *Quality & Quantity*, 41(5), 711–735.
- Tölke, A. (1989). Möglichkeiten und Grenzen einer Edition bei retrospektiven Verlaufsdaten. In K. U. Mayer, & E. Brückner (Eds.), *Konzeption, Design und Methodik der Erhebung von Lebensverläufen der Geburtsjahrgänge 1929–1931, 1939–1941, 1949–1951*. (Vol. 1, pp. 173–226). Berlin: Max-Planck-Institut für Bildungsforschung.
- Tourangeau, R. (2000). Remembering what happened: Memory errors and survey reports. In A. A. Stone, J. S. Turkkan, C. A. Bachrach, J. B. Jobe, H. S. Kurtzman, & V. S. Cain (Eds.), *The science of self-report* (pp. 29–47). Mahwah, NJ: Erlbaum.

About the authors

K. Drasch

Chair for Methods of Empirical Social Research, University of Erlangen
Nuremberg, Kochstraße 4, 91054 Erlangen, Germany.

R. Künster

Social Science Research Center Berlin (WZB),
Reichpietschufer 50, 10785 Berlin, Germany.

B. Matthes

Institute for Employment Research (IAB),
Weddigenstr. 20–22, 90478 Nuremberg, Germany.

M. Ruland

Institute for Applied Social Sciences (infas),
Department Social Research, Friedrich-Wilhelm-Str. 18, 53113 Bonn, Germany.
e-mail: M.Ruland@infas.de

A. Steinwede

Institute for Applied Social Sciences (infas),
Department Social Research, Friedrich-Wilhelm-Str. 18, 53113 Bonn, Germany.

Measurement of Further Training Activities in Life-Course Studies

Florian Janik, Oliver Wölfel and Merlind Trepesch

Abstract

The existing research on further education in Germany provides contradictory results. One reason for this is the complex structure of the further training system; another, the lack of reliable data on further education activities. This article describes in detail all aspects of surveyed items in “NEPS Stage 8—Adult Education and Lifelong Learning” concerning further training and the underlying concept. Three types of further education can be distinguished: formal, nonformal, and informal education. Formal education refers to further education leading to generally recognized educational credentials. Formal education programs include a diverse array of fully qualifying vocational education and training degrees, such as a master craftsman’s diploma or a technician’s certificate, a university degree, or an officially recognized partial qualification. Nonformal education is organized in courses or training programs. Participants may or may not receive certificates of attendance upon course completion, but generally recognized credentials are not awarded. Typical examples include courses on presentation skills or a company’s in-house training in accounting software. Informal learning neither provides generally recognized credentials nor takes place as an organized course. Informal learning activities may include the reading of specialized periodicals or attendance at conferences or fairs. Using the advantages of a life-course study, all three types of further education are measured in NEPS Stage 8. As people remember the acquisition of certificates fairly well, measuring formal education processes is rather simple. Measuring nonformal and informal education, however, is more demanding. In NEPS Stage 8, the main idea is to use the stimuli given in the life-course interview (e. g., current employment status) in order to help the interviewed person remember further training activities. The advantages of this strategy, the design of the data, and the analytic potential are presented in this article.

1 Introduction

Demographic trends and changing skill requirements of the labor market have strengthened the focus on the further education and training sector. Policymakers, business leaders, and researchers now emphasize the relevance of further education and its benefits for society, individuals, and the economy. However, the current state of research and factual knowledge on further education leaves a lot to be desired.

The most frequently used data (Socio-Economic Panel, Mikrozensus, BIBB/BAuA-Erwerbstätigenbefragung, Adult Education Surveys) differ in several important aspects in regard to further education, for example, in sample population, survey participation, length, and retrospective interval of participation in further education. Furthermore, the data differ in their underlying concepts of further education and in the design and number of specific items covering different aspects of further education (Eisermann et al., 2014).

Indeed, several studies and research projects do exist, but their findings differ widely in specific areas and indicate contradictory messages. In general, participation rates vary greatly, between 13 % (Mikrozensus) and almost 60 % (BIBB/BAuA-Erwerbstätigenbefragung) depending on the dataset (Hall & Krekel, 2008; Statistisches Bundesamt, 2010), which explains the existence of different opinions on further education (see also Bilger & Vollmer, 2011). With regard to income returns of further education, several studies find positive effects on employees' wages and salaries (Pannenberg, 1997; Wolter & Schiener, 2009). Nevertheless, still other studies have been unable to identify any significant effect on income (Görlitz, 2011; Jürges & Schneider, 2006). Regarding the motives for participation in further education—such as prevention of unemployment or integration of nonworking individuals into the labor market—the results also depend on (sample) definitions (e. g., Beicht et al., 2006; Fleige, 2007). Overall, efforts to consolidate these areas as a prerequisite for explaining people's further education behavior have been scant.

However, all of the previous studies fail to comprehensively cover the various types of further education, the motives to engage in it, the differences in the participation rates of different social groups, and the essential context variables (e. g., individual characteristics, household information). The mixed results regarding the various dimensions and aspects of further education may to a significant extent be attributed to different datasets, heterogeneous survey groups, different definitions of further education (see next section), and different estimation methods. In the National Educational Panel Study (NEPS), a different innovative approach is used to close the gap in reliable data about further education activities. This article describes the underlying concept of further education, explains the operationalization of further education in the NEPS, shows a descriptive summary of the data, and illustrates the broad range of possibilities for analyses.

2 Types of Further Education

According to the OECD, three types of further education have to be distinguished and therefore measured in Stage 8 of the NEPS (see, e. g., Kuwan & Larsson, 2008; Von Rosenblatt & Bilger, 2008; Eisermann et al. 2014):

- *Formal education* refers to further education leading to generally recognized educational credentials. Formal education programs include all kinds of fully qualifying vocational education and training degrees, such as a master craftsman's diploma or a technician's certificate, a university degree, or an officially recognized partial qualification. Generally recognized credentials awarded by the private sector (e. g., Microsoft Technology Specialist) are also regarded as formal education in most cases.
- *Nonformal education* is organized in courses or training programs. Participants may or may not receive certificates of attendance upon course completion, but generally recognized credentials are not awarded. Courses on presentation skills or a company's in-house training in accountancy software are typical examples.
- *Informal learning*, finally, neither leads to a generally recognized credential nor takes place as an organized course. Informal learning activities may include the reading of specialized periodicals or attendance at conferences or fairs.

These three types of further education are very different from one another in terms of length, scope, and costs. Whereas informal learning activities are mostly very short-term, nonformal and especially formal education programs typically take a substantial amount of time to complete.

3 Operationalization

The development of the NEPS adult panel included two major goals: the creation of a panel study that surveys the whole life course and, at the same time, integrates all instruments of the five NEPS pillars on the one hand, and the implementation of a completely new instrument to map adult education on the other hand. Unlike regular education, adult education exhibits a couple of features that complicate the measurement of participation in life-course surveys. It can, for example, take place in a multitude of shorter courses and self-learning activities that differ in terms of location and context, content and purpose, duration and intensity, as well as in the formalism and credentials that can be attained. This flexibility is advantageous in many regards, but the missing institutionally predefined formal setting may increase the risk of recall problems, which means that respondents easily forget about their participation and, hence, do not report it in surveys. Despite this complication in measuring adult

education, the central interest within the framework of “NEPS Stage 8—Adult Education and Lifelong Learning” lies precisely with these learning activities.

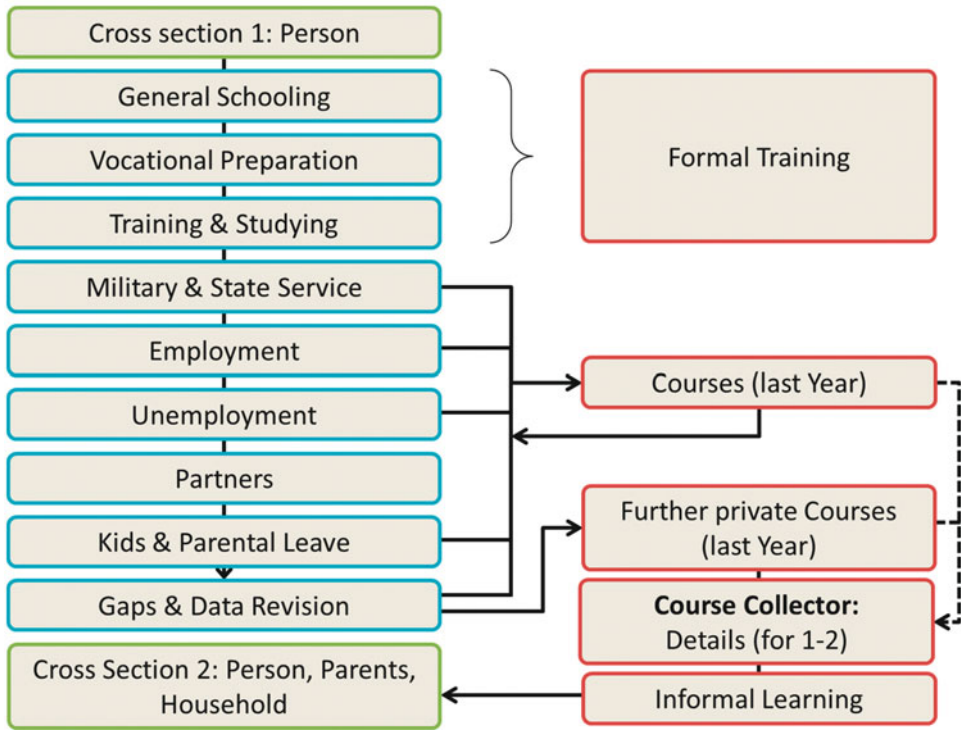
Out of all types of further education, formal training is the one that can be measured most easily. Since it leads to fully qualifying vocational education and training degrees that are very important for careers in Germany, they take place in formal settings and are remembered quite well in life-course interviews. Stage 8 of the NEPS provides instruments for the collection of information about formal training (see Drasch et al., 2016, this volume), including information about the general education history (schooling), vocational preparation schemes, and vocational education history (including studying). Therefore, formal training activities like schooling, training, and studying can be surveyed over the entire life course, dated on a monthly basis, and enriched with detailed additional information about the activity and the degree (see Figure 1).

The development of the instruments concerning nonformal training activities was more demanding. According to Allmendinger et al. (2011) and Kleinert & Matthes (2009), nonformal as well as informal further education and training activities are often forgotten shortly after they take place. Because of the difficulties respondents face when trying to recall these types of further training (Dürnberger et al., 2011), they have rarely been collected in retrospective surveys. We approached this challenge by developing new instruments on adult education and integrating these into the core questionnaire program for Stage 8 that maps our respondents’ life courses. Embedding questions on nonformal training in the life-course episodes means that people reporting military or state service, (un-)employment periods, parental leave, or gaps within the 12 months previous to the last interview are asked if they attended any courses during that episode (see Figure 1),¹ and if so, the content of the course is noted as open-text information.

The participation in nonformal further education is thus surveyed in connection with its context episode. This decision was based on the results of a pretest study (Dürnberger et al., 2011) showing that respondents remember training activities from the recent past well if they can use context-based memory strategies. However, the longer the retrospective interval is, the greater are the recall problems, which cannot be compensated for by context-based retrieval (see also Kuwan & Larsson, 2008). Hence, information on nonformal training activities is collected only for the 12 months previous to the interview (for the period between two interviews). Once the life-course data is completely collected, all reported nonformal courses are listed, thereby giving the respondent the chance to add more (private) courses. Finally, additional questions—including detailed information on duration and intensity, motivation, funding, certificates, support, structure, and challenge of the course—are

1 For more information about modular life course survey instruments, see Reimer & Matthes (2007) and Drasch & Matthes (2009). For information about the underlying concept in terms of cognitive psychology, see Reimer (2001).

Figure 1 Stylized Operationalization of Further Education in NEPS Starting Cohort 6



asked. In order to disburden frequent training participants, this additional information is only asked for a subset of two already-finished courses that are chosen randomly.

The life-course module is then supplemented by a cross sectional part of the questionnaire that includes questions about informal further education (see Figure 1). Here, standardized items about informal learning, like reading specialist literature or attending congresses or using self-learn programs, are surveyed. After the second wave of NEPS Starting Cohort 6, additional information about the content of the informal learning activity is collected. However, for the same reasons as in the case of nonformal training, the questions about informal learning refer only to the 12 months prior to the interview (or rather, to the period between two interviews). Because of the inability to measure unintentional informal learning directly, the standardized questions refer to intentional learning only. We know, though, that unintentional informal learning is also very important not only on the job, but also in the course of voluntary work and political engagement. In Stage 8, we therefore intend to include a special focus on this form of learning. Its effects in terms of competence growth will be assessed indirectly by measuring job requirements, employment experience, as

well as social and political participation. These special modules, however, are not described in this article.

The panel structure of the survey is particularly important for both nonformal and informal learning. In the long term, the sequence of repeated data collection—in which the retrospective time interval is restricted as described above at each interview date—allows for analyzing these training activities over the whole life course. Whereas formal learning activities can already be analyzed over the life course using the data of the first wave, the other two types of further education require the unique database that is created through the panel structure of the NEPS for the very first time.

4 Descriptive Summary and Analytic Potential

In Starting Cohort 6 (NEPS SC6, version 1.0.0),² the sample consists of 11,649 individuals (for a description, see also Allmendinger et al., 2011). The following section shows descriptive (unweighted) summaries and illustrates the analytic potential with regard to the three different types of further education.³

4.1 Formal Activities

The following descriptive statistics are based on a conservative definition of formal education. This includes participation in formal training and courses in order to receive a school degree or start an additional apprenticeship after a first successful vocational training.⁴ For better and clearer understanding, we classified the formal training courses according to four different motives: school as well as vocational, academic, and further courses.

School includes whether an individual has obtained an additional school degree. *Vocational courses* refer to further qualifications at special vocational schools. *Academic courses* indicate a degree at a university (e. g., of applied sciences, of public administration). *Other courses* include remaining qualifications, that is, training courses at an association/chamber of commerce and other types of leaving certificates.⁵

This definition of formal training leads to a total of 882 events reported by 826 individuals (about 7 % of the sample), who participated in at least one of the four differ-

2 doi:10.5157/NEPS:SC6:1.0.0.

3 Unweighted calculations are used to show only the possibilities of the data. Weights should be considered for substantial analyses (see also Aßmann & Zinn, 2011, Leopold et al., 2011). Weighted calculations are available upon request.

4 Hence, in contrast to other definitions, this excludes subsequent certificates within the first vocational training spell.

5 The questionnaire collects precise information about the type of degree, that is, Bachelor, Master, Diploma, doctorate, or post-doctoral lecture qualification.

Table 1 Participation Activities in Formal Training, Activities by Persons

	School	Vocational courses	Academic courses	Other courses
Events	52	126	307	397

Note. Source: NEPS, SC6 (version 1.0.0), own, unweighted calculations.

Table 2 Duration of Completed Other Formal Courses

	1–3 months	4–6 months	7–12 months	13–24 months	Two and more years
Percentage	59	11	11	12	7

Note. Source: NEPS, SC6 (version 1.0.0), own, unweighted calculations.

ent formal training types in the previous year.⁶ As shown in Table 1, only a small percentage of them attended school or vocational training after a first initial vocational training. A larger percentage participated in academic courses and further courses, that is, attended universities and courses strongly related to an occupation.

Since information about the life course is measured on a monthly basis, the duration of school or academic episodes as well as courses at chambers of commerce and industries can be easily calculated as the difference of starting an ending months. Durations for other formal courses vary heavily and range between short-time courses and long-lasting courses of more than two years (Table 2). This fact illustrates the differences between courses in regard to intensity and content. Since information about the type and content of these courses is available, this data can be used for analyses that reveal patterns and determinants of participation (success). Less-demanding training activities, that is, nonformal and informal learning activities, which are described in the next sections, are more common in adult education.

4.2 Nonformal Activities

When summarizing the participation activities in nonformal training for the different modules (compare Figure 1), it is clearly more frequent than participation in formal training (Table 3). For each respondent reporting at least one nonformal course, Table 3 reveals the participation frequencies in affiliation with the different modules,

⁶ Though it is possible to analyze formal training over the whole life course, the descriptive summary refers to the interval of 12 months for the purpose of comparison.

Table 3 Participation Activity in Nonformal Training, Activities by Persons and Modules

Type of module	Employment module	Unemployment module	Other modules	Further mentioned courses
(1) Employment module	2,448	7	4	825
(2) Unemployment module	7	183	0	25
(3) Other modules*	4	0	55	20
(4) Further mentioned courses	825	25	20	1,355
(1)+(2)+(4)	4	4	0	4
(1)+(3)+(4)	3	0	3	3
Total (N = 4,929 Persons)	3,291	219	82	2,232

Note. Source: NEPS, SC6 (version 1.0.0); own, unweighted calculations.

*Other modules include military and civilian service, parental leave, episodes of non-employment, and gaps.

that is, (un-)employment, other modules,⁷ and the number of further mentioned (private) courses after the life-course interview.

Almost 5,000 persons reported participation in nonformal training in one of the four modules. About 80 % (4,041 persons) only reported participation within a single module during the same reference period, that is, only nonformal training participation during either employment episodes (2,448 persons), unemployment episodes (183 persons), other modules (55 persons), or further mentioned courses (1,355 persons). About 20 % (888 persons) reported participation in two or more different modules. The largest percentage refers to respondents reporting courses during an employment episode and additional further mentioned courses (825 persons). As described earlier, data on nonformal activities is collected first within the different modules and second after the life-course part of the interview as a list of all additional courses that have not already been reported. Hence, these further mentioned courses could also have a strong link to an employment or unemployment episode. For this reason, information about motivation and content is also measured to facilitate integration for own analyses. Cooking courses provide a good illustration: On the one hand, they might only be motivated by individual motives (e. g., leisure and own pleasure). On the other hand, they might also be passed up to prepare for new job opportunities and to improve individuals' recently acquired skills. Overall, other courses might not only include privately motivated courses, such as language classes, but also additional job-related courses that have not been mentioned before.

⁷ Other modules are summarized and include military and civilian service, parental leave, non-employment episodes, and gaps.

4.3 Informal Learning Activities

As shown, in NEPS Stage 8, informal learning activities are measured in four different fields in the previous year: attending fairs or conferences, attending lectures or presentations, reading specialized literature, and using self-learning programs. The rate of individuals' participation in different informal learning activities strongly depends on the field and ranges from 19 % to about 66 % (Table 4).

This picture is supported when frequencies of participation in informal learning activities are compared. Reading specialized journals or books (for occupational or private reasons) is very common since the barrier to overcome is low and "participation" is easier than attending conferences or presentations. In summary, the attendance of informal activities (0 when informal learning activity was not mentioned and 1 for each mentioned informal learning activity) reveals that only a minority uses different channels for informal learning. A large part never uses informal learning activities at all, and those who participated only used one or two fields of informal learning within the last year (Table 5).

So far, the conceptual framework and structure as well as the descriptive frequencies have been shown. The next section demonstrates the enormous potential of Starting Cohort 6 and provides some ideas for analyses with regard to further education.

Table 4 Participation in Informal Learning Activities, Activities by Persons

	Attending fairs or conferences	Attending lectures or presentations	Reading specialized literature	Using self-learning programs
Participation rate (in %)	23	26	66	19

Note. Source: NEPS, SC 6 (version 1.0.0), own, unweighted calculations.

Table 5 Participation Frequencies in Informal Learning Activities, Activities by Persons

	Participation rate (in %) (including journals)	Participation rate (in %) (without journals)
No informal participation	30	56
One learning activity	30	25
Two learning activities	22	15
Three learning activities	15	4
Four learning activities	4	0

Note. Source: NEPS, SC6 (version 1.0.0), own, unweighted calculations.

4.4 Analytic Potential

As shown earlier, the knowledge about further education is quite contradictory. Using the NEPS Starting Cohort 6 (version 1.0.0) data can fill this gap. As a survey for all individuals, including working, unemployed, or nonworking (parental leave) individuals, conclusions and analyses about almost the entire employable population are possible. The overall sample consists of more than 11,000 persons and allows for analyses for different subgroups (e. g., sex, educational level).⁸ Several analyses can be conducted by combining the collected information over individuals' life courses (e. g., household and family situation, individuals' context) with the common and broadly accepted definition for further education.

Since the separation of the three types of further education is possible, prevalence and participation patterns can be examined in more detail to reveal differences and similarities between formal, nonformal, and informal further education activities. Participation patterns for different groups, that is, employed, unemployed, or non-working people, can be analyzed and may explain differences in the literature. Effects can be expected for formal training and nonformal further education, in particular. The influence of further education on the prevention of unemployment or the integration of non-working individuals (e. g., parental leave) in the labor market can be compared with appropriate reference groups. When do people participate in further education? Which types are preferred? Does participation depend on one's position in the life course or work life? Furthermore, determinants for a successful degree can be analyzed because information about success and course content is available.

What is also interesting is the influence of further education on individuals' skills and competencies. Information about competencies and cognitive skills are measured in NEPS Starting Cohort 6, and the relationship to further education can be explored. Does further education improve cognitive skills, and if so, which types and for whom?

A great advantage of NEPS Starting Cohort 6 is the panel design of the study. Until now, only analyses for formal education over the life course had been possible. But in the long term, better analyses for nonformal and informal learning activities are possible. By asking the same individuals each year, information about nonformal and informal learning activities is measured frequently. Combining information on further education with information from the life-course interview allows for additional analyses embedded in individuals' life contexts. Thus, in the long term, NEPS Starting Cohort 6 enables not only cross-sectional analyses, but also more sophisticated (panel) models.

The tremendous advantage of a panel study can be used when returns to further education need to be analyzed. Controlling for individuals' context information and changes over time, different aspects of returns, (i. e., monetary, job-related, or social)

8 Birth cohorts 1944 to 1986 are included.

can be distinguished in data and can give a more clear-cut picture of specific returns. Individuals' long-term returns from formal training on career and occupational options can be tracked and offer better policy-relevant conclusions. Do returns to nonformal further education also exist in the short-term, or do they develop in the middle or long-term? What is the real value of informal learning activities with regard to labor market outcomes?

From a non-individual point of view, possible analyses of structure and organization of further education are also an interesting aspect of NEPS Starting Cohort 6. What are the similarities and differences of participants for nonformal further education? Do the courses differ in regard to content or quality? In summary, the possibilities for data users are extensive, and the NEPS Starting Cohort 6 data offer attractive options for analyses in various research topics.

5 Conclusion

What has been missing so far in research on further education are panel data that extend information on individuals, employers, and occupations and that allow for an analysis of personal characteristics and motives, company-related or organizational aspects, and occupational aspects of further education. As shown, the NEPS Starting Cohort 6 data provide specific modules covering specific aspects of formal and nonformal education as well as informal learning and a broad range of context information. This specific design and the advantages of a panel life-course study provide researchers with a new and unique data source dealing with further education.

References

- Allmendinger, J., Kleinert, C., Antoni, M., Christoph, B., Drasch, K., Janik, F., Leuze, K., Matthes, B., Pollak, R. & Ruland, M. (2011). Adult education and lifelong learning. In H.-P. Blossfeld, H.-G. Roßbach, & J. von Maurice (Eds.), *Zeitschrift für Erziehungswissenschaft, 14. Education as a lifelong process: The German National Educational Panel Study (NEPS)* (pp. 283–299). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Aßmann, C., & Zinn, S. (2011). *Data manual - Starting Cohort 6: Adult education and lifelong learning* [Supplement C] (NEPS Research Data Paper No. 3.0.1). Bamberg: University of Bamberg, National Educational Panel Study.
- Beicht, U., Krekel, E. M., & Walden, G. (2006). *Berufliche Weiterbildung: Welche Kosten und welchen Nutzen haben die Teilnehmenden?* Bielefeld: Bertelsmann.
- Bilger, F., & Vollmer, T. (2011). *Zur Situation der Weiterbildungsbeteiligung in Deutschland. Im Gespräch über die Daten des deutschen Adult Education Survey (AES)*. Bonn: Deutsches Institut für Erwachsenenbildung.

- Drasch, K., & Matthes, B. (2009). *Improving retrospective life course data by combining modularized self-reports and event history calendars. Experiences from a large scale survey* (Discussion Paper 21/2009). Nürnberg: Institut für Arbeitsmarkt- und Berufsforschung.
- Drasch, K., Matthes, B., Kleinert, C., & Ruland, M. (2016). Why do we collect data on educational histories across the life course and their contextual conditions the way we do? Core design decisions in NEPS stage 8. In H.-P. Blossfeld, J. von Maurice, M. Bayer, & J. Skopek (Eds.), *Methodological Issues of Longitudinal Surveys. The Example of the National Educational Panel Study*. Wiesbaden: Springer.
- Dürnberger, A., Drasch, K., & Matthes, B. (2011). Kontextgestützte Abfrage in Retrospektiverhebungen: Ein kognitiver Pretest zu Erinnerungsprozessen bei Weiterbildungser eignissen. Methoden, Daten, Analysen. *Zeitschrift für empirische Sozialforschung*, 5(1), 3–35.
- Eisermann M., Janik, F., & Kruppe, T. (2014). Weiterbildungsbeteiligung—Ursachen unterschiedlicher Teilnahmequoten in verschiedenen Datenquellen. *Zeitschrift für Erziehungswissenschaft*, 17(3), 473–495.
- Fleige, M. (2007). Veränderungen des Geschlechterverhältnisses in der Weiterbildung in Deutschland. Weiterbildungsbeteiligung und Angebotsentwicklung 1980–2003. *Hessische Blätter für Volksbildung*, 57(3), 221–231.
- Görlitz, K. (2011). Continuous training and wages: An empirical analysis using a comparison-group approach. *Economics of Education Review*, 30(4), 691–701.
- Hall, A., & Krekel, E. M. (2008). Berufliche Weiterbildung Erwerbstätiger—Zur Erklärungskraft tätigkeitsbezogener Merkmale für das Weiterbildungsverhalten. *Report*, 31(1), 65–77.
- Jürges, H., & Schneider, K. (2006). Dynamische Lohneffekte beruflicher Weiterbildung. Eine Längsschnittanalyse mit den Daten des SOEP. In M. Weiß (Ed.), *Evidenzbasierte Bildungspolitik: Beiträge der Bildungsökonomie* (pp. 131–149). Berlin: Duncker & Humblot.
- Kleinert, C., & Matthes, B. (2009). *Data in the field of adult education and lifelong learning: Present situation, improvements and challenges*. (RatSWD Working Paper No. 91). Berlin: German Council for Social and Economic Data.
- Kuwan, H., & Larsson, A.-C. (2008). *Final report of the development of an international adult learning module (OECD AL Module): Recommendations on methods, concepts and questions in international adult learning surveys*. (OECD Education Working Papers No. 21). Paris: Organisation for Economic Cooperation and Development.
- Leopold, T., Raab, M., & Skopek, J. (2011). *Data manual—Starting Cohort 6: Adult education and lifelong learning* (NEPS Research Data Paper). Bamberg: University of Bamberg, National Educational Panel Study.
- Pannenberg, M. (1997). Financing on-the-job training: Shared investment or promotion based system? Evidence from Germany. *Zeitschrift für Wirtschafts- und Sozialwissenschaften*, 117(4), 525–543.

- Reimer, M. (2001). *Die Zuverlässigkeit des autobiographischen Gedächtnisses und die Validität retrospektiv erhobener Lebensverlaufsdaten. Kognitive und erhebungspragmatische Aspekte*. Berlin: Max-Planck-Institut für Bildungsforschung.
- Reimer, M., & Matthes, B. (2007). Collecting event histories with true tales. Techniques to improve autobiographical recall problems in standardized interviews. *Quality & Quantity: International Journal of Methodology*, 41(6), 711–735.
- Statistisches Bundesamt. (2010). *Bevölkerung und Erwerbstätigkeit. Beruf, Ausbildung und Arbeitsbedingungen der Erwerbstätigen*. Wiesbaden: Statistisches Bundesamt.
- Von Rosenblatt, B., & Bilger, F. (2008). *Weiterbildungsverhalten in Deutschland. Band 1: Berichtssystem Weiterbildung und Adult Education Survey 2007*. Bielefeld: Bertelsmann.
- Wolter, F., & Schiener, J. (2009). Einkommenseffekte beruflicher Weiterbildung. *Kölner Zeitschrift für Soziologie und Sozialpsychologie*, 61(1), 90–117.

Acknowledgement

This paper uses data from the National Educational Panel Study (NEPS): Starting Cohort 6—Adults (Adult Education and Lifelong Learning, doi.org/10.5157/NEPS:SC6:1.0.0). The NEPS data collection is part of the Framework Programme for the Promotion of Empirical Educational Research, funded by the German Federal Ministry of Education and Research and supported by the Federal States.

About the authors

F. Janik
Mayor of the City of Erlangen, Erlangen.

O. Wölfel
Researcher at the Institute for Employment Research (IAB), Nuremberg.
University of Bamberg, Bamberg.
e-mail: oliver.woelfel@iab.de

M. Trepesch
Researcher at the Institute for Employment Research (IAB), Nuremberg.
e-mail: merlind.trepesch@iab.de

IV. Longitudinal Measurement of Skills: Competence Testing

Selecting Appropriate Phonological Awareness Indicators for the Kindergarten Cohort of the National Educational Panel Study: A Theoretical and Empirical Approach

Karin Berendes and Sabine Weinert

Abstract

Language is the central medium for lifelong learning and consequently significantly impacts on the cognitive-academic and socio-emotional development of an individual. Thus, the assessment of language competencies is one major focus of the measurement of competencies in the German National Educational Panel Study (NEPS). Since reading literacy in the lingua franca of society is essential in order to achieve academic goals, acquire knowledge, and participate in society, this competency is assessed coherently over the lifespan in the NEPS. However, in Germany, reading is not taught before formal schooling. Therefore, reading competencies cannot be assessed in preschool or Kindergarten. Instead, phonological awareness is measured as a precursor variable of reading competence and oral language (receptive vocabulary, grammar) as well as more general literacy indicators. Although a broad range of tests and subscales for assessing phonological awareness exist, not all of them are suitable for the assessment within the framework of a large-scale educational study. Most well-established measures for assessing phonological awareness in preschool age are designed as screening instruments and/or indicators within therapeutic settings. Thus, the items are very easy, distinguishing exclusively children who show below-average performances.

In this paper, a theoretical and data-driven approach is presented to select phonological awareness tasks appropriate for the NEPS Starting Cohort 2—Kindergarten cohort. To identify tasks that comprise high psychometric quality, allow for differentiating performances at a broad range of competence levels of phonological awareness, and differ in their relationship to other language indicators, a small study ($n = 164$) was conducted. Based on a two-dimensional model of phonological awareness (Stackhouse & Wells, 1997), five different types of tasks varying in (a) the size of the linguistic unit to be reflected on and (b) the specific cognitive operation to be applied were selected and empirically compared. Statistical analy-

sis showed two tasks to be appropriate for our goals: *blending of onsets and rimes* and *identification of phonemes*.

The results are discussed in line with theoretical considerations concerning the types of tasks and against the background of the two-dimensional model of phonological awareness. The findings suggest that presumably more than two factors must be included in the model for a suitable prediction of task difficulty.

1 Introduction

The German National Educational Panel Study (NEPS) is implementing a large-scale multi-cohort sequence design to build up datasets to investigate the preconditions, consequences, and moderating variables of educational careers in Germany. One of the main questions is how educationally relevant competencies are acquired, how they develop over the lifespan, how and to what extent they are influenced by learning opportunities, and how they impact on educational outcomes. The development of competencies relevant to education and participation in social and political life are to be analyzed in their relation to important aspects of the learning environment, educational decisions, and educational returns. All data will be made available to the national and international scientific community as a Scientific Use File.¹

The NEPS began with six cohorts in parallel: 1) infants, 2) preschool/Kindergarten² children, 3) fifth graders, 4) ninth graders, 5) college students, and 6) adults. These cohorts altogether comprise a total of about 60,000 persons who are followed in their educational careers and life-courses, with measurements taking place nearly every year. In preschool, approximately 3,000 children at the age of 5 took part in a first assessment wave in 2011, and approximately 2,800 children were tested again at the age of 6. As far as possible, these children have been being followed in school since 2012.

Since language is an important means for communicating, storing, and retrieving information as well as for school performance in various school subjects, the assessment of German-language competencies across the lifespan is one major focus of the measurement of competencies in the NEPS (Weinert et al., 2011). The aim is to describe and explain the processes of competence development within and across educational stages while also analyzing their relevance for future prospects.

Some indicators must therefore be assessed coherently across the lifespan (e.g., reading competence), while the assessment of others is restricted to educational stages in which they are of special importance and have a strong predictive impact

1 Data access is possible via download, remote NEPS, and on-site. More information about the data access and user training can be found on the website <https://www.neps-data.de>.

2 Note that the differentiation between *preschool* and *Kindergarten* differs across countries and is used interchangeably in this article to refer to preschool educational institutions for children before formal obligatory schooling starts at the age of six to seven years.

(e.g., phonological awareness; for details concerning the whole conception of the assessment of language competencies within the NEPS, see Berendes, Weinert, Zimmermann, & Artelt, 2013). The coherently assessed measures of the NEPS are thought to be of special educational relevance and ecological validity across a broad age range. This leads to an assessment that heavily relies on everyday problems. The stage-specific measures are assessed in certain educational stages only and allow for further (theoretically and practically relevant) analyses. For example, in the case of phonological awareness, the predictive power (differentiated for various subgroups of children) as well as the interrelation between different language indicators, reading, and education can be analyzed.

In this article, we report on an approach to select appropriate phonological awareness indicators for the Kindergarten cohort of the NEPS (for details concerning all tests and instruments in Kindergarten, see Berendes et al., 2011). First, we briefly summarize theoretical assumptions and empirical results on (different indicators of) phonological awareness and its function in learning to read. Thereafter, the rationale for selecting appropriate tests that assess phonological awareness in the NEPS Kindergarten cohort is presented. Drawing on existing subtests and a model of phonological awareness, we aim to select tasks with a high psychometric quality that test the ability to reflect on different linguistic units while affording different cognitive operations and that differ in their relationship to the language status of the child on the one hand and the family's socioeconomic status (SES) on the other hand.

2 Phonological Awareness and Learning to Read

Phonological awareness refers to the metalinguistic ability to reflect on and manipulate the phonological structure of words independent of their meaning (Tunmer & Hoover, 1992). It is an important precursor variable of the development of written language literacy across languages and orthographies (see for an overview Ziegler & Goswami, 2005). In preschool/Kindergarten, phonological awareness is a high-impact precursor variable of later reading and spelling; later on, in the first years of school,³ it is a key competence for literacy acquisition (for an overview, see Blachman, 2000; Schnitzler, 2008). In later elementary school,⁴ the relevance of phonological awareness diminishes, but nevertheless remains existent (Del Campo, Buchanan, Abbott, & Berninger, 2015; Pfof, 2015; Wagner et al., 1997). In sum, “the discovery of a strong relationship between children's phonological awareness and their progress in learning to read is one of the great successes of modern psychology” (Bryant & Goswami, 1987, p. 439).

3 Alphabetic phase of reading and spelling acquisition (Frith, 1985).

4 Orthographic phase of reading and spelling acquisition (Frith, 1985).

Table 1 Examples of the Levels of Phonological Awareness (Taken from Tertiary Education Commission, 2008, p. 18)

Level	Examples		
Word	bed	black	napkin
Syllable	bed	black	nap-kin
Onset-Rime	b-ed	bl-ack	n-ap k-in
Phoneme	[b]-[ɛ]-[d]	[b]-[l]-[æ]-[k]	[n]-[æ]-[p]-[k]-[ɪ]-[n]

For a detailed analysis of the interrelation between reading and phonological awareness, three different linguistic levels beyond word level on which a person may reflect should be distinguished: syllable, onset-rime, and phoneme.⁵ Table 1 shows some examples for the three levels.

The differentiation of the three levels is important because it is likely that “specific phonological skills have differential effects on specific reading skills” (Christensen, 1997, p. 354).

Moreover, the effects of different forms of phonological awareness on later reading skills depend on the specific orthographical system under study (see “psycholinguistic grain size theory”, Ziegler & Goswami, 2005). Thus, the fact that results based on one language cannot easily be transferred to another one must be taken into account (see Landerl, Wimmer, & Frith, 1997; Wimmer & Goswami, 1994). Comparing the relevance of different linguistic units in alphabetic languages with different orthographic consistency, larger linguistic units can be expected to be less relevant for relatively consistent orthographies (e. g., German) than for relatively inconsistent orthographies (e. g., English; Ziegler, Perry, Jacobs, & Braun, 2001; see also Pfof, 2015), as is shown in the following sections.

Syllable awareness. The relevance of syllable awareness seems to be especially dependent on the characteristics of the specific language under study and is believed to change during reading acquisition. Schnitzler (2008) studied the relationship between early reading and syllable awareness with data from 42 German first graders.

5 A syllable can be categorized in an onset (initial consonant or cluster, optional) and an rime (vowel plus terminal consonant(s), obligatory): examples: “t-eam, dr-eam, str-eam” (Goswami, 2006, p. 489).

“The term ‘rime’ is used because words with more than one syllable have more than one rime, for example, in captain and chaplain, the rimes are -ap and -ain, respectively. The rimes are identical, but these words would not conventionally be considered to rhyme, because they do not share identical phonology after the first onset, as do rabbit and habit, for example” (Goswami, 2006, p. 489).

The results revealed that no unique variance for word and nonword reading could be explained by syllable awareness (see also Fricke, Szczerbinski, Stackhouse, & Fox-Boyer, 2008). Schnitzer (2008) hypothesizes that syllable awareness is no direct predictor of beginning reading competencies (alphabetic phase) but becomes predictive once the orthographic phase of reading acquisition has begun (see p. 60, Figure 4.1). Høien, Lundberg, Stanovich, and Bjaalid (1995) concluded for a Norwegian sample (1,509 first graders) that syllable awareness—in comparison with rhyme and phonemic awareness—“was clearly the weakest predictor” of reading competencies (reading efficiency) and that “the unique variance that it explained was quite small and attained significance only because of the extremely large size of sample” (p. 184). Moreover, they stated that “it is of marginal usefulness as predictor of early reading development if tasks at other levels are available” (p. 184). However, syllable awareness may be a more useful predictor of advanced reading.

Especially in processing long words, syllable-bound processing may be functional, because letter-by-letter processing makes greater demands on working memory (Perfetti, 1985). Using larger functional units during word processing would speed up decoding and, consequently, would free working memory for higher order processes involved in text comprehension. (Wentink, van Bon, & Schreuder, 1997, p. 166)

However, no data (or at least no sufficient data) exist to prove these theoretical considerations. For instance, Schnitzler (2008) conducted regression analyses based on data of 57 German third and fourth graders; within these analyses, syllable awareness did not account for any variance in word and nonword reading (see Schnitzler, 2008, p. 71; Figure 4.4, right column).

Taken together, syllable awareness attained less attention than did onset-rime and phonemic awareness in research on alphabetic writing systems (e.g., German, English) and more attention in syllabary writing systems (e.g., Japanese). This may be due to the reasonable assumption

...that awareness of syllables would be crucial to learning to read in syllabary (a writing system in which there is a unique symbol for each syllable in the spoken language). [...] The available research supports this general picture. For example, measures of syllable awareness are highly correlated with reading ability for Japanese children (whose initial reading involves symbols representing syllables) but not for American children (Mann, 1986). (Nagy & Anderson, 1995, p. 4)

Onset-rime awareness. Onset-rime awareness is believed to be helpful in using analogic reading and spelling strategies and helps the child to build up mental representations of written words (e.g., Goswami, 1986). For German children, it is considered to be of higher importance at the end of elementary school because word recognition at the beginning of reading acquisition heavily depends on grapheme-phoneme cor-

responsiveness. Wimmer, Landerl, and Schneider (1994) tested a total of 183 German-speaking children before they started to learn to read as well as at the end of their first year of schooling and again one and three years later. In accordance with an analogic strategy, they found that preschool phonological awareness at the onset-rime level (rhyme awareness) was significantly related to later reading (speed and accuracy) and spelling at the end of elementary school (Grades 3 and 4) but not at the end of Grade 1 (see also Landerl, Linortner, & Wimmer, 1992). However, in the studies conducted by Schnitzler (2008, see above), onset-rime awareness predicted reading competencies (word and nonword reading) neither at the beginning nor at the end of elementary school. Moreover, the relevance of phonological awareness at the onset-rime level differs between languages. “Cross-language research on children’s reading development has demonstrated quite clearly that rimes are more important orthographic and phonological units for learning to read English than for learning to read orthographically consistent languages like German and Greek” (Goswami, 2001, pp. 25–26).

Phonemic awareness. Phonemic awareness helps to grasp the alphabetic principle that underlies our system of written language (e.g., Muter, Hulme, Snowling, & Taylor, 1998) and thus plays an important role from the very beginning of reading acquisition. Moreover, faced with an alphabetic script,

...the child’s level of phonemic awareness on entering school may be the single most powerful determinant of the success she or he will experience in learning to read and of the likelihood that he or she will fail. Measures of preschoolers’ level of phonemic awareness strongly predict their future success in learning to read, and this has been demonstrated not only for English. (Adams, 1990, pp. 304–305)

Caravolas, Volin and Hulme (2005) conducted path analyses using data from primary-school learners of consistent and inconsistent orthographies; in all models conducted, phonemic awareness turned out to be a unique predictor of reading (speed and comprehension) and conventional spelling. Hulme et al. (2002) state that “good performance on phonemic awareness tasks may be the most direct indicator available that a child’s phonological representations are suitably organized to support the efficient creation of mappings between orthography (graphemes) and phonology (phonemes)” (p. 20).

The NELP (National Early Literacy Panel; Lonigan, Schatschneider, Westberg, & the National Early Literacy Panel, 2008) large-scale meta-analyses indicate that phonological awareness at the phoneme level is most appropriate for the prediction of reading. Phonemic awareness shows a medium correlation ($r_{\text{average}} = .42$) with decoding and with reading comprehension ($r_{\text{average}} = .44$). Moreover, in “terms of the specific levels of linguistic complexity, phonemic awareness had the highest correlation with decoding and reading comprehension” (Lonigan et al., 2008, p. 76). Likewise, Castles and Coltheart (2004) summarize the results of their meta-analyses:

No study that we selected for close scrutiny and that included phonemic awareness measures failed to find evidence for a significant unique contribution to subsequent reading or spelling. This stands in strong contrast with the results for syllabic and rhyme awareness. (p. 91)

For Germany, Schnitzler (2008) studied the relevance of syllable, onset-rime, and phonemic awareness to reading skills in 42 German first graders. Additionally, she included non-verbal intelligence in her analyses. Regression analyses (see Table 4.4., p. 71, left column) showed that phonemic awareness was the single phonological factor suitable for explaining the variance of reading words (37,6%) and nonwords (44,1%).

Although some studies failed to prove the outstanding role of phonemic awareness (e.g., Suggate, Reese, Lenhard, & Schneider, 2014), in sum, all languages have in common the fact that “phoneme awareness is a key component of alphabetic literacy skills in consistent and inconsistent orthographies” (Caravolas et al., 2005, p. 107; see also meta-analytic review of Melby-Lervåg, Lyster, & Hulme, 2012).

3 Development of Phonological Awareness

Many studies have demonstrated that the development of syllable awareness precedes the awareness of phonemes (e.g., Fox & Routh, 1975). The ability to detect onsets and rimes develops later than the conscious awareness of syllables but precedes insights into the phonemic structure of language (Treiman & Zukowski, 1991). Moreover, there is empirical evidence that vowels can be detected and manipulated earlier than consonants (Jansen, 1992; Mannhaupt & Jansen, 1989). This is explained by the fact that vowels are acoustically expandable and thus cover more time in the stream of speech. Additionally, tasks tapping the awareness of the initial sounds of a word are easier than tasks that tap on final sounds, and medial sounds within a word are the most difficult to work on (Yopp, 1988). Jansen (1992) as well as Mannhaupt and Jansen (1989) showed that preschool children’s ability to solve phonological awareness tasks was limited to tasks tapping the level of syllables and onset-rime and to tasks focusing on stressed vowels or very outstanding phonetic characteristics.

Overall, as far as the development of phonemic awareness is concerned, there is

...an unresolved debate in the developmental literature regarding whether phonemic awareness is acquired naturally as part of phonological awareness, or whether it is instead an artefact of reading tuition. This ambiguity affects the interpretation of studies which show that pre-literate phonemic awareness is a powerful predictor of literacy attainment in school. [...] Results suggest that young children can develop phonemic awareness before beginning reading or attending school. (Wood & Terrell, 1998, p. 253)

Likewise, studies with German samples have shown that basic phonemic awareness in general exists before children receive literacy tuition (e. g., Fricke, 2007; Fricke, Stackhouse, & Wells, 2007; Marx, Weber, & Schneider, 2005; Schäfer, Bremer, & Herrmann, 2014; Schäfer et al., 2009; Schäfer, Stackhouse, & Wells, in preparation). However, “[f]ull access to phonemes only develops once children are taught to read and write, irrespective of the age at which reading and writing is taught” (Ziegler & Goswami, 2005, p. 6). Thus, whether or not phonemic awareness is evident before reading tuition depends on the kind of task administered to assess phonemic awareness. Therefore, Moyle, Heilmann, and Berman (2013) requested “that task difficulty needs to be reduced so that younger children can participate in assessments of phoneme-level skills” (p. 682).

Whereas the development of syllable awareness and onset-rime awareness is rather similar across different languages (Goswami, 2006), the development of phonemic awareness differs according to the specific language under study. “Children learning transparent orthographies such as Greek, Finnish, German, and Italian acquire phonemic awareness relatively quickly. Children learning nontransparent orthographies such as English, Danish, and French are much slower to acquire phonemic awareness” (Goswami, 2006, p. 490; see also Goswami, 2008, p. 9, Table 1).

The development of phonological awareness is pictured in a widely acknowledged two-dimensional model (see Figure 1). The assumed developmental trajectory of phonological awareness skills is indicated by the diagonal arrow.

The first dimension is the size of the linguistic unit (beyond the word level) on which a person is able to reflect. As already mentioned, three unit sizes could be distinguished: syllable, onset-rime, and phoneme. The second dimension is the level of explicitness of the cognitive operation needed to solve the task. Four levels can be

Figure 1 Development of Phonological Awareness (see Schäfer et al., 2009, p. 405; Fricke, 2007, p. 11)

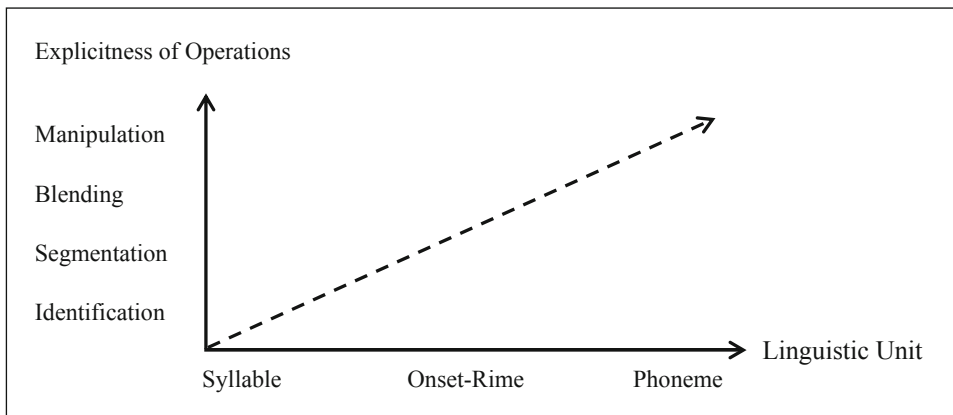


Table 2 Operations Involved in Phonological Awareness Tasks, Their Cognitive Processes, and Examples of Tasks

Operation	Cognitive Processes	Example at the Phoneme Level
Identification	Detection of units	Recognizing the common sound in different words, for example, "Tell me the sound that is the same in bike, boy, and bell" ([b]).
Segmentation	Detection of units and segmenting these units	Breaking a word into its sounds by tapping out or counting the sounds, for example, "How many phonemes are in ship?" (3: [ʃ] [ɪ] [p]).
Blending	Detection of units and synthesizing of these units	Listening to a sequence of separately spoken sounds and combining them to form a recognizable word, for example, "What word is [s] [k] [u:] [l]?" (school).
Manipulation	Detection of units, segmenting these units, manipulation of these units (replacement, elision, addition, reorganization), and synthesizing these units	Recognizing what word remains when a specified phoneme is removed, for example, "What is smile without the /s/?" (mile).

differentiated for this dimension: identification, segmentation, blending, and manipulation (Stackhouse & Wells, 1997). These levels refer to the depth of metalinguistic reflection that is needed to complete a phonological awareness task. Whereas some tasks (e. g., identification) require less awareness and may be regarded as more-or-less implicit tasks, other tasks (e. g., manipulation) require higher and more explicit levels of awareness. In general, the cognitive complexity of a task increases with the explicitness of the operation.

Table 2 depicts the cognitive processes that are involved in the four levels and that differ with respect to their explicitness (according to Fricke & Schäfer, 2008, p. 11). Moreover, for each operation, an example from the phoneme level is given in the table (examples are taken from the National Institute of Child Health and Human Development (NICHD), 2000, 2–10).

The development of phonological awareness has been suggested to continually proceed from larger to smaller linguistic units (syllable—onset-rime—phoneme)⁶ and from simple to complex, explicit operations (identification—segmentation/blending—manipulation). Thus, tasks affording the 'identification of syllables' are expected to be the easiest type of task, and tasks requiring the 'manipulation of phonemes' are expected to be the most difficult of the 12 types of tasks. At present, it is not possible

6 This is termed the "linguistic status hypothesis" (Treiman, 1992). Although the linguistic status of a unit is often confounded with its size (as measured by the number of phonemes) and a longer length of a unit could account for greater accessibility, there is still evidence for the linguistic status hypothesis when units that differ in linguistic level but are equated for their size are compared concerning their item difficulty (see Treiman & Zukowski, 1996).

to make a comparative statement on the relative difficulty of tasks varying according to both classification characteristics if one of the tasks involves an easier, larger linguistic size to work on but at the same time requires a more difficult (more complex, more explicit) operation. So far, no clear consensus or sufficient data exist regarding the question of whether the level of difficulty of a phonological task is determined by the size of the linguistic unit or the explicitness of the operation.

Moreover, task difficulty is influenced by so-called side factors, such as sonority, intonation, the position of the phonological unit to be worked on, and the complexity of the phonological surroundings in which the phonological unit is embedded (Schnitzler, 2008; Smith, Simmons, & Kameenui, 1998). These factors are not taken into account in the two-dimensional model of phonological awareness (see Figure 1). It is unclear how strongly these and other factors (e.g., phrasing of instruction, response format, picture-based or not) influence task difficulty compared with the two main dimensions of phonological awareness. Stanovich, Cunningham, and Cramer (1984) compared the performance on ten different phonological awareness tasks and detected that two tasks affording the manipulation of initial phonemes (stripping and substituting the initial phoneme) differed substantially with respect to their difficulty (25.3 % correct vs. 86.3 % correct). The authors considered specific task characteristics to be responsible for these results. Fricke (2007) discovered unexpected results with regard to the assumed developmental order. In her study, a task requiring the identification of phonemes turned out to be more difficult for the children than did two tasks requiring the synthesis of phonemes. Results conducted by Schäfer, Wessels, and Fricke (2014) also indicate that the performance level children attained in phonological awareness tasks partly depended on other task demands and instructional issues.

Table 3 shows part of a summary proposed by Schnitzler (2008) concerning our empirically based knowledge of the performance level of preschool children concerning different phonological tasks.

Table 3 Performance Level of Preschool Children Concerning Different Types of Tasks to Assess Phonological Awareness (Schnitzler, 2008, p. 52, extract from table 3.11)

	Syllable	Onset-Rime	Phoneme
Manipulation			
Blending/Segmentation	(++)		--
Identification	++	+	

Note. ++ stands for a very good performance (average performance of 75–100% correct), + stands for a good performance (average performance of 50–74% correct), -- stands for very low performance (average performance of 0–24% correct); uncertain declaration is marked with brackets.

Five slots cannot be filled in yet because of a lack of research data. This fact shows that there is a need for further research concerning the phonological awareness abilities of preschool children and that the exact chronological order of development is still unclear (Schäfer, Wessels, & Fricke, 2014).

4 Assessment of Phonological Awareness: A Small Pilot Study

There are various well-known tests and subtests in German as well as in other languages to reliably and validly assess phonological awareness in preschool children. Table 4 presents some of the more or less well-known German test instruments.

However, most of these instruments are designed for and used in therapeutic settings, mainly as screening instruments. Consequently, they focus on children showing below-average performance. For example, using classical test theory, analyses of task difficulty in the well-known *Bielefelder Screening* (Jansen, Mannhaupt, Marx, & Skowronek, 2002) reveal a task difficulty of 0.78 and 0.80 for the most difficult task (10 and 4 months before school entry, respectively).

In the NEPS, we conducted a small study to compare different types of tasks to select suitable ones for our large-scale assessment. In this study, the performance of 6-year-old preschool children was investigated using five different types of tasks. The aim was to identify tasks that would allow us to discriminate performance differences across a broad range of performance levels, that is, to differentiate between lower

Table 4 Examples of German Test Instruments to Assess Phonological Awareness

Acronym	Name of the Test	Authors & Year of Publication
ARS	Anlaute hören, Reime finden, Silben klatschen – Ein Erhebungsverfahren zur phonologischen Bewusstheit für Vorschulkinder und Schulanfänger	Martschinke, Kammermeyer, King, & Forster, 2005
BAKO 1-4	Basiskompetenzen für Lese-Rechtschreibleistungen	Stock, Marx, & Schneider, 2003
BISC	Bielefelder Screening zur Früherkennung von Lese-Rechtschreibschwierigkeiten	Jansen, Mannhaupt, Marx, & Skowronek, 2002
–	Der Rundgang durch Hörhausen	Martschinke, Kirschhock, & Frank, 2001
MÜSC	Münsteraner Screening	Mannhaupt, 2006
PB-LRS	Gruppentest zur Früherkennung von Lese-Rechtschreibschwierigkeiten	Barth & Gomm, 2006
QUIL-D	deutschsprachige Version des Queensland Inventory of Literacy	Hofmann, 2000
TPB	Test für Phonologische Bewusstheitsfähigkeiten	Fricke & Schäfer, 2008

as well as average and above-average performance. The selection of tasks was based on the theoretical framework that describes phonological awareness as a two-dimensional construct (see above). Thus, the tasks differed in terms of the size of the linguistic unit tapped and the cognitive operation required by the task. As already mentioned, little is known about the interaction between the explicitness of operations and the size of the linguistic units in typically developing preschool children.

4.1 Aims of the Study

Overall, the study had the following four aims: The first aim was to select suitable phonological tasks for the NEPS Kindergarten assessment while at the same time contributing to the issue of explaining the difficulty of different phonological tasks by factors that might impact on task difficulty. In addition, we intended to add some information on the interrelations between task performance and other child- and environmental variables that seemed to be especially relevant to large-scale educational assessments. Thus, as a second aim, we examined the interrelation between family background (SES measured by the number of books in the household) and phonological awareness. Since “phonological awareness is highly teachable and modifiable” (Lundberg, Larsman, & Strid, 2012, p. 318; see also Fischer & Pfof, 2015), we expected medium to high correlations with SES. Third, we tested for the interrelation between interindividual differences in phonological working memory capacity as a rather stable child characteristic and phonological awareness since “many tasks devised to tap phonological awareness also impose significant burdens on verbal memory” (Alloway, Gathercole, Willis, & Adams, 2004, p. 88; see also Nithart et al., 2011). Finally, the association of the performance on each of the five awareness tasks with a proxy indicator of the child’s language competencies was considered. We applied a task measuring sentence reproduction because this task is well known as a reliable indicator of child language competencies as it comprises receptive as well as reproductive and reconstructive aspects on the one hand and proved to be a valid predictor of later reading and spelling competencies on the other hand (Ebert & Weinert, 2013; von Goldammer, Mähler, Bockmann, & Hasselhorn, 2010). Since sentence reproduction is partly determined by the capacity of phonological working memory (von Goldammer, Mähler, & Hasselhorn, 2011), we controlled for this variable when analyzing the interrelation between phonological awareness and sentence reproduction.

4.2 Method

Existing more or less well-established test instruments (see Table 4) as well as some subtests of lesser-known tests and test batteries were looked through, and tasks were classified according to the linguistic unit and the dimension of operation tapped by

the respective task. Furthermore, the statistical characteristics (item difficulty, item selectivity, internal consistency) were taken into account, and only tasks with a Cronbach's alpha (if declared) of .80 or higher were included in our study.

4.2.1 Sample

164 children with different language backgrounds (114 German, three Polish, four Russian, eight Turkish, eight other languages, 27 no answer) and a mean age of 5;9 years (min. = 5;3 years, max = 6;5 years) took part in this study. Children were recruited from 15 preschools/Kindergartens in four federal states of Germany: Four in Bavaria ($N = 46$), three in Hamburg ($N = 42$), five in North Rhine-Westphalia ($N = 38$), and three in Thuringia ($N = 38$).

4.2.2 Materials

The following five tasks were included in the data collection to assess phonological awareness:

Identification of syllables: The ability to identify syllables was measured by the subscale *Silbenidentifizieren* (SI) (identification of syllables) from the German version of the *Queensland Inventory of Literacy* (QUIL-D; Hofmann, 2000). Two two-syllable words were presented, and the child was invited to decide whether the two words had a similar beginning (same initial syllable), a similar ending (same final syllable), or no similar part.

Manipulation of syllables: The ability to manipulate syllables was assessed by a modification of the subscale *Silbenzusammensetzen* (reassembling of syllables) from the *Rundgang durch Hörhausen* (Martschinke et al., 2001). Two bisyllabic words (animal names) were presented to the child; pictures of these animals were cut into two parts, and each part was introduced as corresponding to one of the syllables. The child was asked to combine the first syllable of the one word with the second syllable of the other word and vice versa (e. g., <Zie|ge—Ka|mel> → <Zie|mel—Ka|ge>). The task was supported by rearranging the parts of the picture cards to show the corresponding fantasy animal.

Blending of onsets and rimes: The child heard monosyllabic words with a gap between the onset and the rime and was asked to blend these two parts (subscales *Onset-Reim-Synthetisieren—output* (onset-rime synthesis—output) from the TPB, Fricke & Schäfer, 2008).

Identification of phonemes: The ability to identify phonemes was measured with a set of picture-based multiple-choice tasks (subscales *Laut-Wort-Zuordnung* (sound-word

classification) from the *MÜSC*, Mannhaupt, 2006). The child heard a phoneme and then heard three words and was instructed to point to the picture that illustrated the word with the previously heard phoneme.

Manipulation of phonemes: In order to assess the ability to manipulate phonemes, mono- or bisyllabic words were presented to the child, and the child was asked to repeat the word without the initial phoneme (subscale *Anlaute-Manipulieren—output* (manipulation of initial sounds) from the TPB, Fricke & Schäfer, 2008). We included this subtest although we expected it to be rather difficult, or potentially too difficult as indicated by a pilot study with children who had nearly the same age as our sample ($M = 6.0$ years, $N = 38$; Fricke, Stackhouse, & Wells, 2007; see also Fricke & Schäfer, 2008, p. 77) because we wanted to compare the task difficulty with that of the other tasks.

In addition, tasks to assess phonological working memory, sentence repetition, and letter knowledge were included in the data collection:

Phonological working memory: Two tasks to assess phonological working memory were administered, a *digit span* task (taken from the German version of the Kaufman Assessment Battery for Children (K-ABC), Melchers & Preuß, 2009) and a *digit span backward* task. The latter required a change in the order of stimulus material (naming the digits in backwards order) and thus involved the central executive of working memory (taken from *Hamburg-Wechsler-Intelligenztest für Kinder III—HAWIK III*, Tewes, Rossmann, & Schallberger, 1999).

Sentence repetition: The ability to reproduce sentences of increasing grammatical complexity was measured by a subscale of the *Sprachentwicklungstest für drei- bis fünfjährige Kinder* (SETK 3-5; Grimm, 2001).

Letter knowledge: As an indicator of emerging literacy (see Kim, Petscher, Foorman, & Zhou, 2010), we assessed the letter knowledge of the children by giving them a card with all 26 letters of the German alphabet (in a fixed but random order) and asking to name them.

Moreover, the parents were asked about the *number of books* in their household. The number of books is a good indicator for the cultural capital of a family (Paulus, 2009) and is thus often applied as an indicator of the familial SES.

4.2.3 Test Procedure and Training of Test Administrators

Children were tested individually in a quiet room in their preschool. Each child participated in two 30-minute testing sessions on separate days. On the first day, four subtests were presented in the following order: 1) identification of syllables, 2) blending of onsets and rimes, 3) early letter knowledge, and 4) digit span. On the second day, five more subtests were administered: 5) identification of phonemes, 6) digit span backwards, 7) manipulation of syllables, 8) repetition of sentences, and 9) manipulation of phonemes. All tests were instructed as playful games and administered by well-trained test administrators. Stimuli were presented digitally (CD-ROM) to guarantee standardization (e.g., intonation, speech rate) and were spoken by a professional radio speaker to assure high-quality recordings.

All test administrators participated in a two-day test-administrator training conducted by NEPS staff.⁷ Drawing on these training sessions and comprehensive test manuals, all test administrators had to practice and videotape the assessment procedures with two children. These videos were evaluated by NEPS scientific staff to ensure correct handling of test materials, high standardization of the test procedures, and suitable contact with the child. Finally, a third test-administrator training day was arranged to further discuss and train the test administration based on the video evaluations. To ensure high-quality data, only those test administrators who performed well enough during training were recruited for the assessments, which were run by the Data Processing and Research Center (DPC), which is part of the International Association for the Evaluation of Educational Achievement (IEA).

4.3 Results

Phonological awareness tasks. Test results were evaluated and compared using classical test theory (see Table 3 for item difficulties; additional details on the psychometric quality as well as on considerations concerning test selection are given in Berendes et al., 2013). In summary, two tasks emerged as suitable to our study: The subscale identification of phonemes was chosen to differentiate at the lower level of performance (average item difficulty (p_i) = .81; average item selectivity (r_{ii}) = .53; Cronbach's alpha (α) = .83), and the subscale blending of onsets and rimes was chosen as a more difficult task (average item difficulty (p_i) = .23; average item selectivity (r_{ii}) = .74; Cronbach's alpha (α) = .94).

From a theoretical point of view, an overall look at the data suggests that the type of cognitive operation more strongly impacts item difficulty than does the size/type of

7 For more detailed information on tests, test administration, and test administrator training ("train-the-trainer program") in the main studies of the NEPS Kindergarten cohort, see Weinert and Berendes (2012). To acquire this poster, please contact one of the authors.

Table 5 Average Item Difficulty of the Five Phonological Tasks Assessed in the Preliminary Study

	Syllable (p_s)	Onset-Rime (p_r)	Phoneme (p_i)
Manipulation	0.21		0.06
Blending/Segmentation		0.23	
Identification	0.51		0.81

the linguistic unit the child has to reflect on. As Table 5 shows, item difficulty increases from the bottom to the top (identification → blending/segmentation → manipulation), but not from left to right (syllable → onset-rime → phoneme).

A closer look at the data shows that—in line with our expectations—the task *manipulation of phonemes* was the most difficult one. Also in line with our expectations, a task that requires the identification of syllables was easier than tasks that implied the manipulation of syllables, the blending of onsets and rimes, or the manipulation of phonemes. However, contrary to our expectations, the task *identification of syllables* turned out to be more difficult than the identification of phonemes (discussed later).

Interrelation between phonological awareness skills and the number of books at home. The correlations (see Table 6) show that the *blending of onsets and rimes* is significantly related to the SES-indicator ($r = .26^{**}$), whereas the other four tasks show no significant relationship with the number of books at home.

Interrelation between phonological awareness skills, phonological working memory, and sentence reproduction. Four of the five phonological awareness tasks were significantly related to the two tasks measuring phonological working memory (see Table 6).

The interrelation of the five tasks with sentence repetition (with and without control of phonological working memory, see Table 6) proved to be highly task-dependent. Two tasks were significantly related to sentence reproduction, even when controlling for digit span or digit span backwards: identification of syllables ($r = .23^{**}$, $.20^{**}$, $.17^*$) and identification of phonemes ($r = .29^{**}$, $.23^{**}$, $.29^{**}$).

Thus, the two phonological awareness tasks chosen for the NEPS assessments (*blending of onsets and rimes* and *identification of phonemes*) differed with respect to (a) task difficulty, (b) social disparities according to the number of books at home, and (c) their intercorrelation with sentence repetition as a proxy of language competence. However, performance on both tasks was associated with phonological working-memory capacity. Since phonological working memory is also included in the NEPS data assessment, the effect of digit span and digit span backwards can be statistically controlled.

Table 6 Correlations (Pearson) Between Phonological-Awareness Skills and the Number of Books in the Household, Phonological Working Memory (Digit Span and Digit Span Backwards), and Sentence Repetition (Additionally With Differences in Phonological Working Memory Partialled Out)

	Number of Books	Digit Span	Digit Span Backwards	Sentence Reproduction (SR)	SR Controlling for Digit Span	SR Controlling for Digit Span Backwards
Identification of Syllables	.12	.12	.28**	.23**	.20**	.17*
Manipulation of Syllables	.15	.27**	.27**	.11	.05	.03
Blending of Onsets and Rimes	.26**	.30**	.25**	.15	.08	.19*
Identification of Phonemes	.14	.31**	.45**	.29**	.23**	.29***
Manipulation of Phonemes	.16	.25**	.21**	.07	.00	-.14

Note. $n = 164$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

4.4 Discussion

The purpose of our small study was to select appropriate phonological awareness tasks for our NEPS assessment in Kindergarten. Drawing on existing subtests and a model of phonological awareness, we identified two tasks (subtests) with high psychometric quality that test different linguistic units, tap different cognitive operations, and differ in their relationship (a) to the language status of the child and (b) to an indicator of family background (SES) as well as in their task difficulty, suggesting that these tasks may differentiate between children at different performance levels. In addition, both tasks can be expected to be associated with different aspects of later reading competence (see above).

With respect to *phoneme identification*, previous research has shown that this task explains the highest proportion of unique variance in reading compared with various phonemic awareness tasks (Høien, Lundberg, Stanovich, & Bjaalid, 1995). Interestingly enough, in our study, this task shows significant correlations to our language indicator (*sentence repetition*), even after controlling for phonological working memory (*digit span* and *digit span backwards*). The second task, *blending of onsets and rimes*, was not only more difficult, but also proved to be the only task significantly correlated to the SES indicator (*number of books in the household*). This is unexpected since most studies investigating this relationship found SES differences for phonological awareness performance, as we did for the *blending of onsets and rimes* task (e. g., McDowell,

Lonigan, & Goldstein, 2007; Lundberg, Larsman, & Strid, 2012; Lundberg, 2009; Bowey, 1995). However our result might be due to the age and the reading development of the children under study. In fact, McDowell et al. (2007) found evidence that the “effect of SES on phonological awareness is amplified as age increases” (p. 1087). They presume that “the size of this relation will be smaller in younger children because of weaker psychometric properties of the measures, lack of exposure to activities that promote the development of phonological awareness, or both” (p. 1082).

When comparing the difficulty of the different tasks, the pattern does not simply reflect the two-dimensional model underlying our task selection. This is, in fact, of theoretical interest and suggests that additional factors not specified in the model are highly relevant to task performance. Specifically, in our study, the identification of phonemes was the easiest task for the children and was even easier than the identification of syllables. At first sight, this is unexpected because the identification of syllables is—in general—believed to be the easiest phonological task. Moreover, as mentioned above, phonemic awareness is expected to be difficult for children before reading tuition. Thus, a detailed look at the task format and the applied test items is needed to identify relevant additional variables influencing task performance. A detailed look at the test items of the phoneme identification task showed that many of the initial phonemes were vowels (70 %) and/or had syllable quality (40 %, e.g., Ameise à A-mei-se), which facilitates phoneme identification. Additionally, only initial phonemes had to be identified, and the performance on a phoneme task depends on the position of the phoneme within the word (de Graaff, Hasselman, Verhoeven, & Bosman, 2011). Regarding consonants, initial ones are “significantly more identifiable than final consonants” because of their “greater acoustic distinctiveness” (Redford & Diehl, 1999, p. 1555). Moreover, phoneme class could have had an effect on test results. For German children, plosives (b-d-g-k-p-t) are expected to be very difficult to identify because of their acoustic characteristics (short duration of approx. 30–70 msec; Barth, 1999). Furthermore, the identification of initial phonemes in consonant clusters (complex onsets) is more demanding than in a CVC structure (Barth, 1999). The items we applied did not include any plosives or phonemes that were part of a consonant cluster as a target phoneme. Moreover, “perceptual properties, such as sonority levels, greatly influence the development of phoneme awareness” (Yavas & Gogate, 1999, p. 245; see also de Graaff, Hasselman, Bosman, & Verhoeven, 2008). Thus, if more initial phonemes would have been unvoiced plosives in a complex onset, that would likely have resulted in notably higher item difficulty. Moreover, the task *identification of syllables* required a comparison of the initial and final syllables of two words, whereas the identification of phonemes focused on the initial phoneme of one word and was—in addition—picture-based while the task *identification of syllables* was not. This could have influenced the motivation of the child (both tasks were related to working memory). Additionally, the two tasks imply different response formats. The task *identification of phonemes* asked the child to point to a picture, while the task *identification of syllables* required a verbal response (indicating whether the

two words included a similar part or not, and if they did, stating the position of the same syllable).

Although our data support the assumption that the type or explicitness of the cognitive operation impacts more strongly on item difficulty than does the size of the linguistic unit, this presumption is by no means clear-cut and may be relativized when taking task-specific considerations into account. In fact, five tasks—as in our small preliminary study—are not sufficient to support generalized statements, especially when acoustic features and task formats differ widely across the five tasks. Hulme et al. (2002) used a more focused method to compare different phonological awareness skills. By implementing a repeated measurement design, they used multiple measures (detection, deletion, or oddity judgments) to assess the awareness of different phonological units (onset or rime, initial phoneme, final phoneme) while using identical items in each task. In doing so, they were able to control for many item- and child-specific influences. They found that “[m]easures of phoneme awareness were the best concurrent and longitudinal predictors of reading skill with onset-rime skills making no additional predictive contribution once phonemic skills were accounted for” (Hulme et al., 2002, p. 2).

Taken together, we conclude that the two-dimensional model of phonological awareness is not sufficient to represent the underlying demands and interrelationships in order to predict the item difficulty of the five types of phonological tasks. Sound characteristics—among others—should be considered systematically (e. g., by using the five-point scale suggested by Yavas & Gogate, 1999). Moreover, the linguistic surrounding (e. g., simple or complex onset/syllable structure) and the position (initial, medial, final) of the linguistic unit should be considered systematically. Additionally, other aspects of the task (e. g., picture-based or not, response format) should be taken into account. Finally, phonological awareness tasks may be differentially associated with other characteristics of the child as well as with the learning environment, thereby demonstrating that they are possibly a complex multifaceted construct.

References

- Adams, M. J. (1990). *Learning to read: Thinking and learning about print*. Cambridge, MA: MIT Press.
- Alloway, P. T., Gathercole, S. E., Willis, C., & Adams, A.-M. (2004). A structural analysis of working memory and related cognitive skills in young children. *Journal of Experimental Child Psychology*, 87(2), 85–106.
- Barth, K. (1999). *Zur Prophylaxe von Lese-Rechtschreibstörungen: Zeitliche Verarbeitungsprozesse und ihr Zusammenhang mit phonologischer Bewußtheit und der Entwicklung von Lese- Rechtschreibkompetenz* (Doctoral dissertation, University of Dortmund, Department of Special Education and Rehabilitation). Retrieved from <https://eldorado.tu-dortmund.de/bitstream/2003/2922/2/barthge.pdf>

- Barth, K., & Gomm, B. (2006). *Gruppentest zur Früherkennung von Lese- und Rechtschreibschwierigkeiten. Phonologische Bewusstheit bei Kindergartenkindern und Schulanfängern (PB-LRS)*. Reinhardt: München.
- Berendes, K., Fey, D., Linberg, T., Wenz, S., Roßbach, H.-G., Schneider, T., & Weinert, S. (2011). Kindergarten and elementary school. In H.-P. Blossfeld, H.-G. Roßbach, & J. von Maurice (Eds.), *Zeitschrift für Erziehungswissenschaft, 14. Education as a lifelong process: The German National Educational Panel Study (NEPS)* (pp. 203–216). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Berendes, K., Weinert, S., Zimmermann, S., & Artelt, C. (2013). Assessing language indicators across the lifespan within the German National Educational Panel Study (NEPS). *Journal of Educational Research Online, 5*(2), 15–49.
- Blachman, B. A. (2000). Phonological awareness. In M. L. Kamil, P. B. Mosenthal, P. D. Pearson, & R. Barr (Eds.), *Handbook of reading research* (pp. 483–502). Mahwah, NJ: Lawrence Erlbaum Associates.
- Bowey J. A. (1995). Socioeconomic status differences in preschool phonological sensitivity and first-grade reading achievement. *Journal of Educational Psychology, 87*(3), 476–487.
- Bryant, P., & Goswami, U. (1987). Beyond grapheme-phoneme correspondence. *European Bulletin of Cognitive Psychology, 7*, 439–443.
- Caravolas, M., Volin, J., & Hulme, C. (2005). Phoneme awareness is a key component of alphabetic literacy skills in consistent and inconsistent orthographies: Evidence from Czech and English children. *Journal of Experimental Child Psychology, 92*(2), 107–139.
- Castles, A., & Coltheart, M. (2004). Is there a causal link from phonological awareness to success in learning to read? *Cognition, 91*(1), 77–111.
- Christensen, C. A. (1997). Onset, rhymes, and phonemes in learning to read. *Scientific Studies of Reading, 1*(4), 341–358.
- de Graaff, S., Hasselman, F., Bosman, A. M. T., & Verhoeven, L. (2008). Cognitive and linguistic constraints on phoneme isolation in Dutch kindergartners. *Learning and Instruction, 18*(4), 391–403.
- de Graaff, S., Hasselman, F., Verhoeven, L., & Bosman, A. M. T. (2011). Phonemic awareness in Dutch kindergartners: Effects of task, phoneme position, and phoneme class. *Learning and Instruction, 21*(1), 163–173.
- Del Campo, R., Buchanan, W. R., Abbott, R. D., & Berninger, V. W. (2015). Levels of phonology related to reading and writing in middle childhood. *Reading and Writing, 28*(2), 183–198.
- Ebert, S., & Weinert, S. (2013). Predicting reading literacy in primary school: The contribution of various language indicators in preschool. In M. Pfof, C. Artelt, & S. Weinert (Eds.), *The development of reading literacy from early childhood to adolescence: Empirical findings from the Bamberg BiKS longitudinal study* (pp. 93–149) Bamberg: University of Bamberg Press.
- Fischer, M. Y., & Pfof, M. (2015). Wie effektiv sind Maßnahmen zur Förderung der phonologischen Bewusstheit? Eine meta-analytische Untersuchung der Auswirkungen

- deutschsprachiger Trainingsprogramme auf den Schriftspracherwerb. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, 47(1), 35–51.
- Fox, B., & Routh, D. K. (1975). Analyzing spoken language into words, syllables and phonemes: A developmental study. *Journal of Psycholinguistic Research*, 4(4), 331–342.
- Fricke, S. (2007). *Phonological awareness skills in German-speaking preschool children*. Idstein: Schulz-Kirchner Verlag.
- Fricke, S., & Schäfer, B. (2008). *Test für Phonologische Bewusstheitsfähigkeiten (TPB)*. Idstein: Schulz-Kirchner Verlag.
- Fricke, S., Stackhouse, J., & Wells, B. (2007). Phonologische Bewusstheitsfähigkeiten deutschsprachiger Vorschulkinder—Eine Pilotstudie. *Forum Logopädie*, 3(21), 14–19.
- Fricke, S., Szczerbinski, M., Stackhouse, J., & Fox-Boyer, A. V. (2008). Predicting individual differences in early literacy acquisition in German: The role of speech and language processing skills and letter knowledge. *Written Language & Literacy*, 11(2), 103–146.
- Frith, U. (1985). Beneath the surface of developmental dyslexia. In K. E. Patterson, J. C. Marshall, & M. Coltheart (Eds.), *Surface dyslexia* (pp. 301–326). Boston: Routledge.
- Goswami, U. (1986). Children's use of analogy in learning to read: A developmental study. *Journal of Experimental Child Psychology*, 42(1), 73–83.
- Goswami, U. (2001). Rhymes are important: A comment on Savage. *Journal of Research in Reading*, 24(1), 19–29.
- Goswami, U. (2006). Phonological awareness and literacy. In K. Brown (Ed.), *Encyclopedia of language and linguistics* (pp. 489–497). Oxford: Elsevier.
- Goswami, U. (2008). The development of reading across languages. *Annals of the New York Academy of Sciences*, 1145(1), 1–12.
- Grimm, H. (2001). *Sprachentwicklungstest für drei- bis fünfjährige Kinder (SETK 3-5)*. Göttingen: Hogrefe.
- Hofmann, C. D. (2000). *Phonological awareness abilities in German-speaking second graders: Comparison between children with normal literacy and with dyslexia* (Unpublished master-thesis, Department of Speech, University of Newcastle upon Tyne, United Kingdom).
- Høien, T., Lundberg, I., Stanovich, K. E., & Bjaalid, I.-K. (1995). Components of phonological awareness. *Reading and Writing: An Interdisciplinary Journal*, 7(2), 171–188.
- Hulme, C., Hatcher, P. J., Nation, K., Brown, A., Adams, J., & Stuart, G. (2002). Phoneme awareness is a better predictor of early reading skill than onset-rime awareness. *Journal of Experimental Child Psychology*, 82(1), 2–28.
- Jansen, H. (1992). *Untersuchungen zur Entwicklung lautsynthetischer Verarbeitungsprozesse im Vorschul- und frühen Grundschulalter*. Köln: Hänsel-Hohenhausen.
- Jansen, H., Mannhaupt, G., Marx, H., & Skowronek, H. (2002). *Bielefelder Screening zur Früherkennung von Lese-Rechtschreibschwierigkeiten (BISC)*. Göttingen: Hogrefe.
- Kim, Y.-S., Petscher, Y., Foorman, B. R., & Zhou, C. (2010). The contributions of phonological awareness and letter-name knowledge to letter-sound acquisition—A cross-classified multilevel model approach. *Journal of Educational Psychology*, 102(2), 313–326.

- Landerl, K., Linortner, R., & Wimmer, H. (1992). Phonologische Bewusstheit und Schriftspracherwerb im Deutschen. *Zeitschrift für Pädagogische Psychologie*, 6, 17–33.
- Landerl, K., Wimmer, H., & Frith, U. (1997). The impact of orthographic consistency on dyslexia: A German-English comparison. *Cognition*, 63(3), 315–334.
- Lonigan, C. J., Schatschneider, C., Westberg, L., & the National Early Literacy Panel. (2008). *Results of the National Early Literacy Panel research synthesis: Identification of children's skills and abilities linked to later outcomes in reading, writing, and spelling. Report of the National Early Literacy Panel*. Louisville, KY: Author.
- Lundberg, I. (2009). Early precursors and enabling skills of reading acquisition. *Scandinavian Journal of Psychology*, 50, 611–616.
- Lundberg, I., Larsman, P., & Strid, A. (2012). Development of phonological awareness during the preschool year: The influence of gender and socio-economic status. *Reading & Writing*, 25(2), 305–320.
- Mann, V. (1986). Phonological awareness: The role of reading experience. *Cognition*, 24(1-2), 65–92.
- Mannhaupt, G. (2006). *Münsteraner Screening zur Früherkennung von Lese-Rechtschreibschwierigkeiten (MÜSC)*. Berlin: Cornelsen.
- Mannhaupt, G., & Jansen, H. (1989). Phonologische Bewusstheit: Aufgabenentwicklung und Leistungen im Vorschulalter. *Heilpädagogische Forschung*, 15(1), 50–56.
- Martschinke, S., Kammermeyer, G., King, M., & Forster, M. (2005). *Anlaute hören, Reime finden, Silben klatschen (ARS): Erhebungsverfahren zur phonologischen Bewusstheit für Vorschulkinder und Schulanfänger*. Donauwörth: Auer.
- Martschinke, S., Kirschhock, E.-M., & Frank, A. (2001). *Der Rundgang durch Hörhausen: Erhebungsverfahren zur phonologischen Bewusstheit* (Vol. 1). Donauwörth: Auer.
- Marx, P., Weber, J., & Schneider, W. (2005). Phonologische Bewusstheit und ihre Förderung bei Kindern mit Störungen der Sprachentwicklung. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, 37(2), 80–90.
- McDowell, K. D., Lonigan, C. J., & Goldstein, H. (2007). Relations among socioeconomic status, age, and predictors of phonological awareness. *Journal of Speech, Language, and Hearing Research*, 50(4), 1079–1092.
- Melby-Lervåg, M., Lyster, S.A.H., & Hulme, C. (2012). Phonological skills and their role in learning to read: A meta-analytic review. *Psychological Bulletin*, 138(2), 322–352.
- Melchers, P., & Preuß, U. (2009). *Kaufmann-Assessment Battery for Children (K-ABC); Deutsche Version*. Göttingen: Hogrefe.
- Moyle, M. J., Heilmann, B., & Berman, S. S. (2013). Assessment of early developing phonological awareness skills: A comparison of the preschool individual growth and development indicators and the phonological awareness and literacy screening—PreK. *Early Education & Development*, 24(5), 668–686.
- Muter, V., Hulme, C., Snowling, M., & Taylor, S. (1998). Segmentation, not rhyming, predicts early progress in learning to read. *Journal of Experimental Child Psychology*, 71(1), 3–27.

- Nagy, W. E., & Anderson, R. C. (1995). *Metalinguistic awareness and literacy acquisition in different languages* (Technical Report No. 618). Urbana-Champaign, IL: Center for the Study of Reading, University of Illinois at Urbana-Champaign.
- National Institute of Child Health and Human Development (NICHD). (2000). *Report of the National Reading Panel. Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction: Reports of the subgroups* (NIH Publication No. 00-4754). Washington, DC: U.S. Government Printing Office.
- Nithart, C., Demont, E., Metz-Lutz, M., Majerus, S., Poncelet, M., & Leybaert, J. (2011). Early contribution of phonological awareness and later influence of phonological memory throughout reading acquisition. *Journal of Research in Reading, 34*(3), 346–363.
- Paulus, C. (2009). *Die "Bücheraufgabe" zur Bestimmung des kulturellen Kapitals bei Grundschulern*. Saarbrücken: Universität des Saarlandes, Fachrichtung Erziehungswissenschaft. Retrieved from: http://bildungswissenschaften.uni-saarland.de/personal/paulus/Artikel/BA_Artikel.pdf
- Perfetti, C. A. (1985). *Reading ability*. New York: Oxford University Press.
- Pfost, M. (2015). Children's phonological awareness as a predictor of reading and spelling: A systematic review of longitudinal research in German-speaking countries. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 47*(3), 123–138.
- Redford, M. A., & Diehl, R. L. (1999). The relative perceptual distinctiveness of initial and final consonants in CVC syllables. *Journal of the Acoustical Society of America, 106*, 555–1565.
- Schäfer, B., Bremer, M., & Herrmann, F. (2014). Onset and phoneme awareness and its relationship to letter knowledge in German-speaking preschool children. *Folia Phoniatrica et Logopaedica, 66*, 126–131.
- Schäfer, B., Fricke, S., Szczerbinski, M., Fox-Boyer, A., Stackhouse, J., & Wells, B. (2009). Development of a test battery for assessing phonological awareness in German-speaking children. *Clinical Linguistics & Phonetics, 23*(6), 404–430.
- Schäfer, B., Stackhouse, J., & Wells, B. (n. d.). *The development of phonological awareness in German-speaking preschool children: A longitudinal study*. Manuscript in preparation.
- Schäfer, B., Wessels, S., & Fricke, S. (2014). Phonologische Bewusstheit bei Dreijährigen. *Sprache – Stimme – Gehör, 38*, 1–5.
- Schnitzler, C. D. (2008). *Phonologische Bewusstheit und Schriftspracherwerb*. Stuttgart: Thieme Verlag.
- Smith, S. B., Simmons, D. C., & Kameenui, E. J. (1998). Phonological awareness: Research bases. In D. C. Simmons, & E. J. Kameenui (Eds.), *What reading research tells us about children with diverse learning needs* (pp. 61–128). Mahwah, NJ: Lawrence Erlbaum.
- Stackhouse, J., & Wells, B. (1997). *Children's speech and literacy difficulties. A psycholinguistic framework*. London: Whurr Publishers.

- Stanovich, K. E., Cunningham, A. E., & Cramer, B. B. (1984). Assessing phonological awareness in kindergarten children: Issues of task comparability. *Journal of Experimental Child Psychology*, 38(2), 175–190.
- Stock, C., Marx, P., & Schneider, W. (2003). *Basiskompetenzen für Lese-Rechtschreibleistungen. Ein Test zur Erfassung der phonologischen Bewusstheit vom ersten bis vierten Grundschuljahr (BAKO 1-4)*. Göttingen: Beltz Test.
- Suggate, S. P., Reese, E., Lenhard, W., & Schneider, W. (2014). The relative contributions of vocabulary, decoding, and phonemic awareness to word reading in English versus German. *Reading and Writing: An Interdisciplinary Journal*, 27(8), 1395–1412.
- Tertiary Education Commission (2008). *Starting Points—Supporting the learning progressions for adult literacy*. Retrieved from <http://www.tec.govt.nz/Documents/Publications/Learning-progressions-starting-points.pdf>
- Tewes, U., Rossmann, P., & Schallberger, U. (1999). *Hamburg Wechsler Intelligenztest (HAWIK III)*. Bern: Huber.
- Treiman, R. (1992). The role of intrasyllabic units in learning to read and spell. In P. B. Gough, L. Ehri, & R. Treiman (Eds.), *Reading acquisition* (pp. 65–106). Hillsdale, NJ: Erlbaum.
- Treiman, R., & Zukowski, A. (1991). Levels of phonological awareness. In S. Brady, & D. P. Shankweiler (Eds.), *Phonological processes in literacy* (pp. 67–83). Hillsdale: Erlbaum.
- Treiman, R., & Zukowski, A. (1996). Childrens sensitivity to syllables, onsets, rimes and phonemes. *Journal of Experimental Child Psychology*, 61, 193–215.
- Tunmer, W. E., & Hoover, W. A. (1992). Cognitive and linguistic factors in learning to read. In P. B. Gough, L. E. Ehri, & R. Treiman (Eds.), *Reading acquisition* (pp. 175–214). Hillsdale, NY: Lawrence Erlbaum Associates.
- von Goldammer, A., Mähler, C., & Hasselhorn, M. (2011). Determinanten von Satzgedächtnis-Leistungen bei deutsch- und mehrsprachigen Vorschulkindern. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, 43(1), 1–15.
- von Goldammer, A., Mähler, C., Bockmann, A., & Hasselhorn, M. (2010). Vorhersage früher Schriftsprachleistungen aus vorschulischen Kompetenzen der Sprache und der phonologischen Informationsverarbeitung. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, 41(1), 48–56.
- Wagner, R. K., Torgesen, J. K., Rashotte, C. A., Hecht, S. A., Barker, T. A., Burgess, S. R., ... Garon, T. (1997). Changing relations between phonological processing abilities and word-level-reading as children develop from beginning to skilled learners: A 5-year longitudinal study. *Developmental Psychology*, 33(3), 468–479.
- Weinert, S., Artelt, C., Prenzel, M., Senkbeil, M., Ehmke, T., & Carstensen, C. H. (2011). Development of competencies across the life span. In H.-P. Blossfeld, H.-G. Roßbach, & J. von Maurice (Eds.), *Zeitschrift für Erziehungswissenschaft*, 14, *Education as a lifelong process: The German Educational Panel Study (NEPS)* (pp. 67–86). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Weinert, S., & Berendes, K. (2012). *Competence measurement and test administrator training in the Kindergarten cohort of the National Educational Panel Study*. Poster presen-

- tation at the EUCCONET (European Child Cohort Network) & SLLS (Society for Longitudinal and Life Course Studies) International Conference, Paris, France.
- Wentink, H. W. M., van Bon, W. H. J., & Schreuder, R. (1997). Training of poor readers' phonological decoding skills: Evidence for syllable-bound processing. *Reading and Writing: An Interdisciplinary Journal*, 9(3), 163–192.
- Wimmer, H., & Goswami, U. (1994). The influence of orthographic consistency on reading development: Word recognition in English and German children. *Cognition*, 51(1), 91–103.
- Wimmer, H., Landerl, K., & Schneider, W. (1994). The role of rhyme awareness in learning to read a regular orthography. *British Journal of Developmental Psychology*, 12(4), 469–484.
- Wood, C., & Terrell, C. (1998). Preschool phonological awareness and subsequent literacy development. *Educational Psychology: An International Journal of Experimental Educational Psychology*, 18(3), 253–274.
- Yavas, M., & Gogate, L. (1999). Phoneme awareness in children: A function of sonority. *Journal of Psycholinguistic Research*, 28(3), 245–260.
- Yopp, H. K. (1988). The validity and reliability of phonemic awareness tests. *Reading Research Quarterly*, 23(2), 159–177.
- Ziegler, J. C., & Goswami, U. (2005). Reading acquisition, developmental dyslexia, and skilled reading across languages: A psycholinguistic grain size theory. *Psychological Bulletin*, 131(1), 3–29.
- Ziegler, J. C., Perry, C., Jacobs, A. M., & Braun, M. (2001). Identical words are read differently in different languages. *Psychological Science*, 12(5), 379–384.

About the authors

K. Berendes
Hector Research Institute of Education Science and Psychology,
University of Tübingen, Europastraße 6, 72072 Tübingen, Germany.
e-mail: karin.berendes@uni-tuebingen.de

S. Weinert
Department of Psychology I: Developmental Psychology,
University of Bamberg, Markusplatz 3, 96047 Bamberg, Germany.

Assessing Spelling Competence Development in the National Educational Panel Study

Stephan Jarsinski, Sarah Frahm, Inge Blatt, Wilfried Bos and Michael Kandera

Abstract

In the National Educational Panel Study (NEPS), spelling competence is a stage-specific measure for secondary school (Stage 4). This is a challenging task because the current state of research does not offer a theory-based and empirically proven instrument that would allow for longitudinal measurement. In this paper, we present results from NEPS pilot studies in order to illustrate our work. For our research, we used an anchor-item design and IRT-based methods to verify the suitability of the test for the longitudinal survey of spelling competence. We demonstrate the reliability of the tests with regard to dimensionality and discrimination. The person parameter and the item parameter provide an insight into the item difficulty and the students' abilities, thereby guaranteeing that the tests account for individual demands throughout secondary school. Moreover, deviance and the correlational structure of the data are taken into consideration. In summary, the selected test design and the choice of anchor items ensure that an adequate test is administered to each student. We removed about 15–30 % of the items in each grade to obtain a fit model. As the results for Grades 5 to 7 are nearly identical, it can be assumed that the test is reliable. The assumed five-dimensional model proves to be most adequate to measure the subskills for all grades because the deviance is lower.

1 Introduction

In the National Educational Panel Study (NEPS), spelling competence is a stage-specific measure for secondary school. It complements the obligatory measurement of the core areas of reading and mathematics competence because it also constitutes a central aspect of educational success. However, in order to measure spelling competence adequately as well as longitudinally, fundamental research is necessary. This is

due to a change in the linguistic and didactic understanding of spelling during the past decade, resulting in a lack of theory-based and empirically proven tests that allow for the measurement of progress in competence development in spelling (Frahm & Blatt, 2011).

In order to fill this research gap, we focus on spelling as a central stage-specific research field. With an interdisciplinary team of didactic and educational researchers, we adapted a test based on a linguistic framework by Blatt and Voss that was conducted 2006 in a German add-on study of Progress in International Reading Literacy (PIRLS) in Grade 4 (Voss, Blatt & Kowalski, 2007). We changed the test format and developed a less time-consuming, computer-based coding tool. In order to measure the spelling competence longitudinally, we used an anchor-item design. We used methods based on item-response theory for our analysis.

This paper focuses on two research questions:

- 1) How can reliable tests be developed that facilitate a longitudinal data collection and that account for individual demands throughout secondary school?
- 2) Does the structure of spelling competencies of students change over time, and if so, in what way?

In order to answer these questions, we briefly outline the theoretical framework of the test along with the test itself, followed by a description of the test design and methodology. Afterwards, we present first results for Grades 5 through 7 based on preliminary or experimental studies from the NEPS in order to answer our research questions. We then conclude our findings with some remarks on future research.

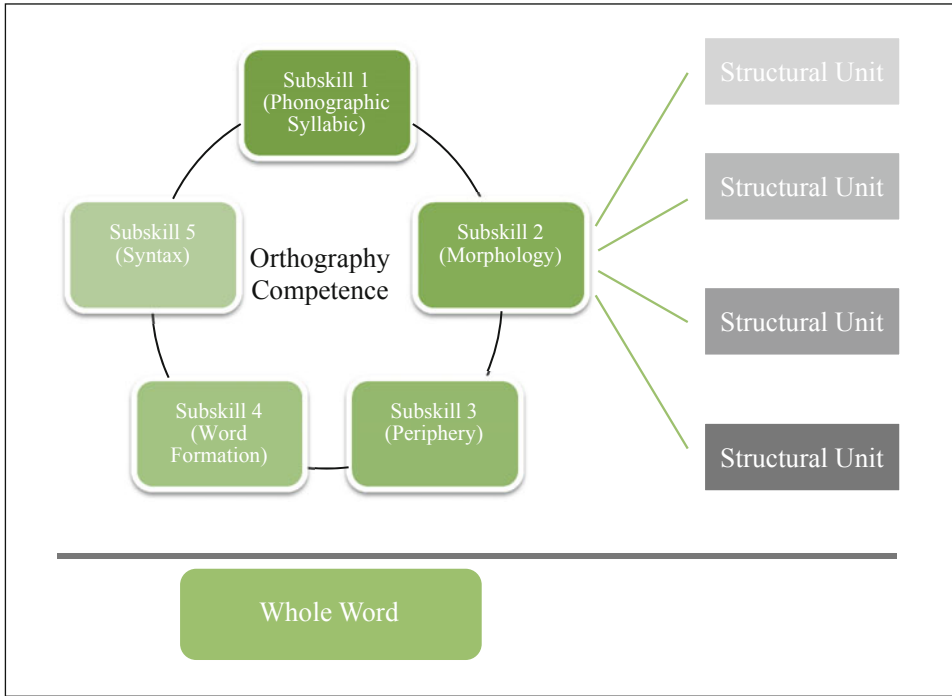
2 Theoretical Framework and Test Development

The framework and test development for the orthography competence test have already been described in Blatt, Voss, Kowalski, and Jarsinski (2011) and in Frahm et al. (2011). Therefore, we give only a brief outline of the framework and the test used in the NEPS.

The framework (Figure 1) distinguishes between five subskills of orthography (phonographic syllabic, morphological, peripheral, derivational, and syntactic subskills). In order to measure these subskills, structural units of words (i. e., reality: #real #ity) are assigned to them. On top of this distinguished model, each word is also assessed at the whole-word level. Therefore, one item can either be a structural unit or a whole word depending on which level is being analyzed. Each level offers information that differs in its preciseness, the whole-word level being less precise.

The five subskills consist of 30 to 60 structural units. According to previous research, a five-dimensional model has proven to be most adequate for modeling the data on the structural-unit level (Voss et al., 2007).

Figure 1 Theoretical framework of spelling competence structure



The words used in the test relate to the curriculum. The content changes due to the emphasis of the syntactic and peripheral subskills in higher grades. New content is added over the course of secondary education—for example, punctuation.

The test consists of a combined cloze test and sentence dictation, which makes use of a compact disc (CD) for the dictation. The test data are first transcribed by the International Association for the Evaluation of Educational Achievement Data Processing Center (IEA DPC) using transcription conventions that were established in the context of the PIRLS study (Frahm et al., 2011). The transcribed data are then coded by a newly developed computer-based tool (SRT-Editor), which facilitates immediate analysis (Frahm et al., 2011).

3 Design and Methodology

We chose the framework and the test content, as described above, to map spelling development in a complex way. As a result, the test accounts for individual demands in different grades. In this paper, we stress the importance of an adequate test design for a valid longitudinal data collection as well as the importance of applying statistical

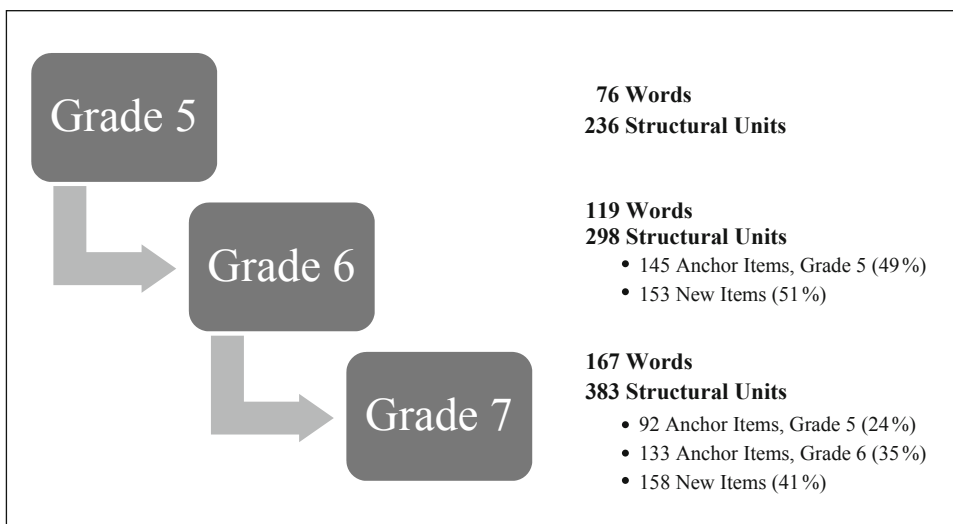
methods to ensure the quality criteria for test development reliability. These methods are well established for studies such as PIRLS and PISA. The design and the methodology are described in the following section.

3.1 Longitudinal Test Design

The test design is an anchor-item design: The test consists of anchor items that bear a relationship to and between each survey. In this way, the development of spelling competence can be determined (Hambleton, Swaminathan, & Rogers, 1991; Lord, 1980). The test is continuously redeveloped for the next grade on the basis of prior results and the curriculum. This process ensures an adequate selection of items based on prior analyses and a variable test development.

The longitudinal data were collected using an anchor-item design (Figure 2). The test for Grade 5 consists of 76 words and 236 structural units. We chose 145 items of these units as anchor items for Grade 6. They were combined with 153 new structural units. Parallel to this item selection, the test in Grade 7 merges anchor items and new items. In order to maintain a relationship with Grade 5 as well as with Grade 6, 59 % of structural units (Grade 5: 24 %; Grade 6: 35 %) were kept in Grade 7.

Figure 2 Test contents



3.2 Methodology

The data analysis is based on item response theory (IRT, Bortz & Döring, 2006). IRT refers to a one- or multiparametric logistic function of person and item parameters and aims to estimate the probability of a correct response (item parameter) as a function of ability (person parameter) (Rasch, 1960). Both parameters can be compared with the item characteristic curve (ICC). Generally, the manifest items feature homogeneity. Hence, it is assumed that the items are measurable on one latent scale (Moosbrugger, 2007). Specific objectivity is of importance for this purpose; the estimation must lead to the same result regardless of item selection and the sample (Bortz & Döring, 2006; Moosbrugger, 2007). In contrast with classical test theory (CTT), IRT is a valid method for modeling data and evaluating how well assessments work. Therefore, its most common application is in educational research (Voss, 2006). Furthermore, IRT is also well established and useful for longitudinal analyses (Moosbrugger, 2007).

Item difficulty and ability can be estimated with the ConQuest software (Wu, Adams, Wilson, & Haldane, 2007) using Rasch's simple logistic model. Just as is carried out by PISA, these items are usually scaled with a mean of 500 and a standard deviation of 100 (OECD, 2005).

In addition, psychometric properties can be investigated by IRT. These properties are unidimensionality and the discriminatory power of the items, reliabilities, and latent correlations of subskills, as well as a comparison of different models based on deviance statistics.

In order to check for unidimensionality, item fits are used as indicators that are determined by weighted mean squares with ConQuest. In line with PISA, the weighted mean square is expected to be 1, thereby allowing for an interval of 0.80–1.20 (Adams, 2002).

Furthermore, discrimination is reviewed. It shows whether or not students with different abilities solve an item. According to earlier beliefs, the discrimination criterion was expected to be higher than 0.25 (OECD, 2005).

The reliability of a test is assumed to be the test's main criterion in the field of psychometrics. Reliability is concerned with the overall consistency of a measure, or rather, how accurately a latent trait is measured within a test. Reliability must be higher than 0.70 (Moosbrugger, 2007).

Latent correlations give information on the co-variation of the test result based on subskills. They show quantified latent relations of different subskills. A high correlation coefficient indicates redundant information. In this case, it is not necessary to differentiate between subskills.

Another criterion for analyzing a competence model is based on deviance statistics that compare the dimensionality of competence models with different complexity. Therefore, deviance is a measure for verifying a theoretical model and its struc-

ture with empirical data. A low deviance indicates higher explanatory power for the information given with the data (Voss, Carstensen, & Bos, 2005).

The methods described above do not only serve the need for adequate test development, but they are also employed to analyze the development of a specific competence longitudinally, which is another topic under discussion in this paper.

In relation to the research questions stated above, we find that research question number one must first be answered from a theoretical point of view. The differential tests are adequate for the individual measurement of spelling in secondary school. The curriculum-based content and the anchor-item design offer an ideal mapping of spelling development. From an empirical point of view, the suitability of the test for longitudinal survey of the spelling competence must still be verified.

The data we used for this paper were drawn from preliminary studies of Grades 5 through 7. However, it is our major aim to use these methods for the longitudinally analysis of the main sample, as well.

4 Data and Results

Data

The data consist of three measurement points from the preliminary or experimental studies in Grade 5 (2009), Grade 6 (2011), and Grade 7 (2011), with about 300 cases (Grade 5/6/7: $N = 298/414/307$) for each grade. Grades 6 and 7 mainly consist of the same population. The data allow for longitudinal analyses with 307 cases.

Results

With regard to the first research question, the item fit and discrimination and the reliability at both levels are taken into consideration. Then, item difficulty and student ability at the whole-word level are mapped.

The second research question is answered on the basis of structural units that present the competence structure. We use maps to outline the results. Furthermore, deviance and correlations are taken into consideration.

Item fit and discrimination

During the estimation of student ability and item difficulty for students in Grades 5 through 7, we removed items from the test design (see Figure 1) for each grade. First, those items that were consistently correct—for example, easy words such as “und” (“and”)—were removed. Second, we deleted those items that deviated from the PISA reference of an item-fit between 0.80 and 1.20 and had a discrimination of less than 0.26.

In Grade 5, we removed five out of 54 items (Table 1). In Grade 6, we removed six items of the initial 75 items, yielding a final number of 69 items. We identified 14 misfit items in Grade 7, leading to a final number of 78 items.

Table 1 Item Misfit—Whole Word

	Original	Optimized
Grade 5	54 words	49 words
Grade 6	75 words	69 words
Grade 7	92 words	78 words

Table 2 Item Misfit—Structural Units

	Original	Optimized
Grade 5	272 structural units	198 structural units
Grade 6	299 structural units	196 structural units
Grade 7	384 structural units	281 structural units

Just as we did at the whole-word level, we also removed misfit items at the structural-unit level.

In Grade 5, we removed 74 items (Table 2). In Grade 6, we eliminated 103 items, leading to a final number of 196 structural units. We also identified 103 misfit items in Grade 7, leading to a final number of 281 items.

In summary, we removed about 10 % of the items at the whole-word level and about 30 % of the items at the structural-unit level in each grade to obtain a fit model.

Reliability

In order to prove the reliability of the tests, the reliabilities of the original and the optimized models for Grades 5 through 7 are presented below, exemplified for the structural-unit level. After removing misfit items, the reliability for all subskills was still higher than the PISA reference of 0.70.

As a result of reducing the number of misfit items, the reliability decreased in most cases (Table 3). Moreover, student ability and the variance of item difficulty decreased slightly because the removed items were mostly too easy. In summary, as the results for Grades 5 to 7 are nearly identical, it can be assumed that the test is reliable.

Item difficulty and student ability

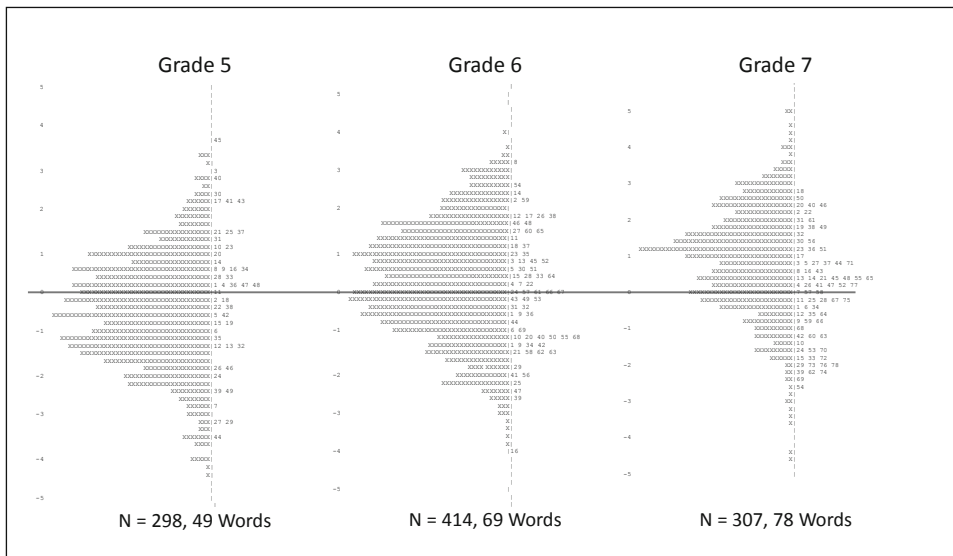
Figure 3 presents the item difficulty and the student ability at the whole-word level for Grades 5 to 7. The ability is shown on the left side, and the item difficulty is shown on the right side. The zero point is indicated by a line.

The items are largely distributed between -3 and $+3$. The variance of the item difficulty is near 2 and increases with higher grades. This underlines the fact that the test caters to low- as well as to high-achieving students. In Grade 5, the average student ability is slightly negative, with a mean of -0.46 , and increases for Grade 6 to 0.28 and for Grade 7 to 0.87 . With increasing student ability, it is vitally important to use items for all students. With the selected test design and the choice of anchor items being the most difficult ones in each grade, we managed to ensure that an adequate test was administered to each student.

Table 3 Reliability

		Phonographic-Syllabic Principle	Morphological Principle	Peripheral Area	Word Formation	Syntactic Principle
Grade 5 (N = 298)	Original (272 struc. units)	0.937	0.926	0.920	0.938	0.859
	Optimized (198 struc. units)	0.927	0.920	0.920	0.932	0.880
Grade 6 (N = 414)	Original (299 struc. units)	0.952	0.945	0.947	0.948	0.922
	Optimized (196 struc. units)	0.926	0.937	0.950	0.945	0.920
Grade 7 (N = 307)	Original (384 struc. units)	0.927	0.927	0.941	0.905	0.946
	Optimized (281 struc. units)	0.918	0.898	0.934	0.945	0.907

Figure 3 Item difficulty and student ability, Grades 5 to 7



The results emphasize the necessity of removing misfit items and selecting the most difficult items from each grade as anchor items. The test difficulty reveals that the test suits the students' competence levels.

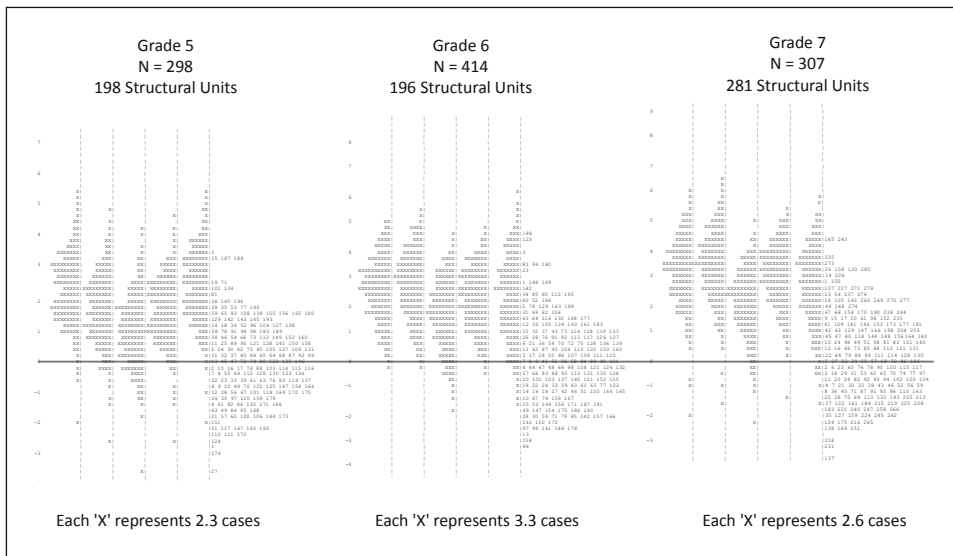
Competence structure

The results concerning the distribution of item difficulty and student ability for each subskill within the competence structure comparing Grades 5 through 7 are shown based on the structural-unit level.

Student ability was quite high overall. In terms of item difficulty, it can be seen that the tests still offered easy as well as difficult items within the range of -3 to $+3$. Compared with the whole-word level, the students solved more structural-unit items than whole-word items. This means that although the students succeeded in writing subskills correctly, they did not manage to write the whole word without any mistakes.

Student ability of Subskill 3 has the lowest mean for all grades. In Grade 5, Subskills 1 and 5 have the highest mean. In Grades 6 and 7, Subskills 1 and 2 have the highest mean. Subskills 3 to 5 remained nearly identical in comparison with Grade 5. Additionally, it can be seen that the relationship of the subskills changed. This relationship is analyzed in the following section.

Figure 4 Competence structure based on structural units



Deviance

The deviance value was used to analyze whether a one-dimensional or a multidimensional model would better represent the collected data at the structural-unit level. Prior results, such as those by PIRLS, showed that a five-dimensional model was most suitable to represent the theoretical framework; however, correlation analyses from different studies demonstrated that all five subskills correlate very highly (Blatt et al., 2011). In order to find a suitable model for the data, different models were analyzed. In line with a didactic and linguistic point of view, we deployed a five-, four-, two-, and one-dimensional model to compare deviance. In a first approach to represent the data differently, a four-dimensional model was applied. In this model, the first two subskills were combined. The next step was a further reduction of dimensions down to a two-dimensional model that differentiated the word-related subskills (1 to 4) and the syntax-related subskill (5). We also used a one-dimensional model.

For all grades, the five-dimensional model has proven to be most adequate for measuring the subskills because the deviance is lower (Table 4). However, the correlation structure must still be taken into consideration.

Correlation

The correlation structure is visualized in Table 5. The triangle matrix represents the correlations for Grade 6 in the upper half. The lower half displays the correlation of subskills in Grade 7.

Except for the syntactic subskill (Subskill 5), all correlations in Grade 6 exceed 0.95 (Table 5). In Grade 7, the same result can be seen, in part even with higher correlations. The syntactic subskill is the most independent one. The high correlations of the first four subskills suggest that they can be merged into one subskill. It is the syntactic principle alone that shows lower correlations with the other subskills. These results speak in favor of a 1–4 + 5 solution resulting in a two-dimensional model with a word-related and a syntax-related dimension.

5 Conclusion and Outlook

This paper confirms that the orthography test is adequate for a longitudinal measurement of spelling competence in the NEPS. This is true from both a theoretical and statistical point of view.

In previous works, we have shown that the differential tests are theoretically adequate for the individual measurement of spelling in secondary school and that the curriculum-based content is useful in this respect.

In this paper, we have outlined the fact that the anchor-item design is also suitable for a longitudinal measurement. In addition, our analysis has proven that the choice of anchor items ensures that an adequate test is administered to each student, even

Table 4 Deviance—Grades 5 to 7

	5D	4D	2D	1D
Grade 5 (<i>N</i> = 298, 198 structural units)	45889.30687	45904.93155	45968.39119	46008.54185
Grade 6 (<i>N</i> = 414, 196 structural units)	53963.00905	53981.63053	54041.54590	54428.04026
Grade 7 (<i>N</i> = 307, 281 structural units)	45445.07219	45499.76225	45619.60747	45841.76617

Table 5 Correlations 5D, Grades 6 to 7

	Phonographic-syllabic principle	Morphological principle	Peripheral area	Word formation	Syntactic principle
Phonographic-syllabic principle		0.972	0.958	0.964	0.839
Morphological principle	0.979		0.982	0.953	0.841
Peripheral area	0.938	0.948		0.971	0.838
Word formation	0.958	0.945	0.966		0.871
Syntactic principle	0.907	0.884	0.895	0.935	

Correlation of subskills in a five-dimensional model.

Pilot study Grade 6: upper half. Pilot study Grade 7: lower half.

after removing a low number of misfit items. The reliability of the tests in the pilot studies is also adequate.

A detailed investigation into the development of five subskills stresses that a differentiated test is necessary to gain a deeper insight into the development of the spelling competence. It has become clear that it is not sufficient to focus only on whole words in order to thoroughly analyze longitudinal development.

The developmental processes have a great research potential because it is not yet clear how competence changes can be modeled in accordance with the theoretical framework. Whether a five- or a two-dimensional model is more useful is a question

that needs to be discussed in more detail in the future. This will become part of our work using the data of the main studies.

The transcribed test data also offer a potential for further qualitative analysis. Moreover, the large variety of context information conducted in the NEPS—that is, test results in other domains, information on language instruction, and engagement—provide a database for a thorough analysis of influencing factors on spelling competence. The results are of importance not only for large-scale assessments, but also for didactic research. The next step is to verify the results presented above with the findings of the NEPS main studies. However, these statistical processes are not the only steps towards developing a reliable test for longitudinal measurement. What must be stressed here is that previous theoretical work and the development of a framework are important prerequisites for a successful test development.

References

- Adams, R. J. (2002). Scaling PISA cognitive data. In R. J. Adams, & M. Wu (Eds.), *PISA 2000 technical report* (pp. 99–108). Paris: OECD.
- Blatt, I., Voss, A., Kowalski, K., & Jarsinski, S. (2011). Messung von Rechtschreibleistung und empirische Kompetenzmodellierung. In U. Bredel (Ed.), *Weiterführender Orthographieunterricht* (pp. 226–256). Baltmannsweiler: Schneider Verlag Hohengehren.
- Bortz, J., & Döring, N. (2006). *Forschungsmethoden und Evaluation für Sozialwissenschaftler*. Heidelberg: Springer.
- Frahm, S., & Blatt, I. (2011). Rechtschreibtests. In U. Bredel (Ed.), *Weiterführender Orthographieunterricht* (pp. 546–567). Baltmannsweiler: Schneider Verlag Hohengehren.
- Frahm, S., Goy, M., Kowalski, K., Sixt, M., Strietholt, R., Blatt, I., ... Kanders, M. (2011). Transition and development from lower secondary to upper secondary school. In H.-P. Blossfeld, H.-G. Rossbach, & J. von Maurice (Eds.), *Zeitschrift für Erziehungswissenschaft, 14. Education as a lifelong process: The German National Educational Panel Study (NEPS)* (pp. 217–232). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park: Sage Productions.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.
- Moosbrugger, H. (2007). Item-Response-Theorie (IRT). In H. Moosbrugger, & A. Kelava (Eds.), *Testtheorie und Fragebogenkonstruktion* (pp. 215–259). Berlin: Springer.
- OECD. (2005). *PISA 2003 technical report*. Paris: OECD.
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Nielsen & Lydiche.
- Voss, A. (2006). *Print- und Hypertextlesekompetenz im Vergleich. Eine Untersuchung von Leistungsdaten aus der Internationalen Grundschul-Lese-Untersuchung (IGLU) und der Ergänzungsstudie Lesen am Computer (LaC)*. Münster: Waxmann.

- Voss, A., Blatt, I., & Kowalski, K. (2007). Zur Erfassung orthographischer Kompetenz in IGLU 2006. *Didaktik Deutsch*, 23, 15–33.
- Voss, A., Carstensen, C., & Bos, W. (2005). Textgattungen und Verstehensaspekte: Analyse von Leseverständnis aus den Daten der IGLU-Studie. In W. Bos, E.-M. Lankes, M. Prenzel, K. Schwippert, R. Valtin, & G. Walther (Eds.), *IGLU. Vertiefende Analysen zum Leseverständnis, Rahmenbedingungen und Zusatzstudien* (pp. 1–36). Münster: Waxmann.
- Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. A. (2007). *ACER ConQuest Version 2.0 generalised item response modelling software*. Melbourne: Acer Press.

About the authors

I. Blatt
University of Hamburg.

W. Bos
Institute for School Development Research (IFS), TU Dortmund University.

S. Frahm
University of Hamburg.

S. Jarsinski
Institute for School Development Research (IFS), TU Dortmund University.
e-mail: stephan.jarsinski@tu-dortmund.de

M. Kanders
Institute for School Development Research (IFS), TU Dortmund University.

Assessment of Immigrant Students' Listening Comprehension in Their First Languages (L1) Russian and Turkish in Grade 9: Test Construction and Validation¹

Aileen Edele, Kristin Schotte and Petra Stanat

Abstract

In large-scale studies, immigrant students' first-language (L1) proficiency is typically measured with subjective instruments, such as self-reports, rather than with objective tests. The National Educational Panel Study (NEPS) addresses this methodological limitation by testing the L1 proficiency of the two largest immigrant groups in Germany, namely students whose families have immigrated to Germany from the area of the Former Soviet Union or Turkey. Listening comprehension tests in Russian and Turkish were developed for this purpose. The current paper describes the general framework and requirements for testing first-language proficiency within the NEPS and describes the construction of the L1 tests for 9th-Grade students. Subsequently, the paper reports on analyses of measurement equivalence indicating that the Russian and Turkish tests assess the same construct (configural equivalence). The ability scores and their correlations with other variables are, however, not directly comparable. Analyses of construct validity confirm the unidimensional structure expected for the test. In addition, the L1 test scores correlate with other indicators of L1 proficiency as well as with factors regarded as crucial for L1 acquisition, such as exposure to L1, in the expected way (convergent validity), and they are not substantially related to measures of general cognitive abilities (discriminant validity). We conclude that the listening comprehension tests developed in the NEPS are valid measures of L1 proficiency.

1 The content of this article significantly overlaps with the following publication; parts of the text have been incorporated verbatim without any further indication: Edele, A., Schotte, K., & Stanat P. (2015). *Listening comprehension tests of immigrant students' first languages (L1) Russian and Turkish in grade 9: Extended report of test construction and validation*. (NEPS Working Paper No. 57). Bamberg: University of Bamberg, National Educational Panel Study.

1 Introduction

The effects immigrant students' first-language (L1) proficiency may have on their social integration and educational success are highly disputed.² On the one hand, some theoretical perspectives and findings suggest positive effects of L1 proficiency on second-language acquisition (e.g., Cummins, 2000; Scheele, Leseman, & Mayo, 2010; Verhoeven, 2007) as well as on third-language learning (Hesse, Göbel, & Hartig, 2008; Rauch, Jurecka, & Hesse, 2010), and bilingualism is assumed to promote cognitive development (Adesope, Lavin, Thompson, & Ungerleider, 2010; Bialystok, 2007). On the other hand, neutral or negative effects of proficiency in L1 are proposed (e.g., Dollmann & Kristen, 2010; Esser, 2006; Mouw & Xie, 1999).

The empirical evidence necessary to elucidate this controversy is, however, inconclusive. This is also a result of the methodological constraints of most studies on this issue. In particular, there is a lack of investigations assessing L1 proficiency with objective tests rather than with subjective measures, especially when it comes to analyses with larger sample sizes. Previous large-scale studies have typically relied on self-report measures of L1 proficiency (e.g., Mouw & Xie, 1999; Portes & Rumbaut, 2012).

The National Educational Panel Study (NEPS; Blossfeld, Roßbach, & von Maurice, 2011) set out to address this research gap by measuring the L1 proficiency of students from the two largest immigrant groups in Germany with objective tests. The project assesses the listening comprehension proficiency of children and adolescents whose families immigrated to Germany from the area of the Former Soviet Union or Turkey. As no suitable instruments for this purpose were available, tests in Turkish and Russian were developed by the Berlin project team within Pillar 4 of the NEPS.³

This paper describes the construction of the L1 tests for 9th-Grade students and reports analyses exploring the tests' validity. The following section delineates the general framework for testing L1 within the NEPS and the requirements the instruments had to meet in order to be suitable for implementation in the study. In section 3, we describe the tests developed for students in Grade 9. Section 4 reports analyses pertaining to the tests' validity.

2 In line with the terminology commonly used in the literature, the term immigrant students refers to the first, second, and third immigrant generation in this paper. The term first language (L1) is used here interchangeably with the language spoken in an immigrant family's country of origin, regardless of whether this language was actually acquired prior to the language of the destination country (in our case, German), or simultaneously.

3 The tests for Grade 9 were developed by Aileen Edele and Petra Stanat.

2 General Framework and Requirements for Assessing L1 Proficiency Within the NEPS

The NEPS assesses L1 proficiency with objective tests at three measurement points. More specifically, the instruments capture listening comprehension skills in Russian and Turkish in Grade 9 (Starting Cohort 4 and later in Starting Cohort 3⁴), in Grade 7 (Starting Cohort 3), and in Grade 2 (Starting Cohort 2). The current paper focuses on the tests developed for the 9th Grade.

2.1 Defining and Identifying the Target Population

The NEPS set out to assess the L1 proficiency of the two largest immigrant groups in Germany, namely students from families who have immigrated to Germany from the area of the Former Soviet Union (e. g., Russia, Kazakhstan) or from Turkey. To ensure that all students from these immigrant groups participating in the NEPS are included in the initial sample, the target population is defined as students of the first, second, and third immigrant generation.⁵

However, not all immigrants maintain their heritage language (Rumbaut, Massey, & Bean, 2006; Strobel & Kristen, 2015), and competence testing is, of course, only meaningful if test-takers possess some proficiency in the tested domain. Therefore, we implemented screening tests with very low item difficulty prior to L1 testing. Only students who demonstrated a minimal level of listening comprehension in these tests were asked to participate in the actual L1 assessment (see section 4.2 for further information).

2.2 Efficiency

As the NEPS assesses a large number of constructs, the testing time available for each competence domain is limited. As a consequence, it was impossible to include multidimensional L1 tests that separately measure the various components of language proficiency, such as vocabulary and grammar. Instead, global and efficient but also comprehensive measures of L1 proficiency had to be developed. Listening comprehension constitutes a complex process requiring the integration of phonological, syntactic, semantic, and pragmatic skills (Anderson, 1995; Flowerdew & Miller, 2005).

4 See Blossfeld, von Maurice, and Schneider (2011) for a description of the starting cohorts and project structure in general.

5 To include all students whose L1 is potentially Russian or Turkish, we initially defined the target population based on country of birth even though Russian or Turkish is not necessarily the L1 of all families from the Former Soviet Union or Turkey.

We therefore decided to assess this aspect as an unidimensional indicator of general language proficiency.

In order to limit the testing time and financial costs, moreover, the NEPS L1 tests were required to be applicable in group settings and to use a paper-and-pencil format. Additionally, in order to avoid costs associated with coding open-response items, all test items had to be in a multiple-choice format.

2.3 Focus on Listening Comprehension

Models of language proficiency and language testing often distinguish between four basic dimensions of language proficiency: listening, reading, speaking, and writing (Harris, 1969; Lado, 1961). Large-scale studies assessing language proficiency focus mainly on reading comprehension. However, children of immigrants typically acquire the L1 in the family context, and the L1 is rarely used in school instruction. A large proportion of immigrant students are therefore unable to read or write in this language. To ensure that students at all levels of L1 proficiency can participate in the assessment, we decided to test the domain of listening comprehension.

Including students who are unable to read and write in their L1 and who may have limited L1 skills overall is important because analyses on most research questions require that a broad spectrum of L1 proficiency be represented in the data. For instance, to identify factors predicting the maintenance or loss of L1, it is crucial to include lower proficiency levels in L1. Similarly, in estimating effects of L1 proficiency on L2 or L3, it may be informative to differentiate between effects of lower and higher levels of L1 proficiency (e. g., Dollmann & Kristen, 2010; Edele & Stanat, 2015).

2.4 Coverage of Proficiency Distribution

In order to ensure that the L1 tests developed for the NEPS would cover a broad range of proficiency levels, we developed listening comprehension texts with varying linguistic difficulty. Based on data from a preliminary study and a larger pilot study, we also ensured that the difficulty of the items varied substantially (for details, see Edele, Schotte, Hecht, & Stanat, 2012; Edele, Schotte, & Stanat, 2015). Due to the tests' limited number of items, however, the instruments differentiate most accurately at intermediate proficiency levels, while their capacity to precisely measure very high or low proficiency levels is somewhat restricted.

2.5 Independence of Test Performance from Previous Knowledge

The L1 tests aim at assessing students' ability to understand spoken language in L1. To ensure that the test does, in fact, measure language proficiency rather than prior knowledge, we used either texts that cover topics that should be familiar to all students alike, such as everyday situations in school and family contexts, or topics that are likely to be equally unfamiliar to all participants, such as events in a fictitious narration written specifically for the test.

One aspect that needs to be taken into account in testing immigrant students is the possible impact of cultural knowledge on their performance. There is evidence that text processing is contingent upon the fit between cultural knowledge and the content of the text. Steffensen, Joag-Dev, and Anderson (1979), for instance, found in their study that participants recalled more information and needed less time when the text they read described content consistent with their cultural knowledge rather than content typical of another culture. Thus, it can be assumed that a text is easier to process when it is in line with test takers' culturally shaped prior knowledge.

This knowledge is likely to vary considerably in the target population for testing L1 in the NEPS, depending on such factors as students' immigrant generation status and acculturation strategies (Berry, Phinney, Sam, & Vedder, 2006; Edele, Stanat, Radmann, & Segeritz, 2013). In order to avoid biases associated with students' culturally shaped knowledge, the stimuli in the L1 tests were chosen such that they should be equally familiar or equally novel to students with different cultural backgrounds.

2.6 Comparability of the Russian and Turkish Test

Some research questions only focus on one of the two first languages assessed in the NEPS and the respective immigrant group. For other research questions, however, it may be important or interesting to determine whether the expected pattern of findings generalizes across both L1s and immigrant groups. To ensure that the relationships between L1 proficiency and other constructs can be compared across the two groups, the tests in Russian and Turkish need to capture the same construct. We therefore developed tests in Russian and Turkish that are equivalent with regard to the content of the texts, the questions, and the response options. In addition, we tried to keep linguistic features comparable that are likely to affect the difficulty of the texts.

Even the most careful translation process, however, does not necessarily ensure measurement equivalence. Measurement equivalence can be tested with multigroup confirmatory factor analyses (MGCFA). Different forms of equivalence can be distinguished (see Schroeders & Wilhelm, 2011 for a detailed description of testing invariance with categorical data). If configural equivalence is given, it can be assumed that an instrument assesses the same construct in two or more groups or—in our case—that the instruments assess the same construct in the respective target population.

Strong invariance is necessary to directly compare the latent means of a test and its correlations with external criteria. In strictly invariant tests, even raw test scores are comparable. Due to the pronounced linguistic differences between the Russian and the Turkish languages on the one hand and the limited time and financial resources for test development on the other hand, it would have been unrealistic to expect strong or even strict invariance of the two tests. We did, however, strive for configural invariance.

3 Test Construction

The L1 tests developed for Grade 9 consist of short texts that are orally presented to students from a CD. The students are subsequently asked to answer questions about what they have heard. All items have a multiple-choice format. To assess listening comprehension broadly, the tests include dialogues, expository texts, and narrative texts. The text types differ in their linguistic features. While the dialogues have features typical of informal oral language, such as repetitions, redundancies, ellipses, breaks, and fragmented language, the expository and narrative texts involve linguistic features typical of written language, such as more explicit vocabulary, correct grammar, and a lack of redundancy or repetition (Grotjahn, 2005; Shohamy & Inbar, 1991). A preliminary version of the tests was included in a pilot study (see Sections 4.2 and 4.4 for more details on the pilot study) for the purpose of item selection.

The final tests for Grade 9 consist of seven text units, namely two dialogues, two narrative texts, and three expository texts. The audio-recorded texts and questions are presented to the students once before they answer the questions about what they have heard. Every text unit is followed by three to six multiple-choice questions, resulting in a total of 31 test items⁶, each with four or five response options. The administration of the tests takes 30 minutes (Russian version) and 32 minutes (Turkish version). For further details on the test construction, see Edele, Schotte, et al. (2015).

4 Validity of the L1 Tests

4.1 Validation Strategy

To investigate the construct validity of the L1 tests, we examined their dimensionality and correlated students' scores with other indicators of L1 proficiency as well as with a nomological net of relevant constructs (Cronbach & Meehl, 1955).

6 Of the 32 items originally included in the final test version, one item was later excluded due to a poor model fit in the main study (see Edele et al., 2012 for further information on the scaling of the tests).

As a first step, we tested whether our L1 tests possess the expected unidimensional structure. To establish the convergent validity (Campbell & Fiske, 1959) of our L1 measures, we then correlated the test scores with another indicator of proficiency in Russian or Turkish. As both instruments are objective tests of language proficiency, we expected the correlations to be substantial. However, as the instruments examine different aspects of language proficiency (see section 4.3), we did not expect a particularly close association between them.

As another indicator of convergent validity, we correlated our L1 test scores with subjective measures of L1 proficiency. Even though subjective assessments, and particularly self-assessments, are susceptible to biases (Edele, Seuring, Kristen, & Stanat, 2015), we expected at least a moderately high correlation.

General cognitive abilities served as criteria in our analyses of discriminant validity. Reasoning ability, which constitutes a key aspect and prototypical indicator of general cognitive abilities, is assumed to influence the efficiency of language acquisition and should therefore relate positively with L1 proficiency (e. g., Esser, 2006; Spolsky, 1989). In addition, text comprehension requires deductive reasoning. However, listening comprehension depends on a multitude of other factors and should be clearly distinguishable from reasoning. We therefore expected a significant yet moderate relationship between listening comprehension in L1 and reasoning ability. Speed of perception, by contrast, should be unrelated to performance in our L1 tests as they do not contain a speed component.

To extend the nomological net for the construct validation of the L1 tests, we draw on models of language acquisition from different disciplines. These models suggest a number of conditions that should foster the acquisition and maintenance of L1 proficiency in immigrants and their children, such as exposure and motivation for language acquisition and improvement (e. g., Chiswick & Miller, 2001; Esser, 2006; Spolsky, 1989).

Immigrant students are exposed to their L1 in different contexts. The most important environment for the acquisition and improvement of L1 skills is typically the family. In addition, children and adolescents from immigrant families may have the opportunity to improve their L1 in interactions with co-ethnic peers. The use of media in L1 can also present an important opportunity for L1 acquisition. Therefore, students' exposure to L1 in the family, in the peer group, and in the media should be positively related to their L1 proficiency (e. g., Duursma et al., 2007; Scheele et al., 2010).

The immigrant generation status can also be assumed to affect exposure to L1 (e. g., Chiswick & Miller, 2001). Generally, a decrease in L1 use and proficiency and an increase in L2 use and proficiency can be observed in immigrants over time (Rumbaut, 2004; Stanat, Rauch, & Segeritz, 2010; Strobel & Kristen, 2015). First-generation immigrant students thus typically have more opportunities to acquire the L1 in their family context than do students who were born in the country of residence. Moreover, first-generation immigrants may have acquired the L1 in their heritage country. Ac-

cordingly, we expected first-generation immigrants to be more proficient in L1 than successive immigrant generations.

Within the first immigrant generation, students who immigrated at an older age were extensively exposed to L1 while they lived in the heritage country—and the quality of the language input can be assumed to be relatively high. Therefore, we expected age at migration to correlate positively with L1 test scores.

In addition to providing learning opportunities, using L1 with family and peers and in media can foster immigrant students' motivation to further improve their L1 skills. As a strong identification with the heritage culture should boost the motivation to improve in L1, it should be positively related to L1 proficiency.

4.2 Study Design

We draw on data from two studies to examine the validity of our tests. This allows us, on the one hand, to cross-check the findings, as most validation criteria were measured in both studies. On the other hand, the two studies complement each other as some validation criteria were assessed in only one of the studies. In addition, the second study includes a larger sample (for further details on sampling within the NEPS, see Aßmann et al., 2011).

The first investigation (Study 1) is a pilot study that was carried out within the NEPS to select items for the final L1 tests. On the test day, students filled out a questionnaire and subsequently completed a preliminary version of an L1 test in their respective first language. The analyses presented in this paper are based on the texts and items included in the final test version that was subsequently administered in the second study (see below). The L1 tests in both studies are largely identical, with the only exception being that a few false response options (distractors) were excluded from some items in the final tests. This should not substantially affect the patterns of findings relevant for validating the tests. A few months after the assessment, a sub-sample of the students completed another test measuring proficiency in Russian or Turkish.

The second study (Study 2), on which we draw in the following analyses, is the main study of the NEPS for the 9th-Grade sample of Starting Cohort 4 (school and vocational training—education pathways of students in 9th Grade and higher, doi:/10.5157/NEPS:SC4:4.0.0;⁷ see Frahm et al., 2011; von Maurice, Sixt, & Blossfeld, 2011, for further information on this starting cohort). The L1 tests were administered on a separate test day. To ensure that the students had at least a very basic proficiency level in Russian or Turkish, they completed a screening test in the respective lan-

7 From 2008 to 2013, NEPS data were collected as part of the Framework Programme for the Promotion of Empirical Educational Research funded by the German Federal Ministry of Education and Research (BMBF). As of 2014, the NEPS survey is carried out by the Leibniz Institute for Educational Trajectories (LIfBi) at the University of Bamberg in cooperation with a nationwide network.

guage prior to the L1 test. These screening tests consist of recordings of eight simple spoken sentences, such as “the dog walks.” Participants were asked to relate each sentence to the corresponding picture among five options. Test administrators instantly scored the screening tests using a template. Students who answered a minimum of three items correctly were eligible for participation in the L1 tests.

4.3 Assessment of Validation Criteria

For the validation analyses, we draw on a number of variables measured with student questionnaires, student competence tests, and computer-assisted telephone interviews (CATIs) with students' parents.

Objective measure of L1 proficiency

Proficiency in Russian and Turkish was tested in individual testing sessions with the *Bilingual Verbal Ability Test, BVAT* (Muñoz-Sandoval, Cummins, Alvarado, & Ruef, 1998). The goal of the BVAT is to capture bilingual participants' overall language ability—and to thereby avoid underestimating their linguistic capacity—by taking into account language proficiency in L1 and L2. More specifically, the BVAT begins by examining participants' proficiency in L2, typically English. If they fail to solve an item in L2, it is presented in the respective L1. The test results consequently reflect the additive language ability in L1 and L2. The BVAT is available for 17 languages besides English, among them Russian and Turkish. The target population of the test ranges from 5 years to old age. The instrument examines productive language proficiency and includes the four subscales *picture vocabulary*, *synonyms*, *antonyms*, and *verbal analogies*. The test is adaptive by specifying starting items according to participants' age as well as termination criteria after a series of eight (*picture vocabulary*, *verbal analogy*) or six (*synonyms*, *antonyms*) unsolved items.

As we are specifically interested in students' L1 proficiency, we presented students only with items in Russian or Turkish. In addition, we refrained from adaptive testing and instead presented the same item set to all participants in order to obtain comparable test scores. Due to time constraints, we only employed the subscales *picture vocabulary* and *synonyms*. The *picture vocabulary* scale requests participants to name drawings of objects or activities, while the *synonyms* scale requires an active production of synonyms for verbally presented words. We excluded the first eight items of the *picture vocabulary* subscale as we considered them too easy for the targeted age group. In total, we administered 51 items from the subscale *picture vocabulary* and 20 items from the subscale *synonyms* of the BVAT.

Trained test administrators, who were native speakers of Russian or Turkish, coded students' responses during the test session. The test sessions were recorded, and 61 % (Russian sample) or 66 % (Turkish sample) of answers were additionally coded by two other native speakers. Inter-rater reliability was very high, with Yules

$Y = .88$ in the Russian sample and $Y = .87$ in the Turkish sample, confirming that the test administrators coded answers adequately.

To deviate as little as possible from the original BVAT, we kept items even if their discriminatory power was lower ($.30 > d > 0$) than would usually be considered acceptable for psychometric tests (Bortz & Döring, 2002). Only items with a discrimination ≤ 0 in the *picture vocabulary* scale were excluded from further analyses (five items in the Russian test, eight items in the Turkish test), leaving 66 items in the Russian test and 63 items in the Turkish test in total. The sum of the correctly answered items from both scales served as the validation criterion for our L1 proficiency tests.

Despite its limitations (see Edele, Schotte et al., 2015), we decided to use the BVAT as a validation criterion for our L1 tests because, to our knowledge, no other instruments suitable for testing oral language proficiency in Russian and Turkish exist for the target population of our study.

General cognitive abilities

The NEPS examines *perceptual speed* and *nonverbal reasoning* as indicators of general cognitive abilities (Lang, Kamin, Rohr, Stünkel, & Williger, 2014). The *perceptual speed* test requires students to allocate numbers to symbols according to a provided key. The *reasoning* test consists of matrices similar to those of the RAVEN test (Raven, 1977). In our analyses, we use the sum of correct answers for each scale as ability estimates.

Subjective indicators of L1 proficiency

The validation analyses also draw on several subjective indicators of students' L1 proficiency. First, the student questionnaires measured self-reported L1 proficiency of students with a first language other than German ("How good is your command of the other language?"⁸) for the dimensions of listening, speaking, reading, and writing. The 5-point response scale was "very good—rather good—rather poor—very poor—not at all." For the analyses, the arithmetic mean of students' ratings across the four dimensions was computed, resulting in the scale *self-estimated global L1 proficiency*, which could vary from 0 (not at all) to 4 (very good).

Students were additionally asked to estimate the number of items they had answered correctly in the L1 test. The *self-estimated number of items solved* is contingent upon the total number of items in the test and could thus range from 0–32⁹ (see Lockl, 2013 for further information on the assessment of procedural metacognition in the NEPS). We interpret this scale as another subjective indicator of L1 proficiency.

8 Before students reached this item, the questionnaire had defined "the other language." Specifically, students were asked to indicate the language other than German they had learned as a child in their family. Afterwards, they were informed that the questionnaire would subsequently refer to this language as "the other language."

9 Students estimated the number of solved items on the basis of the 32 items originally included in the test of which one was eliminated later on due to poor model fit.

Parents' estimates of their children's L1 proficiency on the dimensions of speaking and writing served as a third subjective indicator of students' L1 proficiency. The rating scale was the same as for the students' self-reported L1 proficiency, and the two dimensions were averaged.

Patterns of language use

Another set of items in the student questionnaires measured the patterns of language use in the family (with mother, with father, with siblings) and with peers (with best friend and with classmates) for students with a first language other than German. An example of these questions is: "What language do you speak with your mother?" The 4-point response scale was "only German—mostly German, sometimes the other language—mostly the other language, sometimes German—only the other language." Students' ratings of language use with the mother and with the father were averaged. Similarly, the ratings of language use with the best friend and with classmates were combined into an indicator of the language used with peers.

The student questionnaires further assessed the language of media consumption with seven items capturing the language in which, among other things, students read books, watch television, or surf the web. The same 4-point scale as that for language use with family and peers was employed. We averaged the seven items to a single indicator of language in media use.

Immigrant generation status and age at immigration

We defined the target persons' immigrant generation status based on the country of birth of the students themselves, of their parents, and of their grandparents. We classified students who were born abroad along with their parents as *first generation*; students who were born in Germany but whose parents were both born abroad as *second generation*; and students who were born in Germany, whose parents were born in Germany, and whose grandparents (at least two) were born abroad as *third generation*. We further defined students with one parent born abroad and one parent born in Germany as *one parent born abroad*. Students' immigrant generation status was only classified when all data necessary for its univocal identification were available. The student questionnaires also assessed the age at which students who were born abroad immigrated to Germany.

Identification with heritage culture

Four items captured students' emotional identification with the heritage culture of their families. One item, for instance, states: "I feel closely attached to this culture."¹⁰

10 Before students reached the item, the questionnaire requested students to indicate the country other than Germany from which their family originates. Afterwards, it explained that subsequent questions would refer to the culture of this country as students' "heritage culture." An example for this was presented.

The 4-point response scale was “does not apply—partially applies—mostly applies—fully applies.”

While most validation criteria (particularly the information from the student questionnaires) were measured in both studies, the BVAT was only administered in Study 1, whereas parents’ estimates of students’ L1 proficiency, the self-estimated number of items solved in the L1 test, as well as general cognitive abilities were only included in Study 2.

4.4 Sample

Both studies tested L1 proficiency of immigrant students (first, second, and third generation) whose families immigrated to Germany from the area of the Former Soviet Union (e. g., Russia, Kazakhstan) or Turkey.¹¹ Table 1 presents descriptive information on the samples of Studies 1 and 2.

Study 1 is based on data from schools located in four federal states (Bavaria, Berlin, Hamburg, North Rhine-Westphalia) attended by high percentages of students speaking Turkish and/or Russian. The Russian L1 test was conducted in 17 schools, and the Turkish L1 test in 15 schools.

Study 1

The Russian sample consists of 224 participants (53 % female). On average, students were 16 years old at the time of data assessment. Of these students, 37 % were enrolled in a *Hauptschule* (lowest school track of German secondary education), 31 % in a *Gesamtschule* (comprehensive school), 17 % in a *Schule mit mehreren Bildungsgängen* (school with several educational tracks), and 15 % in a *Gymnasium* (highest track leading to university-entrance degree). More than two-thirds (71 %) were classified as first-generation immigrant students, while only 20 % were second generation.

The Turkish sample consists of 310 participants (50 % female). Participants’ mean age was 15.7 years. The majority of students either attended the *Hauptschule* (28 %) or the *Gesamtschule* (31 %), while a somewhat lower proportion of the students attended a *Schule mit mehreren Bildungsgängen* (20 %) or a *Gymnasium* (22 %). The majority of students in the Turkish sample (66 %) were second generation, while only 7 % were first-generation immigrants.

A subsample of 79 participants in the Russian sample and of 101 participants in the Turkish sample completed the BVAT.

11 For the sake of brevity, the former group is subsequently referred to as “the Russian sample,” and the latter as “the Turkish sample.” These terms do not allude to citizenship or the like, but rather to the L1 that was tested in the sample.

Table 1 Gender, Age, Attended School Track, and Immigrant Generation Status of the Russian and Turkish Samples in Studies 1 and 2

	Study 1				Study 2				
	Russian		Turkish		Russian		Turkish		
	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	
Total	224		310		502		662		
Gender									
Male	106	47.3	155	50.0	248	49.4	342	51.7	
Female	118	52.7	155	50.0	254	50.6	320	48.3	
School Track									
Hauptschule	82	36.6	87	28.1	206	41.0	330	49.8	
Realschule	–	–	–	–	124	24.7	130	19.6	
Gymnasium	34	15.2	67	21.6	98	19.5	94	14.2	
SMB	38	17.0	61	19.7	27	5.4	11	1.7	
Gesamtschule	70	31.2	95	30.6	47	9.4	97	14.7	
Immigrant Generation									
1st Generation	159	71.0	23	7.4	234	46.6	62	9.4	
2nd Generation	44	19.6	205	66.1	206	41.0	430	64.9	
3rd Generation	–	–	3	1.0	–	–	19	2.9	
One Parent Born Abroad	6	2.7	52	16.8	28	5.6	132	19.9	
Not Determinable	15	6.7	27	8.7	34	6.8	19	2.9	

Note. SMB = Schule mit mehreren Bildungsgängen (school with several educational tracks).

Study 2

Study 2 draws on data from schools located in all federal states of Germany. The Russian L1 tests were administered in 257 schools, and the Turkish L1 tests in 237 schools.

The Russian sample includes 502 students in total (51 % female) (for details on missing values, see Edele, Schotte, et al., 2015). On average, the students were 15.8 years old at the time of testing. The largest proportion of students attended the Hauptschule (41 %), followed by students attending the Realschule (25 %) and the Gymnasium (20 %). Almost equal proportions of students belonged to the first immigrant generation (47 %) and to the second generation (41 %). On average, the students born abroad were 5.3 years old when they came to Germany.

The Turkish sample consists of 662 students (48 % female). On average, students in this sample were 15.7 years old. Half of them attended the Hauptschule (50 %), 20 % the Realschule, and 14 % the Gymnasium. In this sample, the majority of students (65 %) were second-generation immigrants; only 9 % were first generation.

4.5 Results

We analyzed the data from the Russian and Turkish samples separately as these groups differ with regard to several important characteristics, such as the proportion of first and second immigrant generation students and the attended school types.

Scaling, item difficulty, and reliability

IRT scaling and item analyses show that the L1 tests fit the Rasch model well. In addition, the item difficulty and the target populations' L1 proficiency generally matches well. In Study 1, the mean item difficulty was $b = -0.57$ for the Russian test and $b = -0.17$ for the Turkish test. In Study 2, the mean item difficulties were $b = -0.12$ (Russian test) and $b = -0.23$ (Turkish test). The range of item difficulty is, however, somewhat limited. It ranges from -2.01 to 0.61 (Russian test) and from -2.61 to 1.23 (Turkish test) in Study 1 and from -1.48 to 1.41 (Russian test) and -1.78 to 1.30 (Turkish test) in Study 2. The tests proved to be highly reliable. In Study 1, the WLE-reliability was .86 for the Russian test and .79 for the Turkish test. In Study 2, the reliability coefficients were .85 for the Russian test and .83 for the Turkish test (see Edele, Schotte, et al., 2015, and Edele et al., 2012 for more details on scaling, item difficulty, and reliability).

Measurement equivalence

To test for measurement equivalence, we conducted a MGCFA on the L1 tests in Study 2 (see Table 2). The results show that the fit indices of the model assuming configural invariance are acceptable (see Yu, 2002). The more restrictive models assum-

Table 2 Tests of Measurement Equivalence of the Russian and the Turkish L1 Tests

	χ^2/df	p	CFI	TLI	RMSEA	$\Delta\chi^2/\Delta df$	p
Configural Equivalence	1315.8/868	.00	.95	.95	.03		
Strong Equivalence	1714.0/898	.00	.91	.91	.04	367.72/30	.00
Strict Equivalence	1775.3/929	.00	.91	.91	.04	108.39/31	.00

Note. In computing these models with MPlus 6.1 (Muthén & Muthén, 2009), we employed a robust weighted least squares approach (estimator: WLSMV) and estimated model parameters based on Theta parameterization; CFI = comparative fit index; TLI = Tucker-Lewis Index; RMSEA = root mean square error of approximation.

ing strong and strict invariance, however, do not hold since the model fit indices are not satisfactory and the test of change in model fit is significant.

These findings confirm configural equivalence of the Russian and the Turkish tests, implying that they measure the same construct. Because more restrictive models of equivalence are not supported, however, neither the ability scores from the Turkish and the Russian tests nor their correlation coefficients with other variables are directly comparable.

Dimensional structure

To examine whether our L1 tests exhibit the expected unidimensional structure, a 1-dimensional model was tested against an alternative, theoretically plausible 2-dimensional model. The 2-dimensional model assigns items on dialogues to the first dimension and items on expository as well as narrative texts to the second dimension. We computed these models with ConQuest 2.0 (Wu, Adams, Wilson, & Haldane, 2007), using Marginal Maximum Likelihood (MML) estimation with Gauss-Hermite quadrature.

In Study 1, the two dimensions correlate very highly, with .97 in the Russian sample and .94 in the Turkish sample. The 1-dimensional model fits the data better than the 2-dimensional model in both language groups, as is demonstrated by two indicators of model fit, namely Akaike's information criterion (AIC; Akaike, 1974) and the Bayesian information criterion (BIC; Schwarz, 1978) (see Table 3). The two dimensions are also highly correlated in Study 2, with .94 in the Russian sample and .98 in the Turkish sample. The model fit indices suggest that the 2-dimensional model fits negligibly better according to the AIC and slightly more poorly according to the BIC in the Russian sample. In the Turkish sample, the 2-dimensional model fits somewhat more poorly, as indicated by both indicators (see Table 3). The very high correlation between the two dimensions, which indicates their near identity, and the very simi-

Table 3 Results of the 1-dimensional and 2-dimensional Scaling: AIC and BIC Model Selection Criteria

	Study 1				Study 2			
	Russian Sample		Turkish Sample		Russian Sample		Turkish Sample	
	<i>1-dim</i>	<i>2-dim</i>	<i>1-dim</i>	<i>2-dim</i>	<i>1-dim</i>	<i>2-dim</i>	<i>1-dim</i>	<i>2-dim</i>
AIC	7720.9	7724.0	11466.2	11468.7	19201.5	19196.3	25767.9	25772.1
AIC-Diff	3.1		2.5		-5.1		4.2	
BIC	7830.1	7840.0	11585.8	11595.7	19340.7	19344.0	25916.2	25929.4
BIC-Diff	9.9		9.9		3.3		13.2	

Note. AIC = Akaike information criterion; BIC = Bayesian information criterion

lar model fit, which is slightly in favor of the 1-dimensional model, suggest that the construct measured with the L1 tests is unidimensional rather than 2-dimensional.

Criterion validity: Other indicators of L1 proficiency and general cognitive abilities

To examine the convergent validity of the L1 tests, in a first step, we correlated the L1 test scores with students' scores on the BVAT. In the Russian sample, the two objective indicators correlate quite highly; in the Turkish sample, the correlation is moderate (see Table 4).

In a second step, we correlated immigrant students' L1 test scores with a number of subjective indicators of L1 proficiency. As expected, students' self-estimated

Table 4 Pairwise Correlations Between Immigrant Students' L1 Test Scores (WLEs) and Validation Criteria

	L1 test score (WLE)			
	Study 1		Study 2	
	Russian	Turkish	Russian	Turkish
BVAT	.72*** (79)	.41*** (101)	–	–
Subjective Measures of L1 Proficiency				
Self-estimated Global L1 Proficiency	0.50*** (220)	0.28*** (302)	0.43*** (426)	0.26*** (572)
Self-Estimated Number of Solved Items	–	–	0.59*** (488)	0.54*** (636)
Parents' Estimates of L1 Proficiency	–	–	0.57*** (182)	0.40*** (232)
General Cognitive Abilities				
Perceptual Speed	–	–	0.11* (483)	0.03 (632)
Reasoning	–	–	0.16*** (482)	0.28*** (628)
Age at Immigration	0.49*** (140)	0.04 (14)	0.34*** (263)	0.30* (48)
Language Use				
With Parents	0.38*** (224)	0.09 (305)	0.31*** (437)	0.14*** (581)
With Siblings	0.27*** (205)	0.07 (293)	0.31*** (399)	0.17*** (545)
With Peers	0.34*** (223)	0.18** (304)	0.10* (437)	0.04 (572)
in Media Use	0.46*** (222)	0.23*** (304)	0.28*** (420)	0.23*** (559)
Identification With Heritage Culture	0.12 (212)	0.17** (290)	0.17*** (465)	0.15*** (566)

Note. Correlations are given as Pearson's r . Spearman's rank correlation coefficients, which we additionally computed because of the non-normal distribution of some validation criteria, yielded almost equal results. Number of cases (N) in parentheses.

* $p < .05$, ** $p < .01$, *** $p < .001$

global L1 proficiency is positively related to their L1 test score. Students' self-estimated number of items solved in the test is also substantially related to their results in the L1 tests, strong correlations emerged in both language groups. Similarly, parents' estimates of their children's L1 proficiency are also strongly (Russian sample) or moderately (Turkish sample) associated with the L1 test scores.

Analyses exploring the discriminant validity of the L1 tests indicate that, as expected, the L1 test scores correlate moderately with reasoning but only weakly and inconsistently with perceptual speed.

Criterion validity: L1 exposure and motivation

Analyses of the L1 tests' validity with indicators of exposure to L1 and motivation for L1 acquisition as criteria generally also show the expected pattern of results. In the Russian group, students' L1 test scores in both studies correlate substantially with their age at immigration as well as with the language they use with parents, siblings, peers, and in media consumption. In the Turkish group, the age at immigration is also positively related to L1 test scores in Study 2 but not in Study 1. However, the coefficient in Study 1 is based on a very small number of students and may therefore not be reliable. The language of media use also shows the expected correlation with the L1 test scores in the Turkish group. Overall, language use in the family and with peers also shows the expected pattern, yet the correlation coefficients are somewhat smaller and less consistent in the Turkish group than in the Russian sample.

In a next step, we examined whether the duration of the family's residence in Germany is associated with the L1 test scores. Because the number of third-generation students is very small, the analyses focus on the first and second generations. As expected, the test scores in Russian are higher for first immigrant generation students than for second-generation students (see Table 5). In the Turkish sample, however, the first generation does not show higher L1 test scores than the second generation.

Table 5 L1 Test Scores (WLEs) by Immigrant Generation Status

	Study 1				Study 2			
	First Generation		Second Generation		First Generation		Second Generation	
	<i>M</i>	<i>SE</i>	<i>M</i>	<i>SE</i>	<i>M</i>	<i>SE</i>	<i>M</i>	<i>SE</i>
Mean L1 Test Score Russian	.31	.10	-.81***	.17	.39	.08	-.34***	.07
Mean L1 Test Score Turkish	.27	.23	.09	.06	-.19	.13	.07	.05

Note. Differences between the first and second immigrant generation groups were tested separately within the Russian and Turkish samples with Mann-Whitney U-tests.

*** Significant difference between first and second immigrant generation ($p < .001$)

In addition, we expected that students' identification with the heritage culture should motivate them for L1 acquisition and should consequently correlate with L1 test scores. The expected pattern emerges in both groups, although the coefficients are rather small (see Table 4).

5 Discussion

The analyses presented in this article confirm that the L1 listening comprehension tests developed for the NEPS are valid measures of 9th-Grade students' proficiency in Turkish and Russian. The L1 tests show convergent validity as evidenced by correlations with another L1 test (BVAT) as well as with a number of subjective estimations of L1 proficiency. As expected, the correlations of our tests with the subjective indicators are somewhat weaker than those with the BVAT, once again suggesting that the subjective proficiency estimates are biased (see also Edele, Seuring, et al., 2015; Finnie & Meng, 2005).

The correlations of our tests with the BVAT are more substantial. However, given that both instruments are objective tests of proficiency in Russian or Turkish, even higher correlations could have been expected, particularly in the Turkish group, for which we only observed a moderate correlation. However, unlike our L1 tests, the BVAT assesses productive language proficiency and examines a linguistic subcomponent (vocabulary). In addition, the BVAT suffers from a number of conceptual and psychometric limitations (for further details, see Edele, Schotte, et al., 2015). These factors most likely limited the correlations with our tests.

Further validation analyses using indicators of exposure to L1 and motivation to acquire L1 as criteria generally also yielded the expected pattern of results, thereby corroborating the tests' validity. In the Turkish group, however, some criterion variables did not show the expected correlations with the L1 test scores. In particular, the first generation did not score higher on the Turkish test than the second generation. This could indicate that the construct validity of the Turkish test is limited. However, the Turkish test was correlated with various other validation criteria in the predicted way, particularly with other subjective as well as objective indicators of Turkish proficiency. The lack of significant relationships with some of the criteria may therefore instead indicate that some of our theoretical assumptions do not fully apply to the Turkish group. Indeed, we found some evidence that second-generation students of Turkish origin may not have significantly fewer opportunities for L1 acquisition than first-generation students (for further details, see Edele, Schotte, et al., 2015), suggesting that this group was more reluctant to give up the use of Turkish than assumed by the models of language acquisition (e. g., Chiswick & Miller, 2001).

In general, the correlations of our L1 tests with the criterion variables were higher for the Russian than for the Turkish test. As strong measurement equivalence could

not be confirmed for the two tests, however, these correlations are not directly comparable.

Our L1 tests assess listening comprehension, which allows them to be administered to students at practically all levels of L1 proficiency, including those with low skills in reading and writing. Moreover, the tests assess students' L1 proficiency efficiently and comprehensively. To further extend the analytical potential of the NEPS, it would be interesting to develop tests of students' reading and writing proficiency in Russian and Turkish, which may be particularly relevant for certain outcomes like labor market success. Tests assessing linguistic subcomponents, such as grammar and vocabulary, as well as tests of other L1s frequently spoken by immigrants in Germany, such as Polish, would further complement the database.

Taken together, our tests present valid measures of an important aspect of L1 proficiency. The instruments will facilitate research on the effects of L1 proficiency on immigrant students' educational development and other outcome variables in a methodologically more sound way than was possible in the past.

References

- Adesope, O. O., Lavin, T., Thompson, T., & Ungerleider, C. (2010). A systematic review and meta-analysis of the cognitive correlates of bilingualism. *Review of Educational Research, 80*, 207-249. doi: 10.3102/0034654310368803
- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*(6), 716-722.
- Anderson, J. R. (1995). *Cognitive psychology and its implications*. New York: Freeman.
- Aßmann, C., Steinhauer, H. W., Kiesl, H., Koch, S., Schönberger, B., Müller-Kuller, A., ... Blossfeld, H.-P. (2011). Sampling designs of the National Educational Panel Study: Challenges and solutions. In H.-P. Blossfeld, H.-G. Roßbach, & J. von Maurice (Eds.), *Zeitschrift für Erziehungswissenschaft, 14. Education as a lifelong process: The German National Educational Panel Study (NEPS)* (pp. 1-65). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Berry, J. W., Phinney, J. S., Sam, D. L., & Vedder, P. (2006). Immigrant youth: Acculturation, identity, and adaptation. *Applied Psychology: An international review, 55*, 303-332. doi: 10.1111/j.1464-0597.2006.00256.x
- Bialystok, E. (2007). Cognitive effects of bilingualism: How linguistic experience leads to cognitive change. *The International Journal of Bilingual Education and Bilingualism, 10*(3), 210-223.
- Blossfeld, H.-P., Roßbach, H.-G., & von Maurice, J. (Eds.). (2011). Education as a lifelong process: The German National Educational Panel Study (NEPS) [Special issue]. *Zeitschrift für Erziehungswissenschaft, 14*. Wiesbaden: VS Verlag für Sozialwissenschaften.

- Blossfeld, H.-P., von Maurice, J., & Schneider, T. (2011). The National Educational Panel Study: Need, main features, and research potential. In H.-P. Blossfeld, H.-G. Roßbach, & J. von Maurice (Eds.), *Zeitschrift für Erziehungswissenschaft, 14. Education as a life-long process: The German National Educational Panel Study (NEPS)* (pp. 5–17). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Bortz, J., & Döring, N. (2002). *Forschungsmethoden und Evaluation: Für Human- und Sozialwissenschaftler*. Heidelberg: Springer.
- Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the Multitrait-Multimethod-Matrix. *Psychological Bulletin, 56*(2), 81–105.
- Chiswick, B. R., & Miller, P. W. (2001). A model of destination language acquisition: Application to male immigrants in Canada. *Journal of Political Economy, 38*(3), 391–409.
- Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281–302.
- Cummins, J. (2000). *Language, power and pedagogy: Bilingual children in the crossfire* (Books Opinion Papers No. ED446442). Clevedon: Multilingual Matters Ltd., UTP.
- Dollmann, J., & Kristen, C. (2010). Herkunftssprache als Ressource für den Schulerfolg? – Das Beispiel türkischer Grundschul Kinder. In C. Allemann-Ghionda, P. Stanat, K. Göbel, & C. Röhner (Eds.), *Zeitschrift für Pädagogik, Beiheft, 55*, 123–146.
- Duursma, E., Romero-Contreras, S., Szuber, A., Proctor, P., Snow, C., August, D., & Calderón, M. (2007). The role of home literacy and language environment on bilinguals' English and Spanish vocabulary development. *Applied Psycholinguistics, 28*, 171–190. doi: 10.1017/S0142716406070093
- Edele, A., Schotte, K., Hecht, M., & Stanat, P. (2012). *Listening comprehension tests of immigrant students' first languages (L1) Russian and Turkish in grade 9: Scaling procedure and results*. (NEPS Working Paper No. 13). Bamberg: University of Bamberg, National Educational Panel Study.
- Edele, A., Schotte, K., & Stanat P. (2015). *Listening comprehension tests of immigrant students' first languages (L1) Russian and Turkish in grade 9: Extended report of test construction and validation*. (NEPS Working Paper No. 57). Bamberg: University of Bamberg, National Educational Panel Study.
- Edele, A., Seuring, J., Kristen, C., & Stanat, P. (2015). Why bother with testing? The validity of immigrants' self-assessed language proficiency, *Social Science Research, 52*, 99–123. doi:10.1016/j.ssresearch.2014.12.017
- Edele, A., & Stanat, P. (2015). The role of first-language listening comprehension in second-language reading comprehension. *Journal of Educational Psychology*. Advance Online Publication. doi:10.1037/edu0000060
- Edele, A., Stanat, P., Radmann, S., & Segeritz, M. (2013). Kulturelle Identität und Lesekompetenz von Jugendlichen aus eingewanderten Familien. In N. Jude, & E. Klieme (Eds.), *Zeitschrift für Pädagogik, Beiheft, 59*, 84–110.
- Esser, H. (2006). *Sprache und Integration: Die sozialen Bedingungen und Folgen des Spracherwerbs von Migranten*. Frankfurt am Main: Campus Verlag.

- Finnie, R., & Meng, R. (2005). Literacy and labour market income: self-assessment versus test score measures. *Applied Economics* 37(17), 1935–1951.
- Flowerdew, J., & Miller, L. (2005). *Second language listening: Theory and practice*. Cambridge: Cambridge University Press.
- Frahm, S., Goy, M., Kowalski, K., Sixt, M., Strietholt, R., Blatt, I., ... Kandera, M. (2011). Transition and development from lower secondary to upper secondary school. In H.-P. Blossfeld, H.-G. Roßbach, & J. von Maurice (Eds.), *Zeitschrift für Erziehungswissenschaft, 14. Education as a lifelong process: The German National Educational Panel Study (NEPS)* (pp. 217–232). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Grotjahn, R. (2005). Testen und Bewerten des Hörverstehens. In M. Ó. Dúill, R. Zahn, & K. D. C. Höppner (Eds.), *Zusammenarbeiten: Eine Festschrift für Bernd Voss* (pp. 115–144). Bochum: AKS-Verlag.
- Harris, D. P. (1969). *Testing English as a second language*. New York: McGraw-Hill.
- Hesse, H.-G., Göbel, K., & Hartig, J. (2008). Sprachliche Kompetenzen von mehrsprachigen Jugendlichen und Jugendlichen nicht-deutscher Erstsprache. In DESI-Konsortium (Ed.), *Unterricht und Kompetenzerwerb in Deutsch und Englisch: Ergebnisse der DESI-Studie* (pp. 208–201). Weinheim: Beltz.
- Lado, R. (1961). *Language testing: The construction and the use of foreign language tests*. London: Longman.
- Lang, F. R., Kamin, S., Rohr, M., Stünkel, C., & Williger, B. (2014). *Erfassung der fluiden kognitiven Leistungsfähigkeit über die Lebensspanne im Rahmen des Nationalen Bildungspanels: Abschlussbericht zu einer NEPS-Ergänzungsstudie*. (NEPS Working Paper No. 43). Bamberg: University of Bamberg, National Educational Panel Study.
- Lockl, K. (2013). *Assessment of procedural metacognition: Scientific Use File 2013*. Bamberg: University of Bamberg, National Educational Panel Study.
- Mouw, T., & Xie, Y. (1999). Bilingualism and the academic achievement of first- and second-generation Asian Americans: Accommodation with or without assimilation? *American Sociological Review*, 64(2), 232–252.
- Muñoz-Sandoval, A. F., Cummins, J., Alvarado, C. G., & Rief, M. L. (1998). *Bilingual Verbal Ability Tests, Comprehensive Manual*. Chicago: Riverside Publishing.
- Muthén, L. K., & Muthén, B. O. (2009). Mplus (Version 6.1) [Computer software]. Los Angeles, CA: Muthén & Muthén.
- Portes, A., & Rumbaut, R. G. (2012). *Children of Immigrants Longitudinal Study (CILS), 1991–2006 (ICPSR20520-v2)* [Codebook]. Retrieved from <http://doi.org/10.3886/ICPSR20520.v2>
- Rauch, D., Jurecka, A., & Hesse, H.-G. (2010). Für den Drittspracherwerb zählt auch die Lesekompetenz in der Herkunftssprache. In C. Allemann-Ghionda, P. Stanat, K. Göbel, & C. Röhner (Eds.), *Zeitschrift für Pädagogik, Beiheft, 55*, 78–100.
- Raven, J. C. (1977). *Standard Progressive Matrices: Sets A, B, C, D & E*. San Antonio, TX: Harcourt.

- Rumbaut, A. G. (2004). Ages, life stages, and generational cohorts: Decomposing the immigrant first and second generations in the United States. *International Migration Review*, 38(3), 1160–1205.
- Rumbaut, R. G., Massey, D. S., & Bean, F. D. (2006). Linguistic life expectancies: Immigrant language retention in southern California. *Population and Development Review*, 32(2), 447–460.
- Scheele, A. F., Leseman, P. M., & Mayo, A. Y. (2010). The home language environment of monolingual and bilingual children and their language proficiency. *Applied Psycholinguistics*, 31, 117–140. doi: 10.1017/S0142716409990191
- Schroeders, U., & Wilhelm, O. (2011). Equivalence of reading and listening comprehension across test media. *Educational and Psychological Measurement*, 71, 849–869. doi: 10.1177/0013164410391468
- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6(2), 461–464.
- Shohamy, E., & Inbar, O. (1991). Validation of listening comprehension tests: The effect of text and question type. *Language Testing*, 8(1), 23–40.
- Spolsky, B. (1989). *Conditions for second language learning: Introduction to a general theory*. Oxford: Oxford University Press.
- Stanat, P., Rauch, D., & Segeritz, M. (2010). Schülerinnen und Schüler mit Migrationshintergrund. In E. Klieme, C. Artelt, J. Hartig, N. Jude, O. Köller, M. Prenzel, ... P. Stanat (Eds.), *PISA 2009: Bilanz nach einem Jahrzehnt* (pp. 200–230). Münster: Waxmann.
- Steffensen, M., Joag-Dev, C., & Anderson, R. (1979). A cross-cultural perspective on reading comprehension. *Reading Research Quarterly*, 15(1), 10–29.
- Strobel, B., & Kristen, C. (2015). Erhalt der Herkunftssprache?—Muster des Sprachgebrauchs in Migrantenfamilien. *Zeitschrift für Erziehungswissenschaft*, 18, 125–142. doi: 10.1007/s11618-014-0607-1
- Verhoeven, L. T. (2007). Early bilingualism, language transfer and phonological awareness. *Applied Psycholinguistics*, 28, 425–429. doi: 10.1017/S0142716407070233
- von Maurice, J., Sixt, M., & Blossfeld, H.-P. (2011). *The German National Educational Panel Study: Surveying a cohort of 9th graders in Germany*. (NEPS Working Paper No. 3). Bamberg: University of Bamberg, National Educational Panel Study.
- Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. A. (2007). ACER ConQuest version 2.0: Generalised item response modelling software. Camberwell, Victoria.
- Yu, C.-Y. (2002). *Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes* (Doctoral dissertation, University of California). Retrieved from <http://www.statmodel.com/download/Yudissertation.pdf>

About the authors

A. Edele

Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany.
e-mail: aileen.edele@iqb.hu-berlin.de

K. Schotte

Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany.

P. Stanat

Institute for Educational Quality Improvement (IQB),
and Berlin Institute for Integration and Migration Research (BIM)
Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany.

Metacognitive Knowledge in Young Children: Development of a New Test Procedure for First Graders

Kathrin Lockl, Marion Händel, Kerstin Haberkorn and Sabine Weinert

Abstract

Declarative metacognition, that is, explicit knowledge about memory, comprehension, and learning processes, has been found within many studies to be related to memory development and strategy use (Schneider, 2015). Given its importance in the educational context, the National Educational Panel Study (NEPS) aims at assessing metacognitive knowledge over the life span. Considering metacognitive knowledge in a longitudinal perspective allows for investigating how metacognitive knowledge evolves and how its development is influenced by other competencies. The present chapter describes the development and evaluation of a new test instrument on metacognitive knowledge that is appropriate for first graders. Comparable with tests for other educational stages investigated in the NEPS, the newly constructed instrument consists of several scenarios that refer to different aspects of strategy knowledge. In the process of test development, an item pool of 20 scenarios was established and pretested in a pilot study with 195 first graders in a group setting. Various criteria were taken into account in the selection of items for the final instrument. The 10 scenarios in the final test covered a wide range of difficulties, and the test exhibited good reliability. The selected items showed good item-fit as well as appropriate item characteristic curves and item total correlations. Moreover, differential item functioning analyses have revealed that the final test was fair for the considered subgroups. In summary, the final instrument demonstrates good psychometric properties and thus serves as an important tool to describe metacognitive knowledge and to analyze its relevance within the educational context.

1 Introduction

The National Educational Panel Study (NEPS) aims at assessing competencies that are considered to be of particular importance for educational pathways and participation in society. In addition to longitudinal measurements of reading competence, listening comprehension, mathematical competence, and scientific literacy, so-called meta-competencies, such as the ability to handle information technologies (ICT) and metacognition, are part of the assessment program (cf. Weinert et al., 2011). Metacognition is considered a central component in the process of self-regulated learning (e. g., Boekaerts, 1997) and is defined as “any knowledge or cognitive activity that takes as its cognitive object, or that regulates, any aspect of any cognitive activity” (Flavell, Miller, & Miller, 1993, p. 150). This very broad conceptualization includes two components, namely declarative and procedural metacognition. While the declarative component refers to people’s knowledge about memory, comprehension, and learning processes, the procedural component comprises executive skills related to monitoring and self-regulation of one’s own cognitive activities (Nelson & Narens, 1990). In the NEPS, declarative and procedural aspects of metacognition are assessed over the life span (see Händel, Artelt, & Weinert, 2013, for an overview). In this chapter, we focus on declarative metacognition in younger children and describe the construction and evaluation of a test instrument that is to be administered in group settings for the assessment of metacognitive knowledge in first graders.

According to Flavell and Wellman (1977), declarative metacognition refers to conscious, explicit knowledge about person-, task-, and strategy variables. Thus, it includes knowledge about the strengths and weaknesses of one’s own memory and learning, knowledge about task characteristics as well as knowledge about ways and means of attaining cognitive learning and achievement goals. With respect to strategy variables, Paris, Lipson, and Wixson (1983) make a further distinction and differentiate between declarative, procedural, and conditional strategy knowledge. Declarative strategy knowledge is the awareness of strategies, that is, the awareness that a certain strategy exists. Procedural knowledge describes how a strategy effectively works, and conditional knowledge helps us understand which strategies are useful for solving a certain task. Whereas declarative and procedural knowledge about strategies can be considered prerequisites for strategic learning, conditional knowledge additionally enables the learner to choose an adequate strategy in a given situation and to be responsive to changing circumstances. Therefore, conditional strategy knowledge provides an important basis for the selection of adequate strategies in concrete learning situations (Neuenhaus, 2011).

The importance of declarative metacognition in the educational context has been documented in many studies. Accordingly, metacognitive knowledge (e. g., knowledge about variables that affect memory performance or knowledge about memory strategies that support retention) has been found to be related to memory development and strategy use within many cross-sectional studies (Schneider & Pressley

1997; Schneider, Schlagmüller, & Visé, 1998). Similarly, more specific metacognitive knowledge and metacognitive skillfulness regarding text processing and mathematics proved to be substantial predictors of test performance in their respective domains even after differences in intellectual abilities were taken into account (Artelt, Schiefele, Schneider, & Stanat, 2002; Veenman, Kok, & Blöte, 2005). Training studies give further evidence for the impact of metacognitive knowledge in the educational context. For instance, various intervention approaches providing metacognitive information in addition to strategy training have revealed that metacognitive information about the value of being strategic increases the probability that children will learn the strategy and later use it (see Joyner & Kurtz-Costes, 1997).

1.1 Development of Metacognitive Knowledge

Despite the rather traditional view that metacognition does not emerge before primary school, at which point children encounter formal learning (see Veenman, Van Hout-Wolters, & Afflerbach, 2006), studies involving younger children generally show that the acquisition of metacognitive knowledge begins as early as in Kindergarten. Children from the age of four years onwards seem to have at least some basic understanding of memory and learning processes. For instance, they begin to understand that they can forget things, that it is harder to remember more items compared with only a few items, and that additional study time may be helpful (Ebert, 2011; Kreuzer, Leonard, & Flavell, 1975; Lockl & Schneider, 2006, 2007; Wellman, 1977). Generally speaking, these children begin to appreciate the active role of the mind in learning and remembering (Wellman & Hickling, 1994). Longitudinal studies provide evidence that children's metacognitive knowledge is influenced by earlier theory-of-mind competencies, that is, by their developing ability to attribute mental states to themselves and others as well as by the amount of mental-state language used by their mothers (Ebert, 2011, 2015; Lockl & Schneider, 2006, 2007). This is also consistent with the view that the origin of metacognitive knowledge lies in interaction with other, more knowledgeable persons (e.g., Bruner, 1990; Vygotsky, 1978).

Once children enter school, their metacognitive knowledge—especially their knowledge about the importance of task characteristics and memory strategies—increases rapidly. For instance, several studies that focused on organizational strategies have reported a major shift in strategy knowledge between Kindergarten and Grade 6 (e.g., Justice, 1985; Schneider, 1986; Sodian, Schneider, & Perlmutter, 1986). Nevertheless, even adolescents and young adults lack knowledge about strategies when the task is to read, comprehend, and memorize complex text materials (Schneider, 2008). Hence, metacognitive knowledge continues to develop beyond adolescence during the entire life span (see Alexander & Schwanenflugel, 1996; Artelt, Neuenhaus, Lingel, & Schneider, 2012; Baker, 1989; Hasselhorn, 2006; Schneider & Lockl, 2006).

1.2 Assessment of Metacognitive Knowledge

In order to measure metacognitive knowledge in younger children, most approaches have applied interviews that include questions about strategies and memory processes, as was done in the classic interview study by Kreutzer et al. (1975) as well as in subsequent studies (e.g., Cavanaugh & Borkowski, 1980; Schneider, 1986). However, there are several limitations to these assessment procedures, especially with respect to the assessment criteria in the NEPS. First, the test-retest correlations and internal consistencies of these measures were often only moderate (Hasselhorn, 1994; Kurtz, Reid, Borkowski, & Cavanaugh, 1982). Second, interview assessments may be questioned, particularly when considering younger children, because these assessment methods rely to a great extent on language (Cavanaugh & Perlmutter, 1982; Fritz, Howie, & Kleitman, 2010; Joyner & Kurtz-Costes, 1997). In order to answer the questions correctly, children need a certain degree of receptive and productive language skills. That is, they have to understand the questions and the task requirements, and they need the vocabulary and the grammatical skills necessary to explain why it is difficult to remember something or why a strategy is useful, especially if open-ended or justification questions are asked. Third, interview methods may also put strong requirements on children's working memory capacity because they have to keep track of what the experimenter says. To deal with these problems, several studies have included ranking methods instead of open-ended questions and/or supplementary visual material illustrating the content of the questions (Annevirta & Vauras, 2001; Fritz et al., 2010; Lockl & Schneider, 2007). Finally, interviews are not suitable for the special demands within the NEPS because they cannot be administered in a group setting.

For older children and adolescents, several tests suitable for group settings have been developed in recent years (e.g., the reading strategy knowledge test for Grades 7 to 12 (WLST) by Schlagmüller & Schneider, 2007; or the metacognitive strategy knowledge test concerning reading strategies implemented in several languages within the PISA-OECD states, see Artelt, Beinicke, Schlagmüller, & Schneider, 2009). These tests typically focus on conditional metacognitive knowledge and consist of a number of scenarios describing challenging situations. Each scenario is followed by a list of approaches of differing strategic quality, and participants are asked to rate the usefulness of each alternative. For the scoring of the test, pair comparisons (option X is more or less useful than option Y) are judged with reference to experts' ratings of the relative usefulness of the presented strategies (see Händel et al., 2013). It has been shown that these test instruments are reliable and economic in use, that they refer to concrete learning situations, and that they are interpretable against a well-defined standard (Artelt & Schneider, 2015). However, these tests are not suitable for younger children for several reasons (e.g., test difficulty, test length, and dependence on reading abilities).

Therefore, a new instrument had to be developed to assess metacognitive knowledge in first graders. Because of the advantages mentioned above, we chose the gen-

eral rationale of the scenario-based approach, which is also used with students in secondary schools within the NEPS (Händel et al., 2013). As the NEPS aims to track metacognitive development across long stretches of the life span, this approach also allows for a consistent longitudinal empirical assessment. The scenarios and the proposed strategies refer to memory and learning in general (domain-general metacognitive knowledge) because the NEPS, as a longitudinal study, intends to assess metacognitive knowledge in a domain-general way detached from particular school subjects or content domains. At the same time, the test instrument covers a broad range of scenarios, including school-relevant and leisure-time activities, and focuses on knowledge about the appropriateness of different strategies in these situations (conditional metacognitive knowledge). However, compared with the test instrument for older students, some characteristics of the test had to be modified in order to be appropriate for first graders. In the following section, we report on a pilot study that was carried out to investigate the appropriateness of the newly constructed scenarios for first graders and to provide an empirical basis for the selection of items for the final test instrument on metacognitive knowledge. Furthermore, we evaluate the psychometric properties of the final test instrument and investigate whether we have succeeded in the construction of a homogeneous one-dimensional test for the assessment of general metacognitive knowledge.

2 Method

2.1 Participants

In total, 195 first graders (43 % female) participated in the pilot study. The majority of the children (96 %) were 6 or 7 years of age. The children were recruited from 5 different schools in Bavaria, Germany, and came from families of diverse social backgrounds. All children had written parental consent to participate in the study.

2.2 Materials

To assess metacognitive knowledge, a total item pool of 28 scenarios was initially constructed. The scenarios as well as the corresponding strategies were partly based on former studies investigating metacognitive knowledge in Kindergarteners and young school children (Haberhorn, Lockl, Pohl, Ebert, & Weinert, 2014; Kreutzer et al., 1975; Lockl & Schneider, 2007; Wellman, 1977). In the process of test construction, the scenarios were modified and optimized through cognitive interviews with children as well as by discussions in an expert team. For the pilot study, 20 of the initial 28 scenarios were selected. The scenarios focused on conditional metacognitive knowledge, that is, knowledge about the appropriateness of different strategies


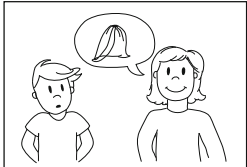
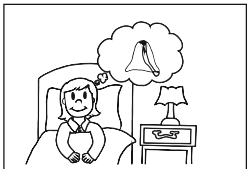
in varying situations, and included cognitive, metacognitive, and resource management strategies. Accordingly, the items assessed knowledge about solving cognitive tasks like remembering or organizing information, about planning and regulating, and about general learning requirements, such as using resource management strategies. As in the test for secondary-school students (see above), some of the scenarios were related to school-relevant activities, whereas the remaining scenarios were embedded in out-of-school contexts and described leisure-time activities. Because a previous study with secondary students had shown that using a third-person perspective seemed more suitable for measuring metacognitive knowledge in contrast to a first-/second-person perspective, the scenarios were phrased in the third-person perspective (Händel et al., 2013). That is, in each scenario, a female or male actor (labeled with a typical male/female name and illustrated as a girl/boy in the respective strategy alternatives) dealing with a specific situation was described.

Taking into account children's restricted reading abilities in the first grade, we decided to present the scenarios and proposed strategies orally, accompanied by pictures. That is, the experimenter read the scenarios and the corresponding strategies aloud, and the children could follow each approach by looking at the pictures. In comparison with the test instrument developed for sixth and ninth graders in the NEPS, we also decided to reduce the number of the presented alternatives, which was intended to result in lower demands on children's working memory capacity. Moreover, the format of the answer scale was changed in a child-appropriate manner: Children had to rate each strategy on a 3-point Likert scale labeled with a different number of stars (1, 2, or 3) indicating the usefulness of the strategy (with 1 star representing low usefulness and 3 stars representing high usefulness; see sample item in Figure 1). To control for fatigue or test order effects, four booklets were composed that differed in the order of the scenarios. In two of the four booklets, scenarios 1 to 10 were in the first half of the booklet, whereas in the remaining two booklets, scenarios 11 to 20 were presented in the first half of the booklet. In addition, the items within each half of the booklets were provided in two different orders (forwards and backwards).

Figure 1 Example Scenario

Eva has gym class the following day. She already put all her clothes into her gym bag the previous evening. What can Eva do in order not to forget her clothes the next morning?

The following three pictures show what Eva can do in order not to forget her sports clothes.

	<p>Eva hangs her bag on the front door.</p>	<p>☆ ☆☆ ☆☆☆</p> <p><input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/></p>
	<p>Eva asks her little brother to remind her.</p>	<p>☆ ☆☆ ☆☆☆</p> <p><input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/></p>
	<p>Eva thinks hard about her bag before she falls asleep.</p>	<p>☆ ☆☆ ☆☆☆</p> <p><input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/></p>

Note: In the test booklets for the children, no text is provided, but the text is read aloud to them.

2.3 Procedure

The pilot study was administered in a group setting at children’s schools. To ensure that all children would follow the instructions and could sit at their own table, the group size was restricted to 14 children (this group size is also realized in the NEPS main study). To acquaint the children with the general procedure, a practice scenario was presented via a poster in front of the class. The experimenter explained the use of the rating scale and asked the children not to tick the boxes of the rating scale before all approaches for a given scenario had been read aloud. Following these instructions, the children turned the page in their booklets to the first scenario, and the experimenter read aloud the scenario with the corresponding approaches. Then, the children were given about 45 seconds to mark the boxes for the scenario. The procedure was repeated for all 20 scenarios. Overall, the completion of the test took approximately 30 minutes.

In addition to the metacognitive knowledge task itself, we collected data on gender, age, and language background. That is, children were asked whether they speak only German, German and another language, or only another language with their families.

2.4 Expert Ratings

To establish content validity for the test instrument, experts were asked to comment on the scenarios (see below) and to provide their judgments on the effectiveness of the proposed strategies. A test booklet that contained all relevant information (text about the scenarios and strategy alternatives as well as the images illustrating the answer options and the rating scale) was constructed for the experts. Thirteen experts (scientists in the field of educational psychology and learning strategies) rated the strategy options according to the options' appropriateness for the respective scenarios of the test. These ratings served as basis for developing an objectified scoring procedure for the students' responses, that is, the 20 scenarios with three strategy options each were scored with reference to the pairwise comparisons provided by the experts. A pair comparison is considered to be valid for the assessment of metacognitive knowledge if experts agreed to at least 75 % in the direction of the pair comparison (i. e., 75 % or more of the experts rated a strategy option as superior or subordinate to another). This procedure resulted in 50 valid pair comparisons. While some scenarios included two valid pair comparisons (as was the case in the example scenario in Figure 1, in which Strategy One is rated superior to both Strategy Two and Strategy Three), other scenarios include three valid pair comparisons.

2.5 Analysis Procedure

Analyses based on Item Response Theory (IRT) were conducted to examine the psychometric properties of the item pool and the final test instrument using the software ConQuest (Wu, Adams, & Wilson, 1997). The Rasch model (Rasch, 1980) was chosen for scaling the data as this preserves the equal weights of pair comparisons intended by construction. Marginal maximum likelihood estimation was used for estimating the parameters. The missing values on the variables were modeled as missing responses, as suggested by Gräfe (2012) as well as by Pohl, Gräfe, and Rose (2014). Extensive analyses on the item pool were performed in order to construct an appropriate final instrument on metacognitive knowledge. Additionally, a detailed quality check of the final instrument was undertaken that included analyses about its dimensionality. To evaluate the fit of the items to the underlying model, weighted mean squares (WMNSQ) and the respective *t*-values, point-biserial correlations between the item score and the total score, and the item characteristic curves were taken into

account. According to the rules of thumb given by Pohl and Carstensen (2012) and with regard to the small sample size of the pilot study, items with $0.85 < WMNSQ < 1.15$ and an item total correlation of $> .30$ were judged as having a good item-fit. Furthermore, the fairness of the test was examined for the variables of gender, language background, and the order of the scenarios in the test by performing differential item functioning (DIF) analyses. DIF occurs when subgroups differ in their probability of a correct response to an item after their overall differences have been controlled for. The size of DIF was reviewed with respect to the criteria by Pohl and Carstensen (2012). Differences in estimated difficulties greater than 1 were judged as a very strong DIF, absolute differences between 0.6 and 1 as worthy of further investigation, differences between 0.4 and 0.6 as mildly considerable DIF, and differences below 0.4 as not considerable DIF. Moreover, the fit of the models including only main effects and the models that additionally estimated DIF were compared for the two variables of gender and language background based on the Akaike's (1974) criterion (AIC) and the Bayesian information criterion (BIC, Schwarz, 1978). Since the final instrument consisted of tasks describing school activities and tasks referring to leisure-time activities, a two-dimensional model was finally compared with the uni-dimensional model using AIC and BIC.

3 Results

We first present the results of the analyses on the total item pool. Next, we report the criteria that were applied for the item selection of the final instrument. Finally, we show the psychometric properties of the final metacognitive knowledge test, including analyses about its dimensionality.

3.1 Preliminary Analyses on the Item Pool

The item pool consisted of 20 scenarios to assess metacognitive knowledge with 50 pair comparisons. These pair comparisons were scored as dichotomous variables, with 1 indicating a correct response (judgment on a strategy pair in line with the experts' ratings) and 0 indicating an incorrect response (judgment on a strategy pair contrary to the expert ratings, or the two strategies of a pair were considered as equal). The difficulty of the pair comparisons had a considerable range from -2.22 to 1.53 logits. In addition, the test targeting revealed that the items of the item pool covered a wide range of the persons' abilities. Many items were located in the medium ability distribution and thus yielded differentiate estimates for most of the subjects. To evaluate the fit of the items to the model, different fit statistics were investigated. The WMNSQ of the items ranged from 0.88 to 1.17 , with respective t -values from -2.70 to 3.30 . Only for one pair comparison did the t -value indicate significant deviances

between the empirical and the model-implied probabilities. All other items exhibited a good item fit with $0.85 < WMNSQ < 1.15$ and a non-significant t -value. The item total correlations varied between .05 and .52, with an average discrimination of .31. Fourteen out of 50 pair comparisons had a rather low item total correlation of $< .30$. All other items had a good point-biserial correlation above .30.

To examine the fairness of the test, that is, to test for possible item bias favoring one group or the other, DIF analyses for the variables of gender, language background, and position of the items in the test were conducted. Considerable differences in difficulty between boys and girls above 0.6 logits occurred for three pair comparisons after controlling for group differences. Twelve pair comparisons also showed a position DIF above 0.6 logits. The results concerning position-related DIF point to fatigue effects since many items were comparatively easier when the items were presented in the first half of the test booklet. Additionally, a substantial language background DIF emerged for eight items.

Altogether, the IRT analyses revealed important information about the items' quality and their functioning. In addition to other criteria described in the next paragraph, we drew on the statistical results to select appropriate items for the final test.

3.2 Selection of Items

The test for the NEPS main study is scheduled to take 15 minutes—which means that only half of the scenarios from the pilot study can be administered here. Accordingly, 10 scenarios were selected for the main study. The selection of the scenarios for the final test instrument was based on different criteria that were carefully weighed up against each other.

First, we considered the psychometric quality of the paired comparisons as described above. That is, we excluded items that showed an unsatisfactory item fit with $WMNSQ > 1.15$ and a rather low discrimination value (item total correlation) below .20. As the pair comparisons are part of the scenarios, we only included scenarios if at least two corresponding pair comparisons with good item fit and discrimination values above .20 were obtained. We also drew on the DIF analyses to select items. Contents of all pair comparisons that showed considerable DIF were checked carefully to detect sources of unfairness/bias within the tasks. The investigation of items for which gender DIF was observed provided evidence that specific traits or preferences of boys and girls might have influenced children's responses. Scenarios containing these pair comparisons were therefore not included in the final instrument. All items with considerable DIF above 0.6 logits due to differences in gender and language background that revealed sources of unfairness/group-specific bias when checking their contents were excluded from the final instrument.

As a second criterion, the test was intended to be well targeted to a person's abilities and to cover a wide range of difficulties of the pair comparisons.

Third, we took into account the additional comments made by the experts with regard to the proposed strategies and scenarios. A few experts raised concerns that four of the scenarios might not be well suited to assess children's conditional metacognitive knowledge about strategies. Hence, these scenarios were excluded. (Two of these scenarios also showed rather poor values with regard to psychometric quality.)

Fourth, we aimed to achieve a balanced assessment of different areas of metacognitive strategy knowledge and intended to make sure that the remaining scenarios covered a broad range of possible strategies. Accordingly, the items were selected in such a way that half of the scenarios were related to school-relevant activities, whereas the other half of the scenarios described leisure-time activities. Regarding specific aspects of children's knowledge about strategies, the final instrument included scenarios related to prospective memory, organizational strategies, planning activities, remembering information, metacognitive control, and the impact of study time.

3.3 Psychometric Properties of the Final Instrument

The final item set included 10 scenarios with 22 pair comparisons. The pair comparisons' difficulties, fit indices, and differences in difficulty, which were obtained from the differential item functioning analyses, are depicted in Table 1.

After constraining the mean of the latent ability of the participants to zero, the item difficulty of the final item set varied between -1.58 and 1.19 logits. All items had a satisfactory item fit of $0.85 < WMNSQ < 1.15$, and no t -value indicated significant differences between the model implied and the empirically estimated probabilities of responses to each item. A good fit of the items to the models was also obtained when inspecting the item characteristic curves. All items showed a point-biserial correlation above $.20$ with the total score, and the average correlation was good ($r = .41$).

In order to judge the appropriateness of the test for first graders, test targeting was evaluated. Figure 2 presents the match between persons' abilities and the item difficulty on the same scale. The items were well targeted to the persons' ability distribution. They covered a wide range of difficulties, and the children with medium ability, in particular, were measured precisely. The final instrument had a satisfactory reliability (EAP/PV reliability = $.77$ and WLE reliability = $.74$).

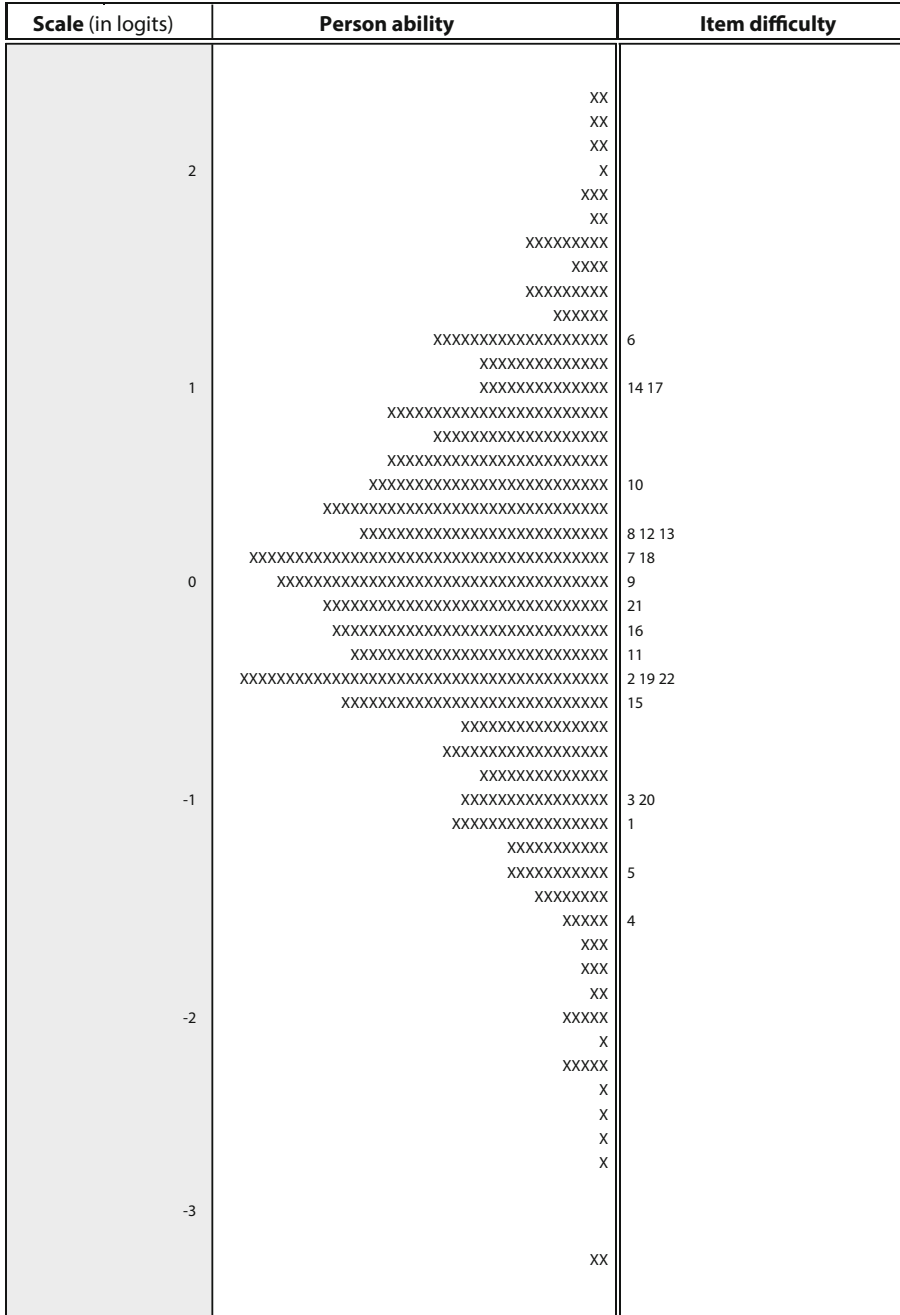
With regard to the difficulty of the items, it was found that pair comparisons related to prospective memory (pair comparisons 1, 2, 19, 20) and pair comparisons dealing with the impact of study time (pair comparisons 3, 4, 5) were relatively easy. In contrast, pair comparisons that referred to metacognitive control strategies (pair comparisons 17, 18), semantic categorization (pair comparisons 13, 14), or other strategies that supported remembering (pair comparisons 6, 7, 8, 9, 10) were relatively difficult (see Table 1 and Figure 2).

For the final item set, DIF was investigated at the item- as well as the test level. At the item level, the DIF values per item for the variables of gender and language

Table 1 Item Parameters of the Final Test Instrument

Pair comparison	Percentage correct	Difficulty	SE (difficulty)	WMNSQ	t-value of WMNSQ	Pt.bis-Correlation
PC1	71.35	-1.05	0.17	0.93	-0.8	.49
PC2	57.81	-0.36	0.16	1.11	1.9	.31
PC3	70.98	-1.03	0.17	1.00	0.0	.42
PC4	79.79	-1.58	0.19	0.97	-0.2	.42
PC5	76.80	-1.37	0.18	0.93	-0.7	.46
PC6	26.29	1.19	0.17	1.09	1.1	.25
PC7	46.39	0.18	0.16	0.99	-0.2	.43
PC8	44.33	0.27	0.16	0.97	-0.5	.46
PC9	48.45	0.08	0.16	0.96	-0.7	.48
PC10	40.21	0.47	0.16	0.99	-0.2	.44
PC11	56.48	-0.30	0.16	0.89	-2.0	.55
PC12	43.23	0.32	0.16	1.02	0.3	.40
PC13	44.74	0.24	0.16	0.99	-0.2	.43
PC14	29.63	0.99	0.17	1.09	1.2	.26
PC15	59.90	-0.47	0.16	0.89	-2.0	.55
PC16	53.13	-0.15	0.16	0.95	-0.9	.50
PC17	29.02	1.03	0.17	1.11	1.4	.24
PC18	45.60	0.21	0.16	1.05	0.9	.36
PC19	59.38	-0.44	0.16	1.05	0.8	.35
PC20	69.79	-0.97	0.17	0.93	-0.9	.47
PC21	50.56	-0.03	0.16	1.04	0.7	.38
PC22	59.55	-0.46	0.17	1.04	0.7	.37

Figure 2 Test Targeting. Person abilities are depicted on the left side of the graph, item difficulties on the right.



Each 'X' represents 0.3 cases

background were explored. Only one item exhibited a considerable gender DIF, and three items showed a considerable language background DIF above 0.6 logits in absolute differences of difficulty. Nevertheless, no evidence for unfairness (e. g., any gender- or culture-specific contents) was found when inspecting the contents of these items. At the test level, the fit of two models for the variables of gender and language background was compared. One model included main effects only, and the other additionally estimated DIF per item. The AIC as well as the BIC for both variables exhibited lower values for the more parsimonious model, and thus, the model that only embedded main effects was preferred over the more complex model. In summary, there was no substantive indication of unfairness/bias at the item level or the test level.

Finally, whether the scenarios referring to school or leisure-time activities formed a unidimensional or a multidimensional measure was explored. For this purpose, the dimensionality of the metacognitive knowledge test was examined by applying a two-dimensional model to the data. The latent correlations of the dimensions were observed, and the fit between the uni- and multidimensional models was compared. Latent correlations, variances, and fit indices are given in Table 2.

Regarding the fit indices, the multidimensional model fit the data slightly better (AIC = 5173.14, BIC = 5254.96) than did the unidimensional model (AIC = 5187.19, BIC = 5262.47). The AIC as well as the BIC preferred the more complex model with two dimensions. The latent correlation of .75 between scenarios referring to leisure-time vs. school activities also yielded certain multidimensionality. It deviated substantially from a perfect correlation of $> .95$ (see Carstensen, 2013) and thus indicated that the two dimensions measured different albeit highly correlated components of metacognitive knowledge.

Table 2 Latent Correlations and Variances of the Two-Dimensional Model

	Dim 1	Dim 2
Scenarios referring to leisure-time activities (Dim 1) (Number of pair comparisons = 11)	1.04	
Scenarios referring to school activities (Dim 2) (Number of pair comparisons = 11)	0.75	0.79

Note. Variances are given in the diagonal, correlations in the off-diagonal.

4 Discussion

The study presented in this chapter focuses on the development and evaluation of a test instrument on metacognitive knowledge for first graders to be used in the NEPS. On the one hand, the new instrument is appropriate for first graders with little reading ability and limited working memory capacity; on the other hand, it is comparable with the assessment of metacognitive knowledge in group settings with older students within the NEPS. Compared with tests for older students, the scenarios and proposed strategies were presented orally along with pictures, the number of the presented options was reduced, and the format of the answer scale was changed to be more child appropriate. The results of the pilot study indicate that the newly developed test instrument is age-appropriate, reliable, and suitable for a group setting. The pilot study also demonstrated that 10 scenarios may be administered within 15 minutes—the scheduled processing time in the NEPS main survey. Though fatigue effects were observed in the pilot study, these occurred only in the second half of the test booklets, that is, at the end of the test including 20 scenarios. Overall, we succeeded in constructing an economic test instrument that is suitable for the purposes within the NEPS and that allows for the assessment of metacognitive knowledge from a longitudinal research perspective.

The results concerning the difficulties of the pair comparisons are consistent with the findings of previous studies. For instance, in line with Kreutzer et al. (1975) and Yussen and Bird (1979), most of the children in the first grade were able to appreciate the impact of study time on learning outcome. Most children also successfully mastered the pair comparisons regarding prospective memory (e.g., not forgetting to take the sports clothes to school), which has also been shown in other studies (e.g., Cavanaugh & Borkowski, 1980; Kreutzer et al., 1975; Weinert & Schneider, 1999) and indicates that children of this age understand that people forget things and that strategies that prevent forgetting are beneficial. In contrast, less than half of the children seemed to have appropriate knowledge about the usefulness of memory strategies such as semantic categorization. This result is consistent with the findings of many studies, suggesting that this aspect of metacognitive knowledge emerges somewhat later (e.g., Justice, 1985; Schneider, 1986; Sodian et al., 1986).

With regard to the evaluation of the psychometric properties, the items in the final test showed a very good fit to the Rasch model, with appropriate item total correlations and item characteristic curves. DIF analyses provided evidence that the test was fair/unbiased for the considered subgroups composed by their gender and their language background. The appropriateness of the test for the specific target group was confirmed since the items were well targeted to the distribution of person's abilities. Furthermore, the test has been shown to have high reliability. The analyses on dimensionality point to some multidimensionality based on scenarios referring to leisure-time or school activities. These findings are consistent with other studies on metacognitive knowledge, in which rather low correlations between dimensions

are reported (Haberkorn et al., 2014; Neuenhaus, Artelt, Lingel, & Schneider, 2011; Schlagmüller, Visé, & Schneider, 2001), indicating that metacognitive knowledge is a rather heterogeneous construct. Nevertheless, the NEPS aims at assessing a broad and comprehensive construct of metacognitive knowledge. Such a balanced assessment of metacognitive strategy knowledge can only be achieved if a broad range of scenarios with different strategies is included. Therefore, notwithstanding the empirical indications for some multidimensionality, based on theoretical arguments, one metacognitive competence score is formed across the items and provided to the scientific community.

The pilot study described in this chapter has some limitations. First, the size of the sample of the pilot study is relatively small. Second, the information regarding language background was provided only by the children themselves. Therefore, this information cannot be assured to be reliable. Third, the IRT analyses were completed based on the single pair comparisons. However, the fact that the pair comparisons refer to specific scenarios and thus might partially depend on each other has to be taken into account. Further research considering the main survey data may allow for estimating the impact of local item dependence and for applying models that have been developed for such item bundles, such as the partial credit model (Masters, 1982). Fourth, due to practical constraints, no further competencies besides metacognitive knowledge could be assessed in this pilot study. Data from the NEPS main survey will help to answer the question of how metacognitive knowledge is related to other competencies, such as German language competencies, mathematical competence, scientific literacy, and domain-general cognitive abilities. Furthermore, considering metacognitive knowledge in a longitudinal perspective will allow for investigating how metacognitive knowledge contributes to the growth of other competencies and how metacognitive knowledge itself is influenced by cognitive and motivational factors during development.

References

- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6), 716–722.
- Alexander, J. M., & Schwanenflugel, P. J. (1996). Development of metacognitive concepts about thinking in gifted and nongifted children: Recent research. *Learning and Individual Differences*, 8(4), 305–325.
- Annevirta, T., & Vauras, M. (2001). Metacognitive knowledge in primary grades: A longitudinal study. *European Journal of Psychology of Education*, 16, 257–282. doi:10.1007/BF03173029
- Artelt, C., Beinicke, A., Schlagmüller, M., & Schneider, W. (2009). Diagnose von Strategiewissen beim Textverstehen. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, 41(2), 96–103.

- Artelt, C., Neuenhaus, N., Lingel, K., & Schneider, W. (2012). Entwicklung und wechselseitige Effekte von metakognitiven und bereichsspezifischen Wissenskomponenten in der Sekundarstufe. *Psychologische Rundschau*, *63*, 18–25. doi:10.1026/0033-3042/a000106
- Artelt, C., Schiefele, U., Schneider, W., & Stanat, P. (2002). Leseleistungen deutscher Schülerinnen und Schüler im internationalen Vergleich (PISA): Ergebnisse und Erklärungsansätze. *Zeitschrift für Erziehungswissenschaft*, *5*(1), 6–27.
- Artelt, C., & Schneider, W. (2015). Cross-country generalizability of the role of metacognitive knowledge for students' strategy use and reading competence. *Teachers College Record*, *117*(1), 1–32.
- Baker, L. (1989). Metacognition, comprehension monitoring, and the adult reader. *Educational Psychology Review*, *1*, 3–38. doi:10.1007/BF01326548
- Boekaerts, M. (1997). Self-regulated learning: A new concept embraced by researchers, policy makers, educators, teachers, and students. *Learning and Instruction*, *7*, 161–186. doi:10.1016/S0959-4752(96)00015-1
- Bruner, J. S. (1990). *Acts of meaning*. Cambridge, MA: Harvard University Press.
- Carstensen, C. H. (2013). Linking PISA competencies over three cycles—results from Germany. In M. Prenzel, M. Kobarg, K. Schöps, & S. Rönnebeck (Eds.), *Research Outcomes of the PISA Research Conference 2009* (pp. 199–213). New York: Springer.
- Cavanaugh, J. C., & Borkowski, J. G. (1980). Searching for metamemory-memory connections: A developmental study. *Developmental Psychology*, *16*, 441–453. doi:10.1037/0012-1649.16.5.441
- Cavanaugh, J. C., & Perlmutter, M. (1982). Metamemory: A critical examination. *Child Development*, *53*(1), 11–28.
- Ebert, S. (2011). *Was Kinder über die mentale Welt wissen—Die Entwicklung von deklarativem Metagedächtnis aus der Sicht der "Theory of Mind."* Hamburg: Dr. Kovac.
- Ebert, S. (2015). Longitudinal relations between theory of mind and metacognition and the impact of language. *Journal of Cognition and Development*, *16*, 559–586. doi:10.1080/15248372.2014.926272
- Flavell, J. H., Miller, P. H., & Miller, S. A. (1993). *Cognitive development*. Englewood Cliffs, NJ: Prentice-Hall, Inc.
- Flavell, J. H., & Wellman, H. M. (1977). Metamemory. In R. V. Kail, & J. W. Hagen (Eds.), *Perspectives on the development of memory and cognition* (pp. 3–33). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Fritz, K., Howie, P., & Kleitman, S. (2010). How do I remember when I got my dog? The structure and development of children's metamemory. *Metacognition and Learning*, *5*, 207–228. doi:10.1007/s11409-010-9058-0
- Gräfe, L. (2012). *How to deal with missing responses in competency tests? A comparison of data- and model-based IRT approaches* (Unpublished master's thesis). Friedrich-Schiller-University Jena, Jena, Germany.

- Haberkorn, K., Lockl, K., Pohl, S., Ebert, S., & Weinert, S. (2014). Metacognitive knowledge in children at early elementary school. *Metacognition and Learning*, 9, 239–263. doi:10.1007/s11409-014-9115-1
- Händel, M., Artelt, C., & Weinert, S. (2013). Assessing metacognitive knowledge: Development and evaluation of a test instrument. *Journal of Educational Research Online*, 5(2), 162–188.
- Hasselhorn, M. (1994). Zur Erfassung von Metagedächtnisaspekten bei Grundschulkindern. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, 26(1), 71–78.
- Hasselhorn, M. (2006). Metakognition. In D. H. Rost (Ed.), *Handwörterbuch Pädagogische Psychologie* (3rd ed., pp. 480–485). Weinheim: Beltz PVU.
- Joyner, M. H., & Kurtz-Costes, B. (1997). Metamemory development. In N. Cowan (Ed.), *The development of memory in childhood* (pp. 275–300). Hove: Psychology Press.
- Justice, E. M. (1985). Categorization as a preferred memory strategy: Developmental changes during elementary school. *Developmental Psychology*, 21, 1105–1110. doi:10.1037/0012-1649.21.6.1105
- Kreutzer, M. A., Leonard, C., & Flavell, J. H. (1975). An interview study of children's knowledge about memory. *Monographs of the Society for Research in Child Development*, 40(1), 1–60.
- Kurtz, B. E., Reid, M. K., Borkowski, J. G., & Cavanaugh, J. C. (1982). On the reliability and validity of children's metamemory. *Bulletin of the Psychonomic Society*, 19(3), 137–140.
- Lockl, K., & Schneider, W. (2006). Precursors of metamemory in young children: The role of theory of mind and metacognitive vocabulary. *Metacognition and Learning*, 1, 15–31. doi:10.1007/s11409-006-6585-9
- Lockl, K., & Schneider, W. (2007). Knowledge about the mind: Links between theory of mind and later metamemory. *Child Development*, 78, 148–167. doi:10.1111/j.1467-8624.2007.00990.x
- Masters, G. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47(2), 149–173.
- Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In G. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (pp. 125–173). New York: Academic Press.
- Neuenhaus, N. (2011). *Metakognition und Leistung. Eine Längsschnittuntersuchung in den Bereichen Lesen und Englisch bei Schülerinnen und Schülern der fünften und sechsten Jahrgangsstufe*. Bamberg: Opus.
- Neuenhaus, N., Artelt, C., Lingel, K., & Schneider, W. (2011). Fifth graders metacognitive knowledge: General or domain specific? *European Journal of Psychology of Education*, 26, 163–178. doi:10.1007/s10212-010-0040-7
- Paris, S. G., Lipson, M. Y., & Wixson, K. K. (1983). Becoming a strategic reader. *Contemporary Educational Psychology*, 8, 293–316. doi:10.1016/0361-476X(83)90018-8
- Pohl, S., & Carstensen, C. H. (2012). *NEPS technical report—Scaling the data of the competence tests*. (NEPS Working Paper No. 14). Bamberg: University of Bamberg, National Educational Panel Study.

- Pohl, S., Gräfe, L., & Rose, N. (2014). Dealing with omitted and not reached items in competence tests—Evaluating approaches accounting for missing responses in IRT models. *Educational and Psychological Measurement, 74*(3), 423–452.
- Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Chicago: University of Chicago Press.
- Schlagmüller, M., & Schneider, W. (2007). *Der Würzburger Lesestrategie-Wissenstest für die Klassen 7 bis 12 (WLST 7-12)*. Göttingen: Hogrefe.
- Schlagmüller, M., Visé, M., & Schneider, W. (2001). Zur Erfassung des Gedächtniswissens bei Grundschulkindern: Konstruktionsprinzipien und empirische Bewährung der Würzburger Testbatterie zum deklarativen Metagedächtnis. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 33*(2), 91–102.
- Schneider, W. (1986). The role of conceptual knowledge and metamemory in the development of organizational processes in memory. *Journal of Experimental Child Psychology, 42*, 218–236. doi: 10.1016/0022-0965(86)90024-X
- Schneider, W. (2008). The development of metacognitive knowledge in children and adolescents: Major trends and implications for education. *Mind, Brain, and Education, 2*, 114–121. doi: 10.1111/j.1751-228X.2008.00041.x
- Schneider, W. (2015). *Memory development from early childhood through emerging adulthood*. Cham: Springer International Publishing Switzerland.
- Schneider, W., & Lockl, K. (2006). Entwicklung metakognitiver Kompetenzen im Kindes- und Jugendalter. In W. Schneider, & B. Sodian (Eds.), *Kognitive Entwicklung. Enzyklopädie der Psychologie, Serie Entwicklungspsychologie* (pp. 721–767). Göttingen: Hogrefe.
- Schneider, W., & Pressley, M. (1997). *Memory development between 2 and 20*. Hillsdale, NJ: Erlbaum.
- Schneider, W., Schlagmüller, M., & Vise, M. (1998). The impact of metamemory and domain-specific knowledge on memory performance. *European Journal of Psychology of Education, 13*, 91–103. doi:10.1007/BF03172815
- Schwarz, G.E. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*, 461–464. doi: 10.1214/aos/1176344136
- Sodian, B., Schneider, W., & Perlmutter, M. (1986). Recall, clustering, and metamemory in young children. *Journal of Experimental Child Psychology, 41*, 395–410. doi: 10.1016/0022-0965(86)90001-9
- Veenman, M. V. J., Kok, R., & Blöte, A. W. (2005). The relation between intellectual and metacognitive skills at the onset of metacognitive skill development. *Instructional Science, 33*(3), 193–211.
- Veenman, M. V. J., Van Hout-Wolters, B. A. M., & Afflerbach, P. (2006). Metacognition and learning: Conceptual and methodological considerations. *Metacognition and Learning, 1*, 3–14. doi:10.1007/s11409-006-6893-0
- Vygotsky, L. (1978). *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.

- Weinert, F. E., & Schneider, W. (1999). *Individual development from 3 to 12: Findings from the Munich longitudinal study*. Cambridge: Cambridge Univ. Press.
- Weinert, S., Artelt, C., Prenzel, M., Senkbeil, M., Ehmke, T., & Carstensen, C. H. (2011). Development of competencies across the life span. In H.-P. Blossfeld, H.-G. Roßbach, & J. von Maurice (Eds.), *Zeitschrift für Erziehungswissenschaft, 14. Education as a lifelong process: The German National Educational Panel Study (NEPS)* (pp. 67–86). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Wellman, H. M. (1977). Preschoolers' understanding of memory-relevant variables. *Child Development, 48*(4), 1720–1723.
- Wellman, H. M., & Hickling, A. (1994). The minds "I": Children's conception of the mind as an active agent. *Child Development, 65*, 1564–1580. doi:10.1111/j.1467-8624.1994.tb00836.x
- Wu, M. L., Adams, R. J., & Wilson, M. R. (1997). *ACER Conquest: Generalised item response modelling software*. Melbourne: ACER Press.
- Yussen, S. R., & Bird, E. J. (1979). The development of metacognitive awareness in memory, communication, and attention. *Journal of Experimental Child Psychology, 28*, 300–313. doi:10.1016/0022-0965(79)90091-2

About the authors

K. Lockl
Leibniz Institute for Educational Trajectories (LifBi), Bamberg.
e-mail: kathrin.lockl@lifbi.de

M. Händel
University of Erlangen-Nuremberg, Nuremberg.

K. Haberkorn
University of Bamberg, Bamberg.

S. Weinert
Department of Psychology I: Developmental Psychology,
University of Bamberg, Bamberg.

Including Students With Special Educational Needs in the Competence Assessment of the NEPS—Results on the Comparability of Test Scores in Reading

Anna Südkamp, Steffi Pohl, Jana Heydrich and Sabine Weinert

Abstract

Including students with special educational needs in learning (SEN-L) is one of the National Educational Panel Study's (NEPS) challenges. In this study, we address the question of whether the reading competence of students with SEN-L may be assessed reliably with the reading test designed for general-education students. In addition, we ask whether the test scores of students with SEN-L can be compared with the test scores of students without SEN-L. The reading competence of $N = 176$ students with SEN-L and $N = 5,208$ general-education students is assessed with the NEPS standard reading test for students in Grade 5. The results of test targeting and item fit reveal that the items of the NEPS standard reading test are rather difficult for students with SEN-L, while item discrimination is low for many items of the test. With respect to measurement invariance, a substantial number of items show differential item functioning, indicating that the standard reading test measures a different construct for students with and without SEN-L. Implications for further research are indicated in the discussion.

1 Introduction

Today, educational assessments play an important role in society as they inform students, parents, educators, policy-makers, and the public about the effectiveness of educational services (Pellegrino, Chudowsky, & Glaser, 2001). Using results from large-scale assessments, factors influencing the acquisition and development of competencies can be studied and strategies on the improvement of educational systems can be derived. Tests within large-scale assessments aim at a valid and reliable measurement of competencies while—at the same time—being both time- and cost-efficient. In order to assure objectivity, tests are usually administered under standard-

ized conditions. Testing is a highly demanding situation from each of the different perspectives of test-administrators, test-takers, parents, and teachers (Guthrie, 2002). For example, Abrams, Pedulla, and Madaus (2003) report that teachers frequently feel pressured to raise test scores. At the same time, increased levels of anxiety, stress, and fatigue have been observed among students. When it comes to testing students with special educational needs (SEN), the challenges of testing seem to be even higher since there might be specific barriers in large-scale assessments for students with SEN (Bolt & Ysseldyke, 2008). For example, students with visual impairments may not be able to access printed material, and students with learning disabilities may not be acquainted with these kinds of tests. However, giving students with SEN the opportunity to participate in large-scale assessments is an issue of fairness and equality. It is also highly relevant for being able to address important practical as well as theoretical questions in research on the developmental and educational pathways for students with SEN. Therefore, efforts have been made to reduce barriers in large-scale assessments and to include more students with special educational needs. Assessing students' domain-specific competencies (e. g., reading or mathematical competence) is a key aspect of the National Educational Panel Study (NEPS;¹ Weinert et al., 2011). The NEPS is a national large-scale longitudinal study that investigates the development of competencies across the lifespan (Blossfeld & von Maurice, 2011; Blossfeld, von Maurice, & Schneider, 2011). The study aims at providing high-quality, user-friendly data on competence development and educationally relevant processes for the international scientific community (Barkow et al., 2011). Between 2009 and 2012, six representative starting cohorts (Aßmann et al., 2011) were sampled, including about 60,000 individuals from early childhood to adulthood. Specific target groups include migrants (Kristen et al., 2011) and students with special educational needs in learning (SEN-L; Heydrich, Weinert, Nusser, Artelt, & Carstensen, 2013). Following the principles of universal design (Dolan & Hall, 2001; Thompson, Johnstone, Anderson, & Miller, 2005), the NEPS aims at providing a basis for fair and equitable measures of competencies for all individuals. In order to empirically address the question of whether and how students with SEN-L can be tested fairly, the NEPS has set up a series of feasibility studies. These studies focus on the validity of competence assessments. For example, we study the effects of testing accommodations for students with SEN-L on the reliability and comparability of test scores. Testing accommodations are generally defined as changes in test administration that are meant to reduce construct-irrelevant difficulty associated with students' disability-related im-

1 This paper uses data from the National Educational Panel Study (NEPS). The NEPS data collection is part of the Framework Programme for the Promotion of Empirical Educational Research, funded by the German Federal Ministry of Education and Research and supported by the German Federal States. Our research is based on the dedicated work of professors and research assistants, particularly those within the NEPS. We especially wish to thank Cordula Artelt, Claus H. Carstensen, Lena Nusser, and Markus Messingschlager. Our thanks also go to the staff of the NEPS survey administration and to the methods group.

pediments to performance. To test for group-specific effects and the comparability of test results and in order to discern—if necessary—whether test items do not function properly because the accommodations change the test construct or whether students with SEN-L still have problems with the test, we implement a control group of students from the lowest academic track, or *Hauptschule*. In addition, we gather in-depth background information on students with SEN-L in surveys of the students' parents, teachers, and school principals.

1.1 Inclusion of Students With Special Educational Needs (SEN) in Large-Scale Assessments

In Germany, the population of students with SEN comprises more than 485,400 individuals, which is around 6.4% of the entire student population (KMK, 2012). The question at hand is whether students with SEN can be tested reliably with the same test instruments and under comparable testing conditions as students without SEN. In the literature and in the field, this question has often been answered in the negative for theoretical as well as empirical and practical reasons. Therefore, students with SEN are still not being extensively included in large-scale assessments. Schools that are solely attended by students with SEN are excluded at the very beginning of the sampling procedure in the Progress in International Reading Literacy Study (PIRLS) as well as in the Programme for International Student Assessment (PISA) (Joncas, 2007; OECD, 2012). Whether students with SEN who are enrolled in general-education schools are included in these studies is mainly decided upon by local school staff even though all studies provide material to alleviate the decision-making process. In PIRLS, students with SEN are included as far as they are able to participate under standard conditions; otherwise, they are excluded. Contrary to PIRLS, PISA provides an extra “one hour” booklet specifically designed for students with SEN that contains half of the items of the standard test (OECD, 2012, p. 29). Surprisingly, in spite of a thorough description of the test design, main national PISA reports on Germany do not even mention the use of this booklet (cf. OECD, 2010). Despite a lack of studies and research reports on students with SEN, there is evidence that reading problems pose one of the greatest barriers to success in school for students with SEN (Kavale & Reece, 1992; Swanson, 1999).

1.2 Reading Performance of Students With SEN

On average, students with SEN² show a lower reading performance in large-scale assessments in comparison with students without SEN (Thurlow, 2010; Thurlow, Bremer, & Albus, 2008; Ysseldyke et al., 1998). For the 1998 National Assessment of Educational Progress (NAEP) of reading in Grades 4 and 8, Lutkus, Mazzeo, Zhang, and Jerry (2004) report lower average scale scores for students with SEN in comparison with students without SEN. Within the German study “Kompetenzen und Einstellungen von Schülerinnen und Schülern” (Bos et al., 2009), reading competence of seventh graders in special schools was compared with the reading competence of fourth graders attending general-education settings. Results demonstrated that fourth-grade primary-school students outperformed students with SEN in the seventh grade in reading competence, the difference being about one third of a standard deviation. Drawing on data from a three-year longitudinal study, Wu et al. (2012) found that students receiving special educational services were more likely to score below the 10th percentile for several years in a row compared with their general-education peers. In light of these findings, different reasons for the low performance of students with SEN have been discussed (Abedi et al., 2011). First, some students with SEN have difficulties related to the comprehension of text (e.g., a lack of knowledge of common text structures, restricted language competencies, inappropriate use of background knowledge while reading; Gersten, Fuchs, Williams, & Baker, 2001). Second, lower performance could be attributed to low teacher expectations and/or to a lack of opportunities to learn (Woodcock & Vialle, 2011). Third, there could be barriers for students with disabilities that lead to unfair testing conditions in large-scale assessments (Pitoniak & Royer, 2001). According to Thurlow (2010), a combination of all these factors is likely. Taking the norm of test fairness seriously, the NEPS tries to ensure that students with SEN will not be confronted with unfair testing conditions.

1.3 Assessment of Students With SEN With Standard Reading Tests

Providing students with SEN with standard reading tests has the advantage that no changes to the standard test instrument are necessary. Whenever changing a test instrument, there is a risk that test scores will not be comparable between groups tested with the standard test and accommodated test versions. Research on testing accommodations (Lovett, 2010; Pitoniak & Royer, 2001) has shown that testing accommodations may significantly alter standard test instruments, leading to test scores that

2 Note that students with SEN comprise a highly heterogeneous group, including, for example, students with visual impairments, hearing disabilities/impairments, and emotional and behavioral difficulties.

are no longer comparable. Nevertheless, students with SEN are often tested with accommodated test versions in large-scale assessments for practical reasons (Bolt & Ysseldyke, 2008; Pitoniak & Royer, 2001). So far, only a few studies have addressed the question of whether this is actually necessary, that is, whether students with SEN can also be tested validly and reliably with standard reading tests. As an exception, Koretz and Hamilton (2000; see also Koretz, 1997, for more detailed results) report that 19 % (Grade 4), 33 % (Grade 8), and 39 % (Grade 11) of students with SEN were tested without accommodations in the Kentucky Instructional Results Information System assessment. As data of students with SEN tested with and without accommodations were available, item difficulty, item discrimination, and differential item functioning (DIF) were analyzed. Unfortunately, not all results were reported (e. g., exemplifications of the target and reference group in DIF analyses are missing; DIF-values are not presented). Koretz (1997) concluded that item discriminations were comparable for students with and without SEN and that instances of DIF were few and generally minor for students with SEN who were tested without accommodations. In line with these results, Lutkus et al. (2004) did not identify any items with a strong indication of DIF for the 1999 NAEP reading assessment when comparing the results of students with disabilities tested without accommodations with the results of students without disabilities. Here, a split-sample design was implemented: Half of the sample of students with SEN were tested without accommodations, while the other half were tested with accommodations. In contrast, Bielinski et al. (2001) conclude—based on their item analyses including the root mean squared discrepancy and differential item functioning—that the reading test results of non-accommodated assessments of students with a primary disability in reading on the Missouri Assessment Program were not comparable with the results of other examinees. In summary, results on the comparability of test scores for students with and without SEN on standard reading assessments are mixed. Aside from differential item functioning, indicators of item fit are reported scarcely. Although testing accommodations are often used in the assessment of students with SEN in large-scale assessments, we consider it beneficial to first analyze whether testing students with SEN with standard test instruments is appropriate.

1.4 Research Questions

Taking the norm of test fairness seriously, the NEPS wants to ensure that students with disabilities are not confronted with barriers in the assessment. At the same time, we want to ensure reliable and valid measurements of competencies. While the need for specially developed test instruments is obvious for some students with special educational needs (e. g., providing visually-impaired students with tests in Braille), students with SEN-L can, in principle, be tested with standard-competence tests. However, psychometric problems (e. g., differential item functioning) might be expected. As students with SEN-L comprise the largest group of students with special educa-

tional needs (KMK, 2012; Koretz, 1997), the NEPS has decided to specifically focus on this group of students when setting up a series of feasibility studies in order to investigate whether and how valid competence measures can be obtained from students with SEN-L (Heydrich et al., 2013). In this chapter, we focus on the assessment of reading competence and report on an initial set of analyses based on the assessment of SEN-L students with the NEPS standard reading test (see Südkamp, Pohl, Hardt, Jordan, and Duchhardt (2015) for results on the NEPS assessment of mathematical competence). We address the question of whether students with SEN-L can be tested reliably with the NEPS standard reading test and whether the test results of students with SEN-L are comparable with those of general-education students.

2 Method

2.1 Sample and Design

The data of this study were collected within the NEPS. The study draws on two different samples within the NEPS: One concerns students with SEN-L, and the other concerns general-education students from the NEPS main sample. The sample of the feasibility studies comprised $N = 176$ students with SEN-L in fifth grade who were recruited at special schools for children with SEN-L in Germany. On average, these students were $M_{\text{age}} = 11.39$ ($SD_{\text{age}} = 0.65$) years old, and 46% were female. In Germany, students are assigned to the group of students with special educational needs in learning, when their learning, academic achievement, and/or learning behavior is impaired (KMK, 2012). The decision of whether a student is in need of special education is usually made jointly by parents, teachers, consultants, and school administrations. About 78% of the SEN-L students in Germany (KMK, 2012) do not attend regular schools but instead attend special schools with specific schooling programs and trainings tailored to those students who appear to be unable to follow school lessons and subject matter in regular classes. However, it is becoming more and more common to educate students with SEN-L at general-education schools as well. For the present study, students with SEN-L were exclusively drawn from special schools. As a reference group, the study draws on representative data from the NEPS main sample (Starting Cohort 3 in Grade 5; see Aßmann, Steinhauer, & Zinn, 2012, for more information on the sampling), which comprises $N = 5,208$ students in general-education schools ($M_{\text{age}} = 10.95$ years, $SD_{\text{age}} = 0.53$; 48.3% female).

2.2 Measures and Procedures

Reading and mathematical competences were assessed within both samples. Within the NEPS, the assessment of reading competence focuses on text comprehension,

which is often conceived of as the essence of reading (Durkin, 1993; Verhoeven & Van Leeuwe, 2008). Across all ages, starting in Grade 5, individuals read five different texts and are asked questions focusing on the content of these texts (Gehrer, Zimmermann, Artelt, & Weinert, 2013). The standard reading test was designed for students enrolled in the regular school system. The test was developed based on a conceptual framework that comprises five different text functions or text types and three different cognitive requirements (finding information in a text, drawing text-related conclusions, reflecting and assessing content). The items in the test were either multiple-choice (MC) items, complex MC items, or matching items (see Gehrer, Zimmermann, Artelt, & Weinert, 2012, for a description of the item formats in the reading test). Overall, 56 items were included in the analyses; however, subtasks of complex MC and matching items were treated as single items. When combined, there were 33 questions in the standard reading test, which students had to complete within 30 minutes. The test shows good psychometric properties for testing general-education students (Pohl, Haberkorn, Hardt, & Wiegand, 2012).

For the present study, all students were tested in the middle of their fifth-grade year in November and December 2010. Data were collected by the International Association of the Evaluation of Educational Achievement (IEA) Data Processing and Research Center (DPC) in Hamburg, Germany. Students participated in the study voluntarily, so student and parental consent was necessary. Each student who participated in the study received 5 euros.

2.3 Analyses

The model

We scaled the data within the framework of Item Response Theory (IRT). In accordance with the scaling procedure of competence data in NEPS (see Pohl & Carstensen, 2012), we used a Rasch model (Rasch, 1960) estimated in ConQuest (Wu, Adams, Wilson, & Haldane, 2007). As described above, the reading test also included complex MC and matching items. These items consisted of a set of subtasks that were aggregated to a polytomous variable in the final scaling model in the NEPS. When aggregating the responses on the subtasks to a single polytomous super-item, we lose information on the single subtasks. Since we are interested in the fit of the items in this study, we treated the subtasks of complex MC and matching items as single dichotomous items in the analyses.³

3 Note that we do not account for possible local item dependence within each set of subtasks with this analysis strategy.

Test targeting

In order to investigate whether the standard reading test was adequately targeted to the ability of the students with SEN-L, we evaluated test targeting. To do this, the estimated item difficulties were depicted on the same scale as the ability estimates. A test is considered well targeted if the item difficulties cover the whole range of ability estimates and there is no superfluity of items at the lower (too easy) or upper (too hard) end of the ability distribution.

Measures of fit

In order to investigate whether the standard reading test reliably measured reading competence for students with SEN-L, we evaluated different fit measures. For this analysis, we focus on the item discrimination, which describes the correlation of the item with the total score. A well-fitting item should have a high positive correlation, that is, subjects with a high ability should be more likely to score high on the item than subjects with a low ability. We considered a discrimination below .2 as a slight misfit and discriminations smaller than .1 as a strong item misfit.

Differential Item functioning

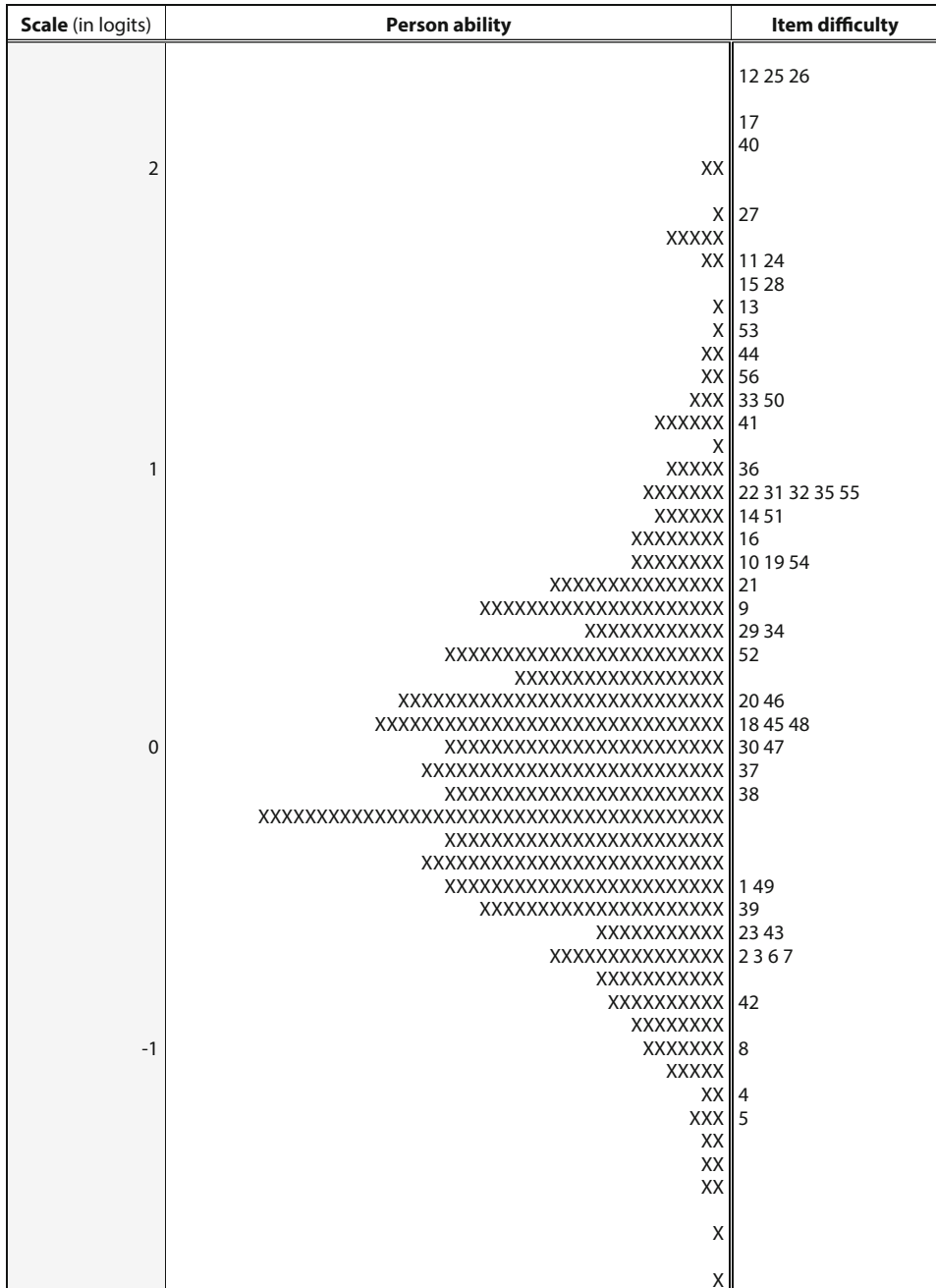
The comparability of the reading score of SEN students with those of general-education students can only be assured when the tests are measurement invariant—that is, when there is no DIF. When measurement invariance holds—and thus there is no DIF—the probability of endorsing an item is the same for students with SEN-L and general-education students who have the same ability. The presence of DIF is an indication that the respective reading test measures a different reading construct for both target groups and thus that the reading scores between the target groups may not be validly compared. We estimated DIF in a multi-group IRT model, estimating and comparing item difficulties for general-education students and students with SEN-L. In line with the benchmarks chosen in the NEPS (Pohl & Carstensen, 2012), we considered absolute differences in item difficulties greater than 0.6 to be noticeable and absolute differences greater than 1 to be strong DIF.

3 Results

3.1 Test Targeting

Figure 1 depicts the estimated item difficulties and the ability estimates of students with SEN-L on the same scale (in logits). In this analysis, the mean of the student's ability is set to zero. Ability estimates greater than zero indicate an above-average reading ability, while ability estimates smaller than zero indicate a below-average reading ability. Test takers with an ability that corresponds to the difficulty of an item have a 50 % probability of solving the item. Items with a lower difficulty are solved

Figure 1 Test targeting of the standard test in the group of SEN-L students. Item difficulties are depicted on the right side, person ability on the left side. Each number represents an item.



Each "X" represents 0.4 cases

with a higher probability, while items with a higher difficulty are solved with a probability lower than 50 %. Figure 1 shows that the item difficulties cover the whole range of students' abilities. However, the test is rather difficult overall. The gross of items is targeted towards students with high reading abilities. As a consequence, students with SEN-L may be overstrained by the test. As a comparison, the test is a bit too easy for students in general education (Pohl et al., 2012).

3.2 Item Fit

In Figure 2, item discrimination is displayed for the standard reading test in the group of students with SEN-L. Overall, item discrimination is relatively small for students with SEN-L. The mean item discrimination is .25. Four items show a slight misfit (discrimination less than .2 and equal to or greater than .1), and 10 items display a strong misfit (discrimination less than .1). As a comparison, there is no item misfit in the group of general-education students with the exception of one item that was excluded from the analyses. The item discrimination levels for general-education students are all above .3 (Pohl et al., 2012).

We further investigated the occurrence of item misfit in the standard test by estimating the correlation of the item difficulty estimated on general-education students (which is thus independent of the measurement model for SEN-L students) and the discrimination in the sample of SEN-L students. Within the group of students with SEN-L, item difficulty and discrimination correlated to $-.492$. The more difficult an item, the lower the discrimination is. That misfit occurs due to a disadvantageous test targeting—that is, due to inappropriate item difficulties for this target group. The items in the standard test are too difficult for students with SEN-L (mean item difficulty = 0.58 logits^4).

3.3 Measurement Invariance

Figure 3 shows the absolute differences in estimated item difficulties between general-education students and students with SEN-L who took the standard reading test. Positive values in the table indicate a higher item difficulty for general-education students as compared with students with SEN-L, while negative values indicate a lower item difficulty.

The results clearly show large differences in estimated item difficulties for students with SEN-L compared with general-education students. 12 out of 56 items have a slight DIF, and 14 items have a strong DIF. The results indicate that the test measures

4 Note that the mean of the reading ability is set to zero.

Figure 2 Discrimination of the items in the standard reading test for students with SEN-L

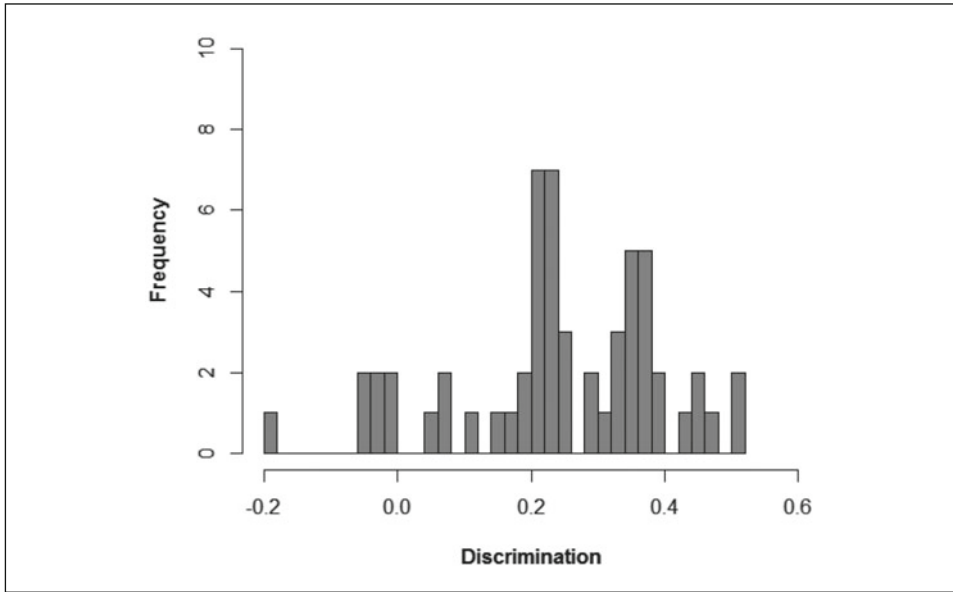
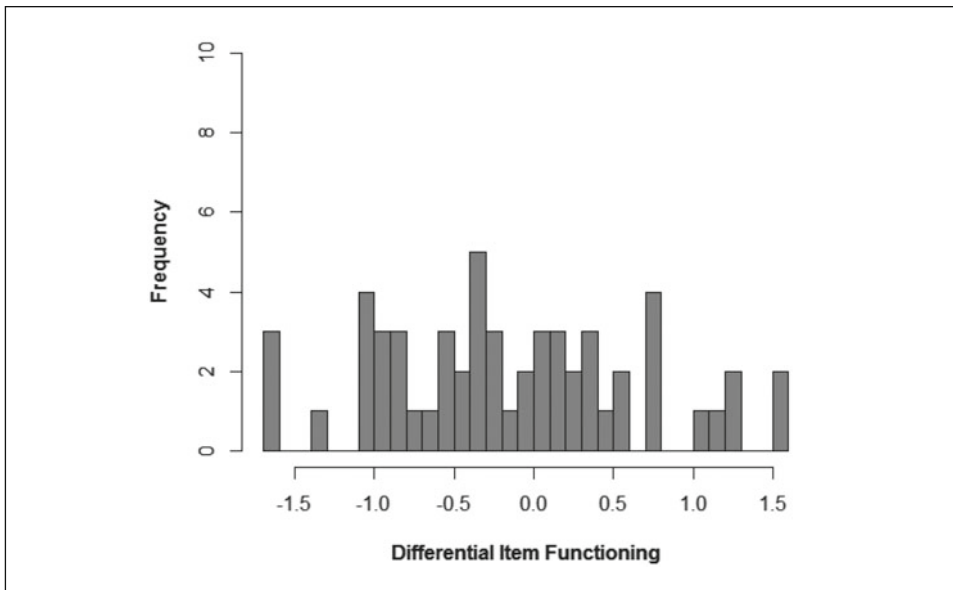


Figure 3 Differential item functioning of the items in the standard reading test. The graph depicts the differences in estimated item difficulties between students with SEN-L and general-education students



a different construct in the group of students with SEN-L as compared with general-education students. Reading-test scores for SEN-L students are thus not comparable with test scores for general-education students.

4 Discussion

The present study is part of a research program dealing with the question of how the competencies of students with SEN-L may be assessed reliably and comparably. In this chapter, we have addressed the question of whether the competencies of students with SEN-L in Grade 5 can be assessed reliably and comparably with the NEPS standard reading test. For this purpose, students with SEN-L were tested with the same test and under the same conditions as general-education students. As mentioned above, the standard reading test has shown good psychometric properties when testing high-achieving as well as low-achieving general-education students (Pohl et al., 2012).

The results on test targeting and item fit reveal that the items of the NEPS standard reading test are rather difficult for students with SEN-L. Item discrimination is low for many items of the test, showing that the items do not differentiate well between low-performing and high-performing students. With respect to measurement invariance, a substantial number of items show DIF, indicating that students with and without SEN-L cannot be measured on the same scale using the NEPS standard reading test.

With the present research, we contribute to the discussion of whether competencies of students with SEN may be assessed reliably and comparably by large-scale assessments. Our research overcomes problems of earlier studies on the assessment of students with SEN (see, e.g., Lovett, 2010). First, we concentrated our research on a specific group of students with SEN, namely students with learning disabilities. As such, we focus on a rather homogenous group of students and are able to disentangle whether the standard reading test is appropriate for a certain group of students with SEN.⁵ In contrast, many other studies on students with SEN include students with various disabilities, which leads to samples that are even more heterogeneous. Second, our sample of students with SEN-L was tested with the age-appropriate standard reading test, regardless of students' disability status. Thus, we were able to study the psychometric quality of the test in a sample of students with SEN-L, while there was no selection of especially capable students with SEN-L. Third, the results of our analyses are based on a relatively large representative sample of students with SEN-L.

5 Please note that the group of students with SEN-L is still a heterogeneous one, including, for example, students with different performance and ability profiles in the cognitive domain. Compared with prior research, however, the target population is rather homogeneous as students with SEN in areas other than learning (e.g., those with physical impairments) are precluded.

There is a complex research program in the NEPS dealing with the question of the testability of students with SEN-L within large-scale assessments. Within this program, the appropriateness of different aspects of testing is systematically investigated in order to identify appropriate testing conditions for students with SEN-L. The analyses reported in this chapter are the basis for further analyses. Südkamp, Pohl, and Weinert (2015), for example, investigated whether different testing accommodations result in reliable and comparable measures of reading competence. Testing accommodations include a reduction in test length as well as a reduction in the test's item difficulty. Further test accommodations draw on a reduction of grammatical and lexical complexity in the texts and items and on a specifically designed test-coaching phase prior to testing. Other research questions motivated by the present study are addressed by Pohl, Südkamp, Hardt, Carstensen, and Weinert (2015). These authors investigated whether there are differences in large-scale testability between students with SEN-L and how these differences are related to individual test-taking behavior.

References

- Abedi, J., Leon, S., Kao, J., Bayley, R., Ewers, N., Herman, J., & Mundhenk, K. (2010). *Accessible reading assessments for students with disabilities: The role of cognitive, grammatical, lexical, and textual/visual features* (CRESST Report 785). Los Angeles, CA: University of California.
- Abrams, L. M., Pedulla, J. J., & Madaus, G. F. (2003). Views from the classroom: Teachers' opinions of statewide testing programs. *Theory Into Practice*, *42*, 18–29. doi:10.1207/s15430421tip4201_4
- Aßmann, C., Steinhauer, H. W., Kiesl, H., Koch, S., Schönberger, B., Müller-Kuller, A., ... Blossfeld, H.-P. (2011). In H.-P. Blossfeld, H.-G. Roßbach, & J. von Maurice (Eds.), *Zeitschrift für Erziehungswissenschaft*, *14*. *Education as a lifelong process: The German National Educational Panel Study (NEPS)* (pp. 51–65). Wiesbaden: VS Verlag für Sozialwissenschaften. doi:10.1007/s11618-011-0181-8
- Aßmann, C., Steinhauer, H. W., & Zinn, S. (2012). *Weighting the fifth and ninth grader cohort samples of the National Educational Panel Study, panel cohorts* (Technical Report). Retrieved from https://www.neps-data.de/Portals/0/NEPS/Datenzentrum/Forschungsdaten/SC3/1-0-0/SC3_SC4_1-0-0_Weighting_EN.pdf
- Barkow, I., Leopold, T., Raab, M., Schiller, D., Wenzig, K., Blossfeld, H.-P., & Rittberger, M. (2011). RemoteNEPS: Data dissemination in a collaborative workspace. In H.-P. Blossfeld, H.-G. Roßbach, & J. von Maurice (Eds.), *Zeitschrift für Erziehungswissenschaft*, *14*. *Education as a lifelong process: The German National Educational Panel Study (NEPS)* (pp. 315–325). Wiesbaden: VS Verlag für Sozialwissenschaften. doi:10.1007/s11618-011-0192-5
- Bielinski, J., Thurlow, M. L., Ysseldyke, J. E., Freidebach, J., & Freidebach, M. (2001). *Read-aloud accommodations: Effects on multiple-choice reading and math items* (NCEO

- Technical Report No. 31). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.
- Blossfeld, H.-P., & von Maurice, J. (2011). Education as a lifelong process. In H.-P. Blossfeld, H.-G. Roßbach, & J. von Maurice (Eds.), *Zeitschrift für Erziehungswissenschaft, 14. Education as a lifelong process: The German National Educational Panel Study (NEPS)* (pp. 19–34). Wiesbaden: VS Verlag für Sozialwissenschaften. doi:10.1007/s11618-011-0179-2
- Blossfeld, H.-P., von Maurice, J., & Schneider, T. (2011). The National Educational Panel Study: Need, main features, and research potential. In H.-P. Blossfeld, H.-G. Roßbach, & J. von Maurice (Eds.), *Zeitschrift für Erziehungswissenschaft, 14. Education as a lifelong process: The German National Educational Panel Study (NEPS)* (pp. 5–17). Wiesbaden: VS Verlag für Sozialwissenschaften. doi:10.1007/s11618-011-0178-3
- Bolt, S. E., & Ysseldyke, J. (2008). Accommodating students with disabilities in large-scale testing: A comparison of differential item functioning (DIF) identified across disability types. *Journal of Psychoeducational Assessment, 26*, 121–138. doi:10.1177/0734282907307703
- Bos, W., Bonsen, M., & Gröhlich, C. (Hrsg.). (2009). *KESS 7: Kompetenzen und Einstellungen von Schülerinnen und Schülern—Jahrgangsstufe 7* [KESS 7: Competencies and attitudes of students in grade 7]. Hamburg: Behörde für Bildung und Sport.
- Dolan, R. P., & Hall, T. E. (2001). Universal design for learning: Implications for large-scale assessment. *IDA Perspectives, 27*(4), 22–25.
- Durkin, D. (1993). *Teaching them to read*. Boston, MA: Allyn and Bacon.
- Gehrer, K., Zimmermann, S., Artelt, C., & Weinert, S. (2012). *The assessment of reading competence (including sample items for Grade 5 and 9)* [Scientific Use File 2012, Version 1.0.0.]. Bamberg: University of Bamberg, National Educational Panel Study.
- Gehrer, K., Zimmermann, S., Artelt, C., & Weinert, S. (2013). NEPS framework for assessing reading competence and results from an adult pilot study. *Journal of Educational Research Online, 5*(2), 50–79.
- Gersten, R., Fuchs, L. S., Williams, J. P., & Baker, S. (2001). Teaching reading comprehension strategies to students with learning disabilities: A review of research. *Review of Educational Research, 71*, 279–320. doi:10.3102/00346543071002279
- Guthrie, J. T. (2002). Preparing students for high-stakes test taking in reading. In A. E. Farstrup, & S. J. Samuels (Eds.), *What research has to say about reading instruction* (pp. 370–391). Newark: International Reading Association.
- Heydrich, J., Weinert, S., Nusser, L., Artelt, C., & Carstensen, C. H. (2013). Including students with special educational needs into large-scale assessments of competencies: Challenges and approaches within the German National Educational Panel Study (NEPS). *Journal of Educational Research Online, 5*(2), 217–240.
- Joncas, M. (2007). PIRLS 2006 sampling design. In M. O. Martin, I. V. S. Mullis, & A. M. Kennedy (Eds.), *PIRLS 2006 technical report* (pp. 35–48). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

- Kavale, K. A., & Reece, J. H. (1992). The character of learning disabilities: An IOWA profile. *Learning Disability Quarterly*, 15, 74–94. doi: 10.2307/1511010
- KMK—Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland [Standing Conference of the Ministers of Education and Cultural Affairs of Germany]. (2012). *Sonderpädagogische Förderung in Schulen 2001–2010* [Special education in schools 2001–2010] (Dokumentation No. 196). Berlin: Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland.
- Koretz, D. M. (1997). *The assessment of students with disabilities in Kentucky* (CSE Technical Report No. 431). Los Angeles, CA: CRESST/RAND Institute on Education and Training.
- Koretz, D. M., & Hamilton, L. (2000). Assessment of students with disabilities in Kentucky: Inclusion, student performance, and validity. *Educational Evaluation and Policy Analysis*, 22, 255–272. doi:10.3102/01623737022003255
- Kristen, C., Edele, A., Kalter, F., Kogan, I., Schulz, B., Stanat, P., & Will, G. (2011). The education of migrants and their children across the life course. In H.-P. Blossfeld, H.-G. Roßbach, & J. von Maurice (Eds.), *Zeitschrift für Erziehungswissenschaft*, 14. *Education as a lifelong process: The German National Educational Panel Study (NEPS)* (pp. 121–137). Wiesbaden: VS Verlag für Sozialwissenschaften. doi:10.1007/s11618-011-0194-3
- Lovett, B. J. (2010). Extended time testing accommodations for students with disabilities: Answers to five fundamental questions. *Review of Educational Research*, 80, 611–638. doi:10.3102/0034654310364063
- Lutkus, A. D., Mazzeo, J., Zhang, J., & Jerry, L. (2004). *Including special-needs students in the NAEP 1998 reading assessment part II: Results for students with disabilities and limited-English proficient students* (Research Report ETS-NAEP 04-R01). Princeton, NJ: ETS.
- OECD. (2010, December). *PISA 2009 results: What students know and can do: Student performance in reading, mathematics and science, 1*. Retrieved from <http://dx.doi.org/10.1787/9789264091450-en>
- OECD. (2012, March). *PISA 2009 Technical Report*. Retrieved from <http://dx.doi.org/10.1787/9789264167872-en>
- Pellegrino, J., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, D. C.: National Academy Press.
- Pitoniak, M. J., & Royer, J. M. (2001). Testing accommodations for examinees with disabilities: A review of psychometric, legal, and social policy issues. *Review of Educational Research*, 71, 53–104. doi:10.3102/00346543071001053
- Pohl, S., & Carstensen, C. H. (2012). *NEPS technical report: Scaling the data of the competence test*. (NEPS Working Paper No. 14). Bamberg: University of Bamberg, National Educational Panel Study.
- Pohl, S., Haberkorn, K., Hardt, K., & Wiegand, E. (2012). *NEPS technical report for reading—Scaling results of Starting Cohort 3 in fifth grade*. (NEPS Working Paper No. 15). Bamberg: University of Bamberg, National Educational Panel Study.

- Pohl, S., Südkamp, A., Hardt, K., Carstensen, C. H., & Weinert, S. (2015). *Testing students with special educational needs—Psychometric properties of test scores and associations with test taking behavior*. Manuscript submitted for publication.
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Nielsen & Lydiche.
- Südkamp, A., Pohl, S., Hardt, K., Duchhardt, C., & Jordan, A.-K. (2015). Kompetenzmessung in den Bereichen Lesen und Mathematik bei Schülerinnen und Schülern mit sonderpädagogischem Förderbedarf [Competence assessment of students with special educational needs in the areas of reading and mathematics]. In P. Kuhl, P. Stanat, B. Lütje-Klose, C. Gresch, H. A. Pant, & M. Prenzel (Eds). *Inklusion von Schülerinnen und Schülern mit sonderpädagogischem Förderbedarf in Schulleistungserhebungen* (S. 243–272). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Südkamp, A., Pohl, S., & Weinert, S. (2015). Competence assessment of students with special educational needs—Identification of appropriate testing accommodations. *Frontline Learning Research*, 3, 1–25. doi:10.14786/flr.v3i2.130
- Swanson, L. (1999). Reading research for students with LD: A meta-analysis of intervention outcomes. *Journal of Learning Disabilities*, 32, 504–532. doi:10.1177/002221949903200605
- Thompson, S. J., Johnstone, C. J., Anderson, M. E., & Miller, N. A. (2005). *Considerations for the development and review of universally designed assessments* (NCEO Technical Report No. 42). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.
- Thurlow, M. L. (2010). Steps toward creating fully accessible reading assessments. *Applied Measurement in Education*, 23, 121–131. doi:10.1080/08957341003673765
- Thurlow, M. L., Bremer, C., & Albus, D. (2008). *Good news and bad news in disaggregated subgroup reporting to the public on 2005–2006 assessment results* (Technical Report No. 52). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.
- Verhoeven, L., & van Leeuwe, J. (2008). Prediction of the development of reading comprehension: A longitudinal study. *Applied Cognitive Psychology*, 22, 407–423. doi:10.1002/acp.1414
- Weinert, S., Artelt, C., Prenzel, M., Senkbeil, M., Ehmke, T., & Carstensen, C. H. (2011). Development of competencies across the life span. In H.-P. Blossfeld, H.-G. Roßbach, & J. von Maurice (Eds.), *Zeitschrift für Erziehungswissenschaft*, 14. *Education as a lifelong process: The German National Educational Panel Study (NEPS)* (pp. 67–86). Wiesbaden: VS Verlag für Sozialwissenschaften. doi:10.1007/s11618-011-0182-7
- Woodcock, S., & Vialle, W. (2011). Are we exacerbating students' learning disabilities? An investigation of pre-service teachers' attributions of the educational outcomes of students with learning disabilities. *Annals of Dyslexia*, 61, 223–241. doi:10.1007/s11881-011-0058-9
- Wu, M., Adams, R. J., Wilson, M., & Haldane, S. (2007). *Conquest 2.0*. [Computer Software]. Camberwell: ACER Press.

- Wu, Y.-C., Liu, K.K., Thurlow, M.L., Lazarus, S.S., Altman, J., & Christian, E. (2012). *Characteristics of low performing special education and non-special education students on large-scale assessments* (Technical Report No. 60). Minneapolis, MN: University of Minnesota, National Centre on Educational Outcomes.
- Ysseldyke, J.E., Thurlow, M.L., Langenfeld, K.L., Nelson, R.J., Teelucksingh, E., & Seyfarth, A. (1998). *Educational results for students with disabilities: What do the data tell us?* (Technical Report No. 23). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

About the authors

A. Südkamp
Rehabilitation Psychology, TU Dortmund University,
Emil-Figge Str. 50, 44227 Dortmund, Germany.
e-mail: anna.suedkamp@tu-dortmund.de

S. Pohl
Methods and Evaluation/Quality Assurance, Free University Berlin,
Habelschwerdter Allee 45, 14195 Berlin, Germany.

J. Heydrich
Formerly University of Bamberg,
Wilhelmsplatz 3, 96045 Bamberg, Germany.

S. Weinert
Department of Psychology I: Developmental Psychology,
University of Bamberg,
Markusplatz 3, 96047 Bamberg, Germany.

Estimation of Plausible Values Considering Partially Missing Background Information: A Data Augmented MCMC Approach

Christian Aßmann, Christoph Gaasch, Steffi Pohl and Claus Carstensen

Abstract

The National Educational Panel Study (NEPS) provides data on the development of competencies across the whole life span. Plausible values as a measure of individual competence are provided by explicitly including background variables that capture individual characteristics in the corresponding Item Response Theory model. Despite tremendous efforts in field work, missing values in the background variables can occur. Adequate estimation routines are needed to reflect the uncertainty stemming from missing values in the background variables with regard to plausible values. We propose an estimation strategy based on Markov Chain Monte Carlo techniques that simultaneously addresses missing values in background variables and estimates parameters characterizing the distribution of plausible values. We evaluate the validity of our approach with respect to statistical accuracy in a simulation study that allows for controlling the mechanism that causes missing data. The results show that the proposed approach is capable of recovering the true regression parameters that describe the relationship between latent competence scores and background variables and thus of recovering the distribution that characterizes plausible values. The approach is illustrated in an example using competence test data on mathematical abilities of Grade-5 students.

1 Introduction

In large-scale studies, such as the National Educational Panel Study (NEPS), an aim is to provide educational researchers with data that support the investigation of various educational research questions. Typical research questions concern, for example, the explanation of competencies and their development over the life course based on individual characteristics. These characteristics could include respondents' gender, so-

cio-economic status, migration experience, or context variables such as school environment. Competencies are assessed in the NEPS via test items in different domains, such as mathematics, reading, and science (Weinert et al., 2011), which are commonly analyzed via the Item Response Theory (IRT) modeling framework. In IRT models, the response to test items is described by a function of the respondents' ability to solve a specific task as well as by the properties of the item. IRT models allow for the aggregation of an individual response pattern towards latent competence scores. Typically, competence scores are provided in the form of plausible values (Mislevy, 1991). As such, plausible values may be used to investigate the relationship between latent competence scores and these background variables. However, despite tremendous efforts in field work, missing values in these background variables might occur, which poses a great challenge to the estimation of parameters characterizing the distribution of plausible values. Missing values in background variables are usually treated via multiple imputation (Rubin, 1987), including relevant variables used in later analyses in the imputation model. However, as the relationship of latent competence scores with background variables are of interest, latent competence scores need to be included as covariates in the imputation model. On the other hand, for the estimation of latent competence scores, background variables (that show missing values) are needed in the measurement model. Furthermore, it is necessary to account for uncertainty due to measurement error in competence data as well as due to the imputation of missing values. We propose using a fully Bayesian approach based on the device of data augmentation (Tanner & Wong, 1987) to deal with this challenge. In the following section, we first introduce the IRT model used for scaling the competence data. We then focus on the problem of missing responses in the background data and introduce the Bayesian approach using Markov Chain Monte Carlo (MCMC) methods. We develop a hybrid estimation algorithm that simultaneously estimates parameters that characterize the distribution of plausible values and imputes missing responses in background variables. The approach is evaluated within a simulation study and demonstrated in a small empirical example that measures one-dimensional mathematical competence using NEPS data.

2 IRT Model for Scaling of Competence Tests

Different competence domains are assessed in the NEPS, including mathematics, reading, and science. The competence domains are assessed with test booklets that contain domain-specific item blocks with varying response formats. Single items may be dichotomously scored, (i. e., a respondent is able to solve an item or not) or consist of a couple of dichotomously scored tasks (complex multiple choice items). For IRT analysis, the complex multiple-choice items are aggregated to a single polytomous super-item (see e. g., Andrich, 1985), which indicates the number of correct responses to the subitems.

The competence data in the NEPS are scaled using the multidimensional random coefficient multinomial logit model (RCMLM) (Adams, Wilson, & Wang, 1997a) (for a description of the scaling model for the competence data in the NEPS, see Pohl & Carstensen, 2012). The RCMLM model encompasses the simple Rasch model (Rasch, 1960) for dichotomous data and its extension towards ordered polytomous data, namely the Partial Credit Model (Masters, 1982), for special cases (see also Adams and Wilson, 1996). Here, Y denotes the data matrix containing responses from $i = 1, \dots, N$ test takers to $j = 1, \dots, J$ items, each having up to M_j categories. Considering unidimensional scaling within the class of RCMLM models, the probability of a response being in category m of item j for individual i (y_{ijm}) is given by

$$P(y_{ijm} = 1|\theta_i) = \frac{\exp(b_{jm}\theta_i + \alpha'_{jm}\xi)}{\sum_{m=1}^{M_j} \exp(b_{jm}\theta_i + \alpha'_{jm}\xi)}, \quad i = 1, \dots, N, j = 1, \dots, J, m = 1, \dots, M_j \tag{1}$$

where θ_i is the scalar ability parameter of person i ,

$$b = (b_{11}, b_{12}, \dots, b_{1M_1}, b_{21}, b_{22}, \dots, b_{2M_2}, \dots, b_{J1}, \dots, b_{JM_J}) \tag{2}$$

is a vector of scoring functions with b_{jm} reflecting the performance level of each possible item category, $\zeta = (\zeta_1, \zeta_2, \dots, \zeta_p)$ is a vector of p item difficulty parameters, and

$$A = (a_{11}, a_{12}, \dots, a_{1M_1}, a_{21}, \dots, a_{2M_2}, \dots, a_{J1}, \dots, a_{JM_J}) \tag{3}$$

is a design matrix of design vectors a_{jm} , each of length p , giving scoring weights of the item categories. Within the class of RCMLM models, θ_i is regarded as a random parameter with density function $g(\theta_i)$ for all i . Typically, this density is assumed to be normal with mean μ and variance σ^2 . Adams, Wilson, and Wu (1997b) provided an interpretation of the RCMLM as a multilevel model with test takers as Level-2 units and test responses as Level-1 observations. Additionally, this model can be extended towards a structural analysis of latent competencies via the inclusion of empirical predictor variables, that is, background variables. The distribution of θ_i is then assumed to depend on X_i , that is, it takes the form $g(\theta_i|X_i)$, typically normal with mean $X_i\gamma$ and variance σ^2 , where X_i denotes a vector of individual background variables explaining differences in achievement. Extensions of this model are available in Adams et al. (1997a). It is important to note that parameter estimation in these models is routinely performed via the marginal maximum likelihood method (for details, see Adams et al., 1997b) or by using weighted likelihood estimation (Warm, 1989).

Alternatively, parameter estimates in a Bayesian framework (see Fox and Glas, 2001; Edwards, 2010) can be obtained as expected values of the posterior distributions. When multiple draws from the posterior distribution of θ_i conditional on the item responses, item parameters, and background variables are provided, this is referred to as the concept of plausible values (Mislevy, 1991). Plausible values are nowa-

days considered to be a state-of-the-art method (OECD, 2009) for characterizing the properties of estimated competencies. It is important to note that the inclusion of certain background variables for estimating plausible values enhances the power of secondary analysis with the same background variables, for which Adams et al. (1997b) include individual characteristics as Level-2 predictors.

The provision of plausible values becomes non-trivial when missing values occur in the background variables. Missing values in questionnaire items are routinely treated via multiple imputation (Rubin, 1987). Since released Scientific Use File data are used for a variety of research questions in the NEPS context that are not known at the time of data release, providing appropriate data for all these analyses is a great challenge. In any case, an appropriate model is required that includes all relevant background variables intended to be considered in secondary analyses. Specifically, questionnaire variables enhance the provision of plausible values, and latent competence scores should be considered in the imputation of missing information in questionnaire variables. Other large-scale studies, such as the Programme for International Student Assessment (PISA) and the National Assessment of Educational Progress (NAEP), deal with this problem by aggregating the questionnaire variables to orthogonal factors and using a set of factors (as many as needed to explain 90 % of the variance of the questionnaire items) as background variables in the IRT measurement model of the competence data (Allen, Carson, Johnson, & Mislevy, 2001; OECD, 2009, 2012). However, this approach is a two-step approach that does not incorporate all questionnaire variables and does not depict the uncertainty stemming from missing values in questionnaire items.

In the following section, we describe a data-analysis and parameter-estimation strategy that applies to the multilevel RCMLM for univariate competence measurement settings. The proposed estimation strategy is designed to cope with missing information for individual-level variables influencing person abilities. To ensure the validity of empirical competence estimates given the uncertainty stemming from missing values in background variables, we adopt a Bayesian estimation scheme that allows for a conceptually stringent treatment of missing values in observed individual characteristics via the device of data augmentation. Bayesian estimation is implemented using MCMC techniques, namely Gibbs sampling, which are ideally suited to deal with the hierarchical structure of the model and the incorporation of a missing data imputation step. In addition, the usage of MCMC simulation methods proves straightforward for complex IRT models relative to marginal maximum likelihood (Patz & Junker, 1999). To illustrate the approach, we restrict the distribution of missing values to the normal distribution, for which nonparametric distributions would provide a valid and highly flexible alternative.

3 Bayesian Inference Using MCMC Techniques

Summarizing all model parameters as ψ and letting S denote the observed sample data, Bayesian inference is concerned with the posterior distribution $p(\psi|S)$ and moments corresponding to it. A general introduction to the basic principles employed in the following section is provided by Geweke (1999) and Koop (2003). Gibbs sampling is a device used to produce a sample from the joint posterior distribution of the parameter vector ψ , which can be used to estimate posterior moments and density estimates. Posterior draws of ψ partitioned into convenient blocks $\psi = \{\psi_1, \dots, \psi_T\}$ are obtained via Gibbs sampling when direct sampling from the posterior distribution is difficult but sampling from the full conditional distributions is directly accessible. The functional forms of the full conditional distributions can be deduced from the joint posterior distribution of parameters ψ and the sample data S ,

$$p(\psi, S) = L(S|\psi)\pi(\psi), \tag{4}$$

where $L(S|\psi)$ denotes the model likelihood and $\pi(\psi)$ denotes the a priori distribution, via isolating the kernel of a single parameter block ψ_t conditional on all other blocks $\psi_1, \dots, \psi_{t-1}, \psi_{t+1}, \dots, \psi_T$, and the data S . Given an initialization $\psi^{(0)}$, the Gibbs sampling algorithm simulates iteratively for $r = 1, \dots, R$ from the full conditional distributions

$$p(\psi_1^{(r)} | \psi_2^{(r-1)}, \dots, \psi_T^{(r-1)}, S), \tag{5}$$

$$p(\psi_2^{(r)} | \psi_1^{(r)}, \psi_3^{(r-1)}, \dots, \psi_T^{(r-1)}, S), \tag{6}$$

...

$$p(\psi_T^{(r)} | \psi_1^{(r)}, \dots, \psi_{T-1}^{(r)}, S). \tag{7}$$

The iterative sampling constitutes a Markov Chain, which ensures convergence to the joint posterior distribution under the general regularity conditions given in Roberts and Smith (1994). Since these conditions are fulfilled within the considered class of RCMLM models, the convergence of the joint distribution of the sample $\{\psi^{(r)}\}_{r=1}^R$ for $R \rightarrow \infty$ towards the posterior distribution $p(\psi|S)$ is ensured. Since the functional forms of the full conditional distributions depend on the assumed prior distributions, these are generally conveniently chosen to facilitate sampling from closed-form full conditional distributions.

4 Estimation Algorithm for Binary Test Data Considering Partially Missing Background Information

To illustrate the proposed treatment of missing values within the class of RCMLM models, we refer to a simplified version of the model outlined in Equation (1). This simplified version, which allows for closed-form sampling from the full conditional distributions employed within the Gibbs sampler, is derived as follows (for a general treatment of Bayesian estimation for binary panel probit models, see also Aßmann & Boysen-Hogrefe, 2011): We consider the likelihood conditional on latent individual abilities of the model stated in Equation (1) given as

$$L(Y|\{\theta_{ij}\}_{i=1}^N, A, b, \xi) = \prod_{i=1}^N \prod_{j=1}^J \prod_{m=1}^{M_j} \left(\frac{\exp(b_{jm}\theta_i + \alpha'_{jm}\xi)}{\sum_{m=1}^{M_j} \exp(b_{jm}\theta_i + \alpha'_{jm}\xi)} \right)^{y_{ijm}} \tag{8}$$

Setting $M_j = 2$ (i. e., considering only dichotomous items), assuming $\alpha'_{jm}\xi = \xi_j$ to be known with $\xi_1 = 0$, normalizing $b_{j1} = 0$ and $b_{j2} = 1$, and changing notation into $y_{ij1} = 1 - y_{ij2} = y_{ij}$ results in the conditional likelihood

$$L(Y|\{\theta_{ij}\}_{i=1}^N, \xi) = \prod_{i=1}^N \prod_{j=1}^J \left(\frac{\exp(\theta_i - \xi_j)^{y_{ij}}}{1 + \exp(\theta_i - \xi_j)} \right). \tag{9}$$

In conjunction with a mixing distribution $g(\theta_i|X_i)$ given as normal, that is,

$$g(\theta_i|X_i) = (2\pi)^{-.5} (\sigma^2)^{-.5} \exp\left(-\frac{1}{2\sigma^2}(\theta_i - X_i\gamma)^2\right), \tag{10}$$

this allows for a derivation of the likelihood

$$L(Y|\gamma, \sigma^2, \xi) = \prod_{i=1}^N \int \prod_{j=1}^J \left(\frac{\exp(\theta_i - \xi_j)^{y_{ij}}}{1 + \exp(\theta_i - \xi_j)} \right) g(\theta_i|X_i) d\theta_i. \tag{11}$$

Since no conjugate priors for parameters of this likelihood exist that facilitate either direct sampling or closed-form sampling from the corresponding full conditional distributions, we further change from logit to probit. This allows for Bayesian estimation via Gibbs sampling along the lines suggested by Albert (1992). The likelihood is then given as

$$L(Y|\gamma, \sigma^2, \xi) = \prod_{i=1}^N \int \prod_{j=1}^J \Phi((2y_{ij} - 1)(\theta_i - \xi_j)) g(\theta_i|X_i) d\theta_i, \tag{12}$$

where $\Phi(\cdot)$ denotes the standard normal cumulative distribution function and allows for derivation of the corresponding full conditional distributions. In addition to itera-

tive sampling from the set of full conditional distributions facilitating parameter estimation, the missing values in background variables are augmented into the parameter vector. To assess the uncertainty of missing values and because the analytical IRT model does not include information concerning the full conditional distribution of the missing values, an ad hoc modelling is adapted based on a linear regression with normal errors. This results in a hybrid sampling scheme that allows for parameter estimation in case of missing values in the background variables. After initializing parameters, this leads to the following iterative scheme for sampling from the set of full conditional distributions.

Step 1) The underlying latent variable y_{ij}^* is sampled from a truncated normal distribution with corresponding parameters

$$\mu_{y_{ij}^*} = \theta_i - \xi_j, \quad \text{and} \quad \sigma_{y_{ij}^*}^2 = 1,$$

where the truncation sphere is $(-\infty, 0)$ for $y_{ij} = 0$ and $(0, \infty)$ for $y_{ij} = 1$.

Step 2) The individual abilities θ_i are sampled from a normal distribution, with moments defined by

$$\mu_{\theta_i} = \left(J + \frac{1}{\sigma^2} \right)^{-1} \left(\sum_{j=1}^J y_{ij}^* + \sum_{j=1}^J \xi_j + X_i \gamma / \sigma^2 \right), \quad \text{and} \quad \sigma_{\theta_i}^2 = \left(J + 1 / \sigma^2 \right)^{-1}.$$

Step 3) The independent conjugate prior for γ is allowed to be multivariate normal, with moments mean vector v_γ and covariance matrix Ω_γ . Draws from the full conditional distribution for γ are then obtained from a multivariate normal distribution, with corresponding moments given as

$$\mu_\gamma = (X'X/\sigma^2 + \Omega_\gamma^{-1})^{-1}(X'\theta/\sigma^2 + \Omega_\gamma^{-1}v_\gamma), \quad \text{and} \quad \Sigma_\gamma = (X'X/\sigma^2 + \Omega_\gamma^{-1})^{-1},$$

where X denotes the $N \times K$ matrix of background variables and $\theta = (\theta_1 \dots \theta_N)$ the $N \times 1$ vector of latent abilities.

Step 4) The independent conjugate prior for σ^2 inverse gamma with parameters α_0 and β_0 is chosen, and the σ^2 is also distributed inverse gamma with corresponding parameter

$$\alpha = N/2 + \alpha_0, \quad \text{and} \quad \beta = (0.5 \sum_{i=1}^N (\theta_i - X_i \gamma)^2 + \beta_0)^{-1}.$$

Step 5) Item nonresponse is imputed in the $N \times K$ matrix of background variables X by specifying a univariate normal full conditional distribution for each of the K vari-

ables contained in X . Within the Gibbs sampler, imputed and hence complete variables are at hand for each iteration r , resulting in the following K regression equations, given as

$$X_k = W_k \varphi_k + \varepsilon_k, \quad k = 1, \dots, K,$$

where $W_k = (X_{-k}, \theta)$, and X_{-k} also subsumes a constant. Each missing value in X_k is replaced via a draw from a univariate normal distribution with moments $\mu = W'_{mis} \hat{\varphi}$ and $\sigma^2 = \hat{\sigma}_\varepsilon^2$. It is important to note that instead of the least squares estimators $\hat{\varphi}$ and $\hat{\sigma}_\varepsilon^2$, draws from the corresponding asymptotic distributions are used for generating draws for the missing values in X . However, as missing values are filled in within each iteration of the Gibbs sampler, the corresponding uncertainty is accounted for. Furthermore, it should be explicitly noted that the estimation scheme introduces the updated draws of the individual abilities θ into the imputation model for each iteration.

The sampler given here assumes knowledge of the item difficulty parameters. Simultaneous estimation of item difficulties is a straightforward extension of the outlined approach. Given a sample of all model parameters obtained via iterative sequential cycling through the set of full conditional distributions, the plausible values for each individual can be directly taken from the provided Gibbs output presupposing that the effect of initialization has been accounted for via discarding a reasonable burn-in phase. Each $\theta^{(r)}$, $r = 1, \dots, R$ could be taken as a vector of plausible values.

5 Simulation Study

To assess the validity of our approach suggested above, we set up a simulation design comparing the data augmented Gibbs sampler when missing values occur with the full sample estimates before deletion. Given this benchmark situation, the relative performance of recovering a set of given parameters in the presence of missing values in the background variables can be evaluated. Replication analysis is a method commonly used for this purpose. The estimation procedure is conducted for $C = 1,000$ replications of a single data-generating process and two missing generating processes. Then, the root mean square error and the proportion of 95 % highest posterior density regions that contain the true parameter values (coverage) are computed as the main criteria for comparison. The detailed conditions of the data generation and missing value-generating processes are as follows:

For each replication $c = 1, \dots, C$, the binary response pattern is simulated using the model in (12) with a sample setup of $N = 1,000$ individuals facing $J = 10$ competence items, for which the item difficulties are specified as draws from a normal distribution, that is, $\xi_j \sim N(0, 0.5)$. Three background variables X explaining differences in individual abilities θ_i are generated from a standard normal distribution and display a correlation of 0.5. The variable X_1 is transformed into a binary variable that takes 1 if

Table 1 Mean Posterior Means and Mean Standard Deviations of $C = 1,000$ Replications for a Data Set Before Deletion (BD), Missing Scenario I and Missing Scenario II for Data Augmented (DA) Estimation Strategy

	True	Mean			Sd		
		BD	DA – I	DA – II	BD	DA – I	DA – II
γ_1	1.000	1.002	1.002	1.001	0.066	0.066	0.067
γ_2	-0.500	-0.504	-0.504	-0.503	0.097	0.097	0.099
γ_3	0.500	0.500	0.500	0.501	0.052	0.053	0.056
γ_4	-0.500	-0.500	-0.500	-0.502	0.052	0.054	0.057
σ^2	1.440	1.449	1.448	1.446	0.098	0.099	0.100

the original value exceeds 0. Then, observations in X_2 and X_3 are deleted via a missing process according to two different Scenarios I and II. In Scenario I, on average, 5% and 10% of missing values result completely at random for the variables X_2 and X_3 . In Scenario II, these rates of missingness increase to 10% and 25% and depend on the values of X_1 . The regression weights of the background variables including an intercept take on the values $\gamma = (1, -0.5, 0.5, -0.5)$, while the individual abilities are distributed with variance parameter $\sigma^2 = 1.2^2$. We use $\nu_y = 0$ and $\Omega_y = 100I_{k+1}$, where I_{k+1} denotes a unity matrix of dimension $K + 1$, $\alpha_0 = 3/10$, and $\beta_0 = 10/3$.

Table 1 shows the means of the posterior expected values and their standard deviations over $C = 1,000$ replications. For both missingness Scenarios I and II, our approach reveals an unbiased estimation of all parameters. Furthermore, with respect to the error and the coverage rate depicted in Table 2, the findings also reveal that there is no notable difference in the full sample estimates before deletion reported in the first block columns of the table (BD). The observed number of intervals covering the particular parameter corresponds approximately to the theoretical values. Thus, our proposed sampler is a suitable solution for the use of partially missing background variables in the context of IRT models, even when a relatively large amount of missing values is present.

Table 2 Root Mean Square Error and Coverage of $C = 1,000$ Replications for a Data Set Before Deletion (BD), Missing Scenario I and Missing Scenario II for Data Augmented (DA) Estimation Strategy

	True	RMSE			Coverage		
		BD	DA – I	DA – II	BD	DA – I	DA – II
Y_1	1.000	0.067	0.068	0.069	0.945	0.949	0.944
Y_2	-0.500	0.097	0.098	0.099	0.957	0.959	0.954
Y_3	0.500	0.053	0.055	0.057	0.943	0.946	0.938
Y_4	-0.500	0.049	0.052	0.057	0.968	0.961	0.950
σ^2	1.440	0.099	0.099	0.101	0.944	0.948	0.945

6 Measuring Mathematical Competence in NEPS Starting Cohort Grade 5

To further illustrate the usefulness of our approach, we apply the structural IRT model to an exemplary research question. We use data from the NEPS Starting Cohort 3–5th Grade, doi:10.5157/NEPS:SC3:2.0.0,¹ on mathematical competence of students in the fifth grade (Blossfeld, Roßbach, & von Maurice, 2011) (for a description of the assessment of mathematical competence in NEPS, see Neumann et al., 2012; for a description of the respective competence data, see Duchardt & Gerdes, 2012; for the data manual, see Skopek, Pink, & Bela, 2012). The data used in this analysis contains information on $N = 5,129$ students who had a valid response to at least three of $J = 23$ binary mathematics test items. Missing values in the test item set were ignored (for a comparison of different approaches for treating missing responses in competence tests, see Pohl et al., 2014). In addition to the test results, we consider two binary variables representing the mathematics test position within the booklet (Position) and the gender of the test takers (Female), as well as self-concept beliefs in mathematical skills (Self-Concept), and satisfaction with school (Schoolsat) as explanatory variables for the analysis. Descriptive statistics for the data considered in the application are displayed in Table 3. Both quantitative variables of self-concept beliefs and satisfaction with schools were measured with a single item with 4 and 11 response options, respectively. High values on these variables indicate a high level of self-concept

1 From 2008 to 2013, NEPS data were collected as part of the Framework Programme for the Promotion of Empirical Educational Research funded by the German Federal Ministry of Education and Research (BMBF). As of 2014, the NEPS survey is carried out by the Leibniz Institute for Educational Trajectories (LIfBi) at the University of Bamberg in cooperation with a nationwide network.

Table 3 Descriptive Statistics: Background Variables

Variable	Min	Max	Mean	Sd	Missing
Position	0	1	0.50	–	0.00
Female	0	1	0.48	–	0.00
Self-Concept	1	4	2.94	0.85	0.06
Schoolsat	1	11	8.72	2.52	0.03

Notes: $N = 5,129$.

and satisfaction. With 6 % and 3 % missing values in the background variables of self-concept beliefs and satisfaction with schools, respectively, the amount of missing data is relatively small.

We applied the proposed data augmented Gibbs sampling approach to estimate the regression coefficients of test position, gender, mathematical self-concept, and school satisfaction with the latent mathematics score. In a next step, we additionally generated one third of each variable X_2 (Self-Concept) and X_3 (Schoolsat) to show missing values completely at random. The data augmented Gibbs sampling approach is able to deal with the missing values in the two background variables while simultaneously estimating plausible values for mathematical competence. The item parameters were estimated beforehand using a conditional likelihood estimation of the simple Rasch model. For each of the considered scenarios, the algorithm showed a good convergence behavior when we assumed the same prior distributions as in the simulation study. As an example, the trace plots for the data augmented approach without additional missings show no indication of convergence problems (Figure 1), the autocorrelations become very low (Figure 2), and the cumulative means converge (Figure 3) with similar findings for the other scenarios. Using a burn-in period of 2,000 draws, the parameter estimates were based on $R = 8,000$ simulated draws. Table 4 depicts the estimated posterior means and standard deviations as well as the 95 % Highest Density Regions (HDR) for the data augmented cases, the complete cases, and the complete cases with additional missings induced at random. The latter is considered a benchmark for our approach to illustrate the efficiency gains in parameter estimation using the data augmented approach. No substantial differences are revealed between the data augmented case analysis and the complete case analysis for both missing scenarios. However, clear efficiency gains in terms of standard deviation and highest density regions of the parameter estimates are documented for the data augmented approach. While the results indicate a lower level of competence for females, both quantitative variables have a positive effect on student mathematical skills. There is no evidence for a significant position effect because the 95 % HDR of the corresponding posterior contains the value of zero. The estimated standard errors of the regression coefficients

Table 4 Parameter Estimates for NEPS Starting Cohort Grade 5: Mathematics Test Data

Variable	Mean	Sd	95 % HDR
Data augmented (N = 5,129)			
Constant (γ_1)	-0.590	0.055	[-0.698; -0.482]
Position (γ_2)	0.041	0.021	[-0.001; 0.083]
Female (γ_3)	-0.135	0.023	[-0.180; -0.090]
Self-Concept (γ_4)	0.241	0.014	[0.215; 0.268]
Schoolsat (γ_5)	0.030	0.005	[0.021; 0.039]
Variance (σ^2)	0.493	0.013	[0.468; 0.518]
Complete cases (N = 4,675)			
Constant (γ_1)	-0.541	0.057	[-0.652; -0.428]
Position (γ_2)	0.037	0.023	[-0.008; 0.081]
Female (γ_3)	-0.144	0.024	[-0.180; -0.098]
Self-Concept (γ_4)	0.236	0.014	[0.208; 0.264]
Schoolsat (γ_5)	0.030	0.005	[0.020; 0.039]
Variance (σ^2)	0.489	0.013	[0.464; 0.515]
Data augmented with additional missings (N = 5,129)			
Constant (γ_1)	-0.583	0.061	[-0.705; -0.465]
Position (γ_2)	0.041	0.022	[-0.002; 0.084]
Female (γ_3)	-0.131	0.024	[-0.177; -0.085]
Self-Concept (γ_4)	0.255	0.016	[0.223; 0.286]
Schoolsat (γ_5)	0.024	0.005	[0.013; 0.034]
Variance (σ^2)	0.491	0.013	[0.467; 0.517]
Complete cases with additional missings (N = 2,135)			
Constant (γ_1)	-0.444	0.086	[-0.613; -0.277]
Position (γ_2)	0.030	0.034	[-0.037; 0.098]
Female (γ_3)	-0.184	0.036	[-0.254; -0.114]
Self-Concept (γ_4)	0.208	0.022	[0.166; 0.251]
Schoolsat (γ_5)	0.032	0.007	[0.018; 0.046]
Variance (σ^2)	0.506	0.020	[0.467; 0.547]

incorporate not only uncertainty due to person sampling, but also uncertainty due to missing values in the predictors.

7 Conclusion

In large-scale assessments, researchers are usually interested in explaining competence scores with individual characteristics and context variables. Simultaneously accounting for measurement error in competence scores and missing values in background variables that capture individual characteristics and context variables is challenging. We propose a data augmented MCMC approach that simultaneously estimates plausible values and accounts for missing values in background variables. With this approach, latent relationships between competence scores and background variables that efficiently incorporate the uncertainty stemming from only partially observed background variables may be estimated. In a simulation study, the proposed approach proved to adequately recover the model parameters to be estimated, even when higher rates of missingness occurred in the data. The applicability to educational research data has been illustrated in an empirical example. The iterative use of updated parameter values from posterior sampling for the imputation model showed an appealing feature of our approach. Future research should focus on the considerations of an alternative imputation step that copes with the often-categorical character of background variables in research questions involving a larger set of variables.

References

- Adams, R. J., & Wilson, M. R. (1996). Formulating the Rasch model as a mixed coefficients multinomial logit. In G. Engelhardt, & M. Wilson (Eds.), *Objective measurement: Theory into practice* (Vol. 3, pp. 143–166). Norwood, NJ: Ablex.
- Adams, R. J., Wilson, M. R., & Wang, W. C. (1997a). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, 21(1), 1–23.
- Adams, R. J., Wilson, M. R., & Wu, M. (1997b). Multilevel item response models: An approach to errors in variables regression. *Journal of Educational and Behavioral Statistics*, 22(1), 47–76.
- Albert, J. H. (1992). Bayesian estimation of normal ogive item response curves using gibbs sampling. *Journal of Educational Statistics*, 17(3), 251–269.
- Allen, N. L., Carson, J. E., Johnson, E. G., & Mislvey, R. J. (2001). Scaling procedures. In N. L. Allen, J. R. Donoghue, & T. L. Schoeps (Eds.), *The NAEP 1998 technical report*. Washington, DC: U. S. Department of Education.
- Andrich, D. (1985). A latent trait model for items with response dependencies: Implications for test construction and analysis. In S. E. Embretson (Ed.), *Test design—Developments in psychology and psychometrics* (pp. 245–275). Orlando: Academic Press.

- Aßmann, C., & Boysen-Hogrefe, J. (2011). A Bayesian approach to model-based clustering for binary panel probit models. *Computational Statistics & Data Analysis*, 55(1), 261–279.
- Blossfeld, H.-P., Roßbach, H.-G., & von Maurice, J. (Eds.). (2011). Education as a life-long process: The German National Educational Panel Study (NEPS) [Special issue]. *Zeitschrift für Erziehungswissenschaft*, 14. Wiesbaden: VS Verlag für Sozialwissenschaften.
- Chib, S. (2001). Markov chain monte carlo methods: Computation and inference. In J. J. Heckmann, & E. Leamer (Eds.), *Handbook of econometrics* (Vol. 5, pp. 3569–3649). Amsterdam, Netherlands: North Holland.
- Duchhardt, C., & Gerdes, A. (2012). *NEPS technical report for mathematics—Scaling results of Starting Cohort 3 in fifth grade*. (NEPS Working Paper No. 19). Bamberg: University of Bamberg, National Educational Panel Study.
- Edwards, M. C. (2010). A Markov chain Monte Carlo approach to confirmatory item factor analysis. *Psychometrika*, 75(3), 474–497.
- Fox, J.-P., & Glas Cees A. W. (2001). Bayesian estimation of a multilevel IRT model using gibbs sampling. *Psychometrika*, 66(2), 271–288.
- Geweke, J. F. (1999). Using simulation methods for bayesian econometric models: Inference, development and communication. *Econometric Reviews*, 18(1), 1–73.
- Koop, G. (2003). *Bayesian econometrics*. Hoboken, NJ: Wiley.
- Masters, G. N. (1982). A rasch model for partial credit scoring. *Psychometrika*, 47(2), 149–174.
- Mislevy, R. J. (1991). Randomization-based inference about latent variables for complex samples. *Psychometrika*, 56(2), 177–196.
- Neumann, I., Duchhardt, C., Ehmke, T., Grüßing, M., Heinze, A., & Knopp, E. (2012). *Modeling and assessing of mathematical competence over the lifespan*. Manuscript submitted for publication.
- OECD. (2009). *PISA 2006 technical report* (Report No. 56393 2009). Paris: OECD Publishing.
- OECD. (2012). *PISA 2009 technical report* (Report No. 59805 2012). Paris: OECD Publishing.
- Patz, R. J., & Junker, B. W. (1999). A straightforward approach to markov chain monte carlo methods for item response models. *Journal of Educational and Behavioral Statistics*, 24(2), 146–178.
- Pohl, S., Gräfe, L., & Rose, N. (2014). Dealing with omitted and not-reached items in competence tests. Evaluating approaches accounting for missing responses in item response theory models. *Educational and Psychological Measurement*, 74(3), 1–30.
- Pohl, S., & Carstensen, C. (2012). *NEPS technical report—Scaling the data of the competence tests*. (NEPS Working Paper No. 14). Bamberg: University of Bamberg, National Educational Panel Study.
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.

- Roberts, G., & Smith, A. (1994). Simple conditions for the convergence of the Gibbs sampler and Metropolis-Hastings algorithms. *Stochastic Processes and its Applications*, 49(2), 207–216.
- Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. Hoboken, NJ: Wiley.
- Skopek, J., Pink, S., & Bela, D. (2012). *Data manual. Starting Cohort 3—From lower to upper secondary school. NEPS SC3 1.0.0.* (NEPS Research Data Paper). Bamberg: University of Bamberg, National Educational Panel Study.
- Tanner, M. A., & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82(398), 528–540.
- Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, 54(3), 427–450.
- Weinert, S., Artelt, C., Prenzel, M., Senkbeil, M., Ehmke, T., & Carstensen, C. H. (2011). Development of competencies across the life span. In H.-P. Blossfeld, H.-G. Roßbach, & J. von Maurice (Eds.), *Zeitschrift für Erziehungswissenschaft*, 14. *Education as a lifelong process: The German National Educational Panel Study (NEPS)* (pp. 67–86). Wiesbaden: VS Verlag für Sozialwissenschaften.

Appendix A: Figures

Figure 1 Trace plots for the regression constant (γ_1), the regression coefficients for test position (γ_2), gender (γ_3), mathematical self-concept (γ_4), and school satisfaction (γ_5), as well as the residual variance (σ^2)

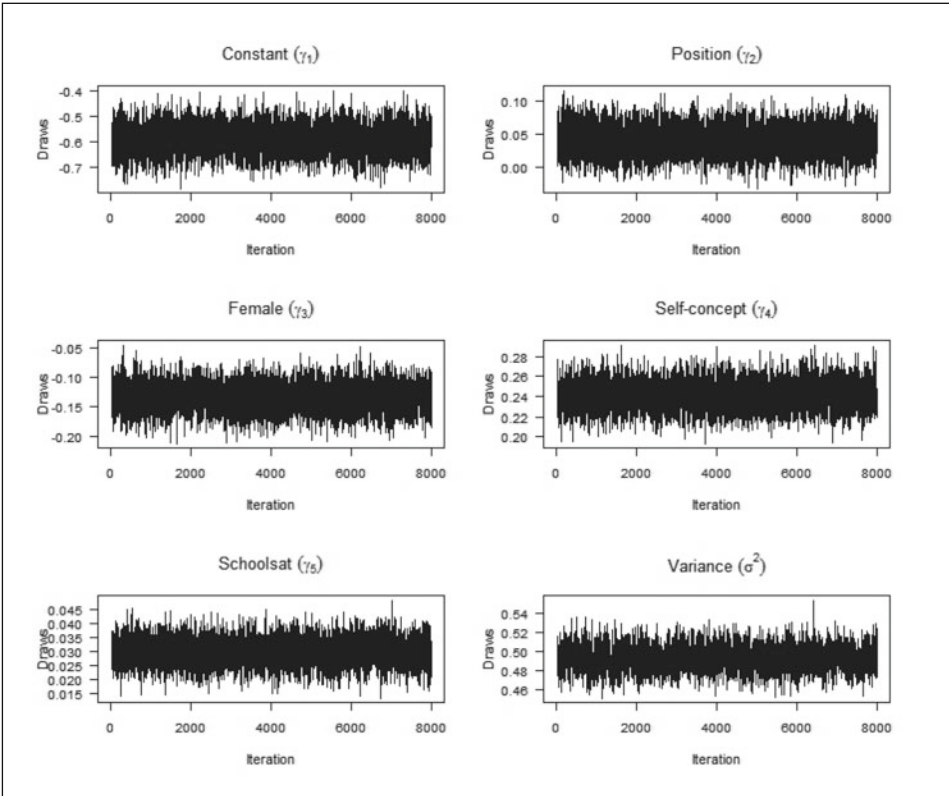


Figure 2 Autocorrelation functions for the regression constant (γ_1), the regression coefficients for test position (γ_2), gender (γ_3), mathematical self-concept (γ_4), and school satisfaction (γ_5), as well as the residual variance (σ^2)

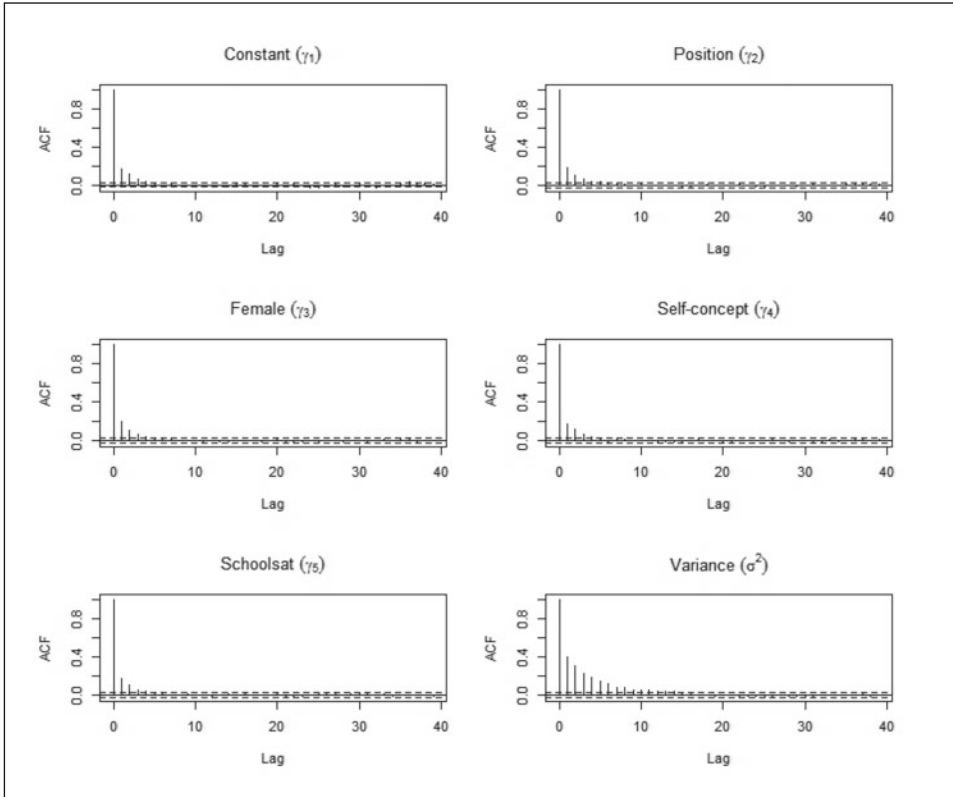
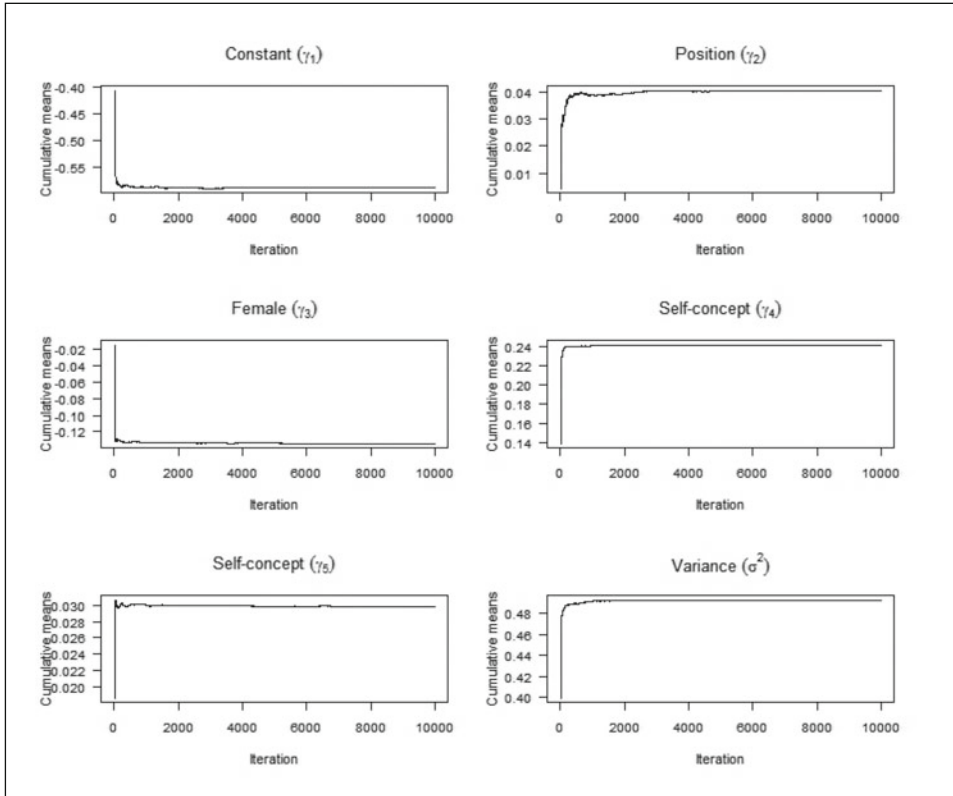


Figure 3 Cumulative mean functions for the regression constant (γ_1), the regression coefficients for test position (γ_2), gender (γ_3), mathematical self-concept (γ_4), and school satisfaction (γ_5), as well as the residual variance (σ^2)



About the authors

C. Aßmann

Leibniz Institute for Educational Trajectories (LifBi), Bamberg.
Chair of Statistics and Econometrics, University of Bamberg, Bamberg.
e-mail: christian.assmann@uni-bamberg.de

C. Carstensen

Leibniz Institute for Educational Trajectories (LifBi), Bamberg.
Psychological Methods in Educational Research, Department of Psychology,
University of Bamberg, Bamberg.
e-mail: claus.carstensen@uni-bamberg.de

C. Gaasch

Leibniz Institute for Educational Trajectories (LifBi), Bamberg.
Chair of Statistics and Econometrics, University of Bamberg, Bamberg.
e-mail: christoph.gaasch@lifbi.de

S. Pohl

Department Methods and Evaluation/Quality Assurance,
Free University Berlin, Berlin.
e-mail: steffi.pohl@fu-berlin.de

Scoring of Complex Multiple Choice Items in NEPS Competence Tests

Kerstin Haberkorn, Steffi Pohl, Claus Carstensen and Elena Wiegand

Abstract

In order to precisely assess the cognitive achievement and abilities of students, different types of items are often used in competence tests. In the National Educational Panel Study (NEPS), test instruments also consist of items with different response formats, mainly simple multiple choice (MC) items in which one answer out of four is correct and complex multiple choice (CMC) items comprising several dichotomous “yes/no” subtasks. The different subtasks of CMC items are usually aggregated to a polytomous variable and analyzed via a partial credit model. When developing an appropriate scaling model for the NEPS competence tests, different questions arose concerning the response formats in the partial credit model. Two relevant issues were how the response categories of polytomous CMC variables should be scored in the scaling model and how the different item formats should be weighted. In order to examine which aggregation of item response categories and which item format weighting best models the two response formats of CMC and MC items, different procedures of aggregating response categories and weighting item formats were analyzed in the NEPS, and the appropriateness of these procedures to model the data was evaluated using certain item fit and test fit indices. Results suggest that a differentiated scoring without an aggregation of categories of CMC items best discriminates between persons. Additionally, for the NEPS competence data, an item format weighting of one point for MC items and half a point for each subtask of CMC items yields the best item fit for both MC and CMC items. In this paper, we summarize important results of the research on the implementation of different response formats conducted in the NEPS.

1 Item Formats and Scaling Model of the NEPS Competence Tests

In the process of test development, the choice of the items' format plays a crucial role for different aspects of validity (Rodriguez, 2002). So far, comprehensive item writing rules and guidelines have been published (Downing & Haladyna, 2006; Haladyna & Rodriguez, 2013; Osterlind, 1998), and a variety of analyses have been performed on different item formats in order to evaluate the strengths and weaknesses of each response format. A main distinction is usually made between selected response (SR) items and constructed response (CR) items. Whereas constructed response items require the examinee to create a response to a specific question or item stem, selected response items require choosing an answer out of a set of options or matching options to several stems that are presented. Most assessments make use of the SR item format (Osterlind, 1998). SR items ensure an efficient and effective measurement, and a large body of research shows that thoroughly and representatively constructed SR items achieve high content validity (Downing, 2006; Haladyna & Downing, 2004; Rodriguez, 2002). Furthermore, the objective, efficient scoring prevents threats to validity, such as construct-irrelevant variance induced by the subjectivity of human raters (Haladyna & Rodriguez, 2013).

In the National Educational Panel Study (NEPS), different types of SR items are used in the competence tests. In the NEPS, the tests measuring mathematical competence, reading competence, scientific literacy, and information and communication technologies (ICT) literacy mainly include simple multiple choice (MC) and complex multiple choice (CMC) items¹ (see Pohl & Carstensen, 2012, for a more detailed description of the different response formats; for an overview of the competencies, see also Weinert et al., 2011). MC items in the NEPS usually consist of four response options, with one being correct and three being incorrect. CMC items in the NEPS are composed of a number of subtasks, with one out of two response options being correct. An example for an MC and a CMC item is presented in Figure 1. The number of subtasks within CMC items varies in the NEPS competence tests.

As CMC items consist of item bundles with a common stimulus, the assumption of local item independence may be violated within CMC items (e. g., Yen, 1993). To account for this local item dependence (LID), the subtasks within CMC items are usually aggregated to polytomous super-items, as suggested by many researchers (e. g., Andrich, 1985; Ferrara, Huynh, & Michaels, 1999). Several psychometric models have been developed for polytomous variables. The item bundles may, for example, be analyzed via a graded response or a partial credit model (Huynh, 1994; Wainer, Sireci, & Thissen, 1991). For scaling the NEPS competence data, a partial credit model (Masters, 1982) was used. The partial credit model was deliberately chosen because

1 Note that some test instruments in the NEPS additionally contain matching items as another type of SR item and constructed response items, but these response formats are rare and thus not considered in the analyses.

Figure 1 Example of (a) an MC item and (b) a CMC item within NEPS competence tests (Neumann et al., 2013)

Mr. Brown owns a rectangular piece of land and wants to fence it in. He has already made some calculations and then bought a 40 m fence. The piece of land has a width of 8 m. How long is the land?

<input type="checkbox"/>	5 m
<input type="checkbox"/>	8 m
<input type="checkbox"/>	12 m
<input type="checkbox"/>	16 m

(a)

Are the following statements about the study's result correct?

	yes	no
Half of the participants showed at least one side effect, because 50 is half of 100.	<input type="checkbox"/>	<input type="checkbox"/>
Sickness occurred less than itching, because $50+40$ is less than $50+70$.	<input type="checkbox"/>	<input type="checkbox"/>
About 53% of the participants showed at least one side effect, because $(50+40+70)/3 \approx 53\%$.	<input type="checkbox"/>	<input type="checkbox"/>
More than half of the participants showing sickness also showed itching, because $50:90 > 50\%$.	<input type="checkbox"/>	<input type="checkbox"/>

(b)

of its membership in the family of Rasch models and the advantageous properties that Rasch models are known to have (Penfield, Myers, & Wolfe, 2008). For scaling the competence data, many large-scale studies, for example, PISA or NEPS, use one-parameter (1PL) models or extensions of this model to preserve the item weights intended by the instrument construction (see Pohl & Carstensen, 2012, for an argumentation of model choice in the NEPS). If the number of items from different conceptual aspects is intentionally chosen, the 1PL scaling model ensures the intended weightings of the conceptual aspects in contrast to the 2PL model, in which the items' weight depends on their empirical factor loadings. Given the 1PL model, we asked ourselves how we could best implement the different response formats in the scaling model and especially how we should score the categories of the CMC items and how we should weight both MC and CMC items.

2 Research on the Implementation of Response Formats Within a Scaling Model

Until now, several methods of implementing items with different response formats in a 1PL-scaling model have been applied in large-scale studies. The scoring procedures for items with different response formats, in particular, differed in their degree of aggregation of categories they used for polytomous variables as well as in their weighting of the item formats. In the following section, first, common aggregation approaches for response categories of CMC items are presented, and second, weightings of different item formats within an Item Response Theory (IRT) framework are described.

2.1 Aggregation

The simple MC items are usually scored dichotomously, with one point given for a correct response and zero points given for the selection of an incorrect response (also called distractor). Reviewing various competence assessments that implemented different response formats, there are two widely applied aggregation methods for polytomous variables. First, the *All-or-Nothing scoring rule* is very common and means that subjects only receive full credit if all answers on subtasks are correct (Ben-Simon, Budescu, & Nevo, 1997). If at least one subtask is answered incorrectly, the person receives no credit. This method makes use of a dichotomous scoring and is implemented for CMC items in the study “Teacher Education and Development Study in Mathematics” (TEDS-M, see Blömeke, Kaiser, & Lehmann, 2010). Another established method of dealing with CMC items is the *Number Correct (NC) scoring rule*, which rewards partial knowledge, meaning that partial credit is given for each correctly solved subtask of a CMC item (see Ben-Simon et al., 1997). To apply the NC scoring rule, the subtasks of CMC items are formed to a composite score, and each of the categories receives partial credit according to the number of correctly answered subtasks. This scoring option is well known and has often been used in large-scale studies, such as PISA (Adams & Wu, 2002).

While several researchers have examined the impact of the two aggregation options for CMC items using parameters of classical test theory (CTT), there are only few results within the field of IRT. Hence, findings of research based on CTT are described first to get an impression of the impact of the two aggregation options before presenting results based on IRT. Based on CTT-analyses, Ben-Simon and colleagues (1997) reported a disadvantage of the All-or-Nothing scoring rule for students with low ability since the students’ partial knowledge is not captured. They pointed out that the NC scoring, in particular, measures lower-performing students more accurately. Hsu (1984) and Wongwiwatthanakit, Bennett, and Popovich (2000) demonstrated advantages of the NC scoring rule regarding reliability and discrimination.

Nevertheless, Hsu found only a slight increase in discrimination and reliability of the NC scoring in comparison with the All-or-Nothing scoring rule and thus argued that the slight gains of the NC scoring do not seem to justify the additional effort involved in this procedure in comparison with dichotomous scoring.

Si (2002) compared the effects of NC scoring and dichotomous scoring using IRT. In his study, he applied several dichotomous and polytomous IRT-models to simulated item-response data and investigated effects on parameter estimation using different model parameterizations (1-, 2-, and 3PL) and degrees of aggregation (dichotomous versus polytomous). His results provided evidence that polytomous models produce more accurate ability estimates than dichotomous models independent of the prior distribution of the persons' abilities. Furthermore, the 1PL model considerably outperformed the 2PL- and 3PL models. Among the polytomous models, the partial credit model exhibited the most accurate ability estimation. Nevertheless, Si only examined the effect of various models on the accuracy of the estimated person abilities.

2.2 Weighting of Different Response Formats

Besides their variation in the degree of aggregation of response categories within polytomous CMC items, competence assessments also differ in their allocation of scores for solving items with different response formats. PISA, for instance, awards one point for correctly solved MC items. The CMC items are given different maximum scores based on theoretical considerations by the test developers (OECD, 2009). There are a few CMC items with special requirements that are therefore scored with a maximum score of two points. Other CMC items are weighted equally to the simple MC items and are hence given a maximum score of one point when all subtasks are solved correctly. During the development of scaling models for the NEPS competence data, the question arose of whether CMC items should receive the same maximum score as simple MC items or whether they should have more impact on the overall competence score. One may argue that CMC items should be scored equally to MC items to make sure that the different items in the test contribute equally to the competence score. Others may suggest that CMC items should be weighted more as they incorporate a set of tasks and each subtask should get the same maximum score as an MC item. CMC items contain two response options, whereas simple multiple choice items consist of four response options. Thus, an appropriate procedure might also be a scoring of half points for each subtask while MC items receive one point when solved correctly.

Up to now, there has been only little research on weighting different types of item formats, especially concerning the item formats implemented in the NEPS competence tests. In contrast, differential weighting of items has received considerable attention in scaling test instruments. In the field of CTT, different methods and prin-

ciples for weighting items have been established (Ben-Simon et al., 1997; Kline, 2005; Stucky, 2009). Overall, the weighting of items is usually performed using a statistical or theoretical approach. If item weighting is based on statistical data, items' reliability and factor loadings may be regarded. Weighting items by objective theoretical criteria involves weighting determined by experts or weights imposed by items' length, difficulty, or assumed validity. In the field of IRT, studies mainly focused on models with an implicit item weighting in 2- or 3-PL-models (Stucky, 2009). However, studies dealing with a priori weighting of response formats in IRT models to preserve the item weighting by construction are limited. Lukhele and Sireci (1995) as well as Sykes and Hou (2003) looked for ways to model different response formats with deliberately chosen weights via IRT. Lukhele and Sireci established a specific weighting of MC and constructed response (CR) items in a 1PL-model using "unweighted" IRT marginal reliabilities for weighting the different formats. Sykes and Hou also applied a priori weighting of MC and CR items to their test data by giving a maximum score of one point for each MC item and a maximum score of two points for each CR item, but they did not examine different weighting schemes to find out the best way to implement the response formats. In sum, these studies used a priori weighting for implementing response formats in an IRT framework, but fit indices of the response formats were not evaluated as important indicators for the appropriateness of the weighting procedure. Furthermore, only constructed response items and simple MC items were implemented, whereas CMC items, which are included in the NEPS competence data, were not.

Given the limited findings on the implementation of response formats in a 1PL model, different analyses were conducted in the NEPS in order to replicate and extend preliminary research into the best way to deliberately model different item formats. Two relevant questions concerning the response formats in the development of the scaling model that were addressed in the NEPS were as follows: *First*, to which degree should the response categories of CMC items be aggregated, and *second*, how should the response formats encompassing CMC and MC items be weighted assuming that both item types assess the same latent trait?

In the following section, we begin by illustrating the empirical study we carried out to find the best aggregation option for the CMC items in the NEPS. Second, we describe the NEPS research of Haberkorn, Pohl, and Carstensen (2015), who looked for the best weighting procedure of different response formats for the NEPS competence tests.

3 Investigating Aggregation for CMC Items in NEPS Competence Tests

3.1 Method

Sample and Instruments

For analyzing the impact of different aggregation schemes for CMC items in the scaling model, data from two competence domains, which were assessed in a main study of ninth graders in the National Educational Panel Study, were used. In the main study in Grade 9, the subjects were engaged in different competence tests. The analyses were conducted using the domains of *scientific competence* and *information and communication technologies (ICT) literacy*. The tests of scientific competence assessed children's scientific knowledge in the contexts of health, environment, and technology (Hahn et al., 2013). The ICT instrument tapped children's ability to locate and use essential information and their knowledge on different kinds of technology, such as hardware and software (Senkbeil, Ihme, & Wittwer, 2012). The competence tests of scientific competence and ICT literacy contained a reasonable amount of MC and CMC items (see Schöps & Saß, 2013; Senkbeil & Ihme, 2012).

Since cases with less than three valid responses were excluded from the IRT analyses, the analyses were undertaken based on 14,301 subjects for scientific competence and 14,312 subjects for ICT literacy.² The test instrument to assess scientific competence consisted of 19 simple MC items and nine CMC items. The number of subtasks within the CMC items varied from four to six items. The test instrument of ICT literacy included 32 MC items and eight CMC items, and there were four to seven subtasks within the CMC items.

Analyses

The partial credit model (Masters, 1982) was used to apply the different scoring approaches to the data. Marginal maximum likelihood estimation was chosen for estimating the models, and all analyses were done using ConQuest (Wu, Adams, Wilson, & Haldane, 2007). If at least one of the subtasks of CMC items contained a missing value, the whole CMC item was coded as missing response. According to Gräfe (2012) as well as Pohl, Gräfe, and Rose (2013), ignoring missing responses in the scaling model yields unbiased item- and person parameter estimates. Therefore, missing responses were ignored in the application of the different scoring procedures. If response categories of the polytomous CMC items had less than 200 cases, adjacent categories were combined to avoid possible estimation problems. This occurred for the lowest categories, in particular, and predominantly if the CMC item consisted of many subtasks. For scientific competence, the two lowest categories of a CMC vari-

2 Note that due to later updates and data-editing processes, the number of persons and items may slightly differ from the number of persons and items found in the Scientific Use File.

able were collapsed into one category and received a score of zero points within four CMC items. For ICT literacy, the lowest categories of zero and one were combined into one category within seven CMC items due to low cell frequencies.

Different aggregation schemes for the categories of polytomous items were applied to the data. The MC items were always scored as zero points for an incorrect answer and as one point for a correct answer. In order to examine the impact of aggregation of response categories, CMC items were scored a) dichotomously, with one point given if all subtasks were answered correctly and zero points otherwise. This resembles the All-or-Nothing scoring rule implemented for most of the CMC items in PISA. In contrast, the second rule b) was a more differentiated scoring according to the NC scoring rule, with a maximum score of one point for a correct response on all subtasks and partial credit for each correctly answered subtask. The partial credit points ranged between zero points and one point in equal intervals. As a consequence, the partial credit steps were different depending on the number of categories within the CMC item. For example, the categories of a CMC item with five categories were scored with a score of $r = 0, 0.25, 0.5, 0.75, \text{ and } 1$, whereas the categories of a CMC item with four categories were scored $r = 0, 0.33, 0.67, \text{ and } 1$.

To get detailed information about changes in item- and test parameters caused by the two aggregation options, the CMC items were first analyzed separately without considering MC items, and different item statistics were investigated. We evaluated difficulty, correlation of the item score of CMC items with the total score (discrimination value as computed in ConQuest), and test reliability of the two aggregation rules. The correlation of the item score with the total score corresponds to the product-moment-correlation between the categories of CMC items and the total score, and the correlation is labeled as discrimination in the following sections. Furthermore, based on analyses of both MC and CMC items, the range of the abilities of test takers with partially correct answers was explored in order to assess the amount of information that is lost by applying a dichotomous scoring. For this purpose, differences between person ability in the second-highest and the lowest response categories were computed for each polytomous item. For example, for a CMC item with 4 subtasks, subjects with only incorrect answers might have a medium ability of -0.54 logits (the estimate of person ability in each category is always computed using the other items in the test only), whereas subjects who solved three out of the four subtasks might have a medium ability of 0.03 logits. Thus, person ability between the lowest and the second-highest response category in this case would vary with a range of 0.57 logits. This range of person ability is combined into one category in the All-or-Nothing scoring rule. Therefore, a computation of the range of person abilities is performed to investigate how much information we lose if we analyze these persons together in one category.

3.2 Results

First, we present the comparison of the two aggregation procedures for the categories of CMC items, the All-or-Nothing scoring, and the NC scoring. In Table 1, the item difficulty and discrimination for the All-or-Nothing scoring and the NC scoring in the Science and ICT domains are depicted.

With regard to item difficulty, high differences between the All-or-Nothing scoring and the NC scoring emerged. The NC scoring for CMC items yielded considerably lower difficulty estimates than the All-or-Nothing scoring. Comparing the two aggregation options by the average item difficulties, their means differed by about 3.17 logits (standard deviation (*SD*) = 0.71) for Science and 3.46 logits (*SD* = 0.69) for ICT. Thus, substantially higher item difficulties were estimated for the All-or-Nothing scoring than for the NC scoring since subjects with partially correct answers were given no credit in the All-or-Nothing scoring and there were consequently more subjects with zero points on the items. Furthermore, the item discrimination varied slightly to moderately between the dichotomous scoring and the NC scoring. For most of the items in Science and ICT, discrimination at the item level increased when applying the NC scoring. For six out of the 17 items, rather equal discriminations oc-

Table 1 Item Location Parameters, Characterizing the Items' Difficulty (in Logits), and Discrimination of the All-or-Nothing Scoring and the NC Scoring

	Science				ICT			
	Location parameter		Discrimination		Location parameter		Discrimination	
	All-or-Nothing scoring	NC scoring	All-or-Nothing scoring	NC scoring	All-or-Nothing scoring	NC scoring	All-or-Nothing scoring	NC scoring
CMC_1	-0.30	-4.11	0.47	0.48	0.38	-2.57	0.50	0.53
CMC_2	1.58	-1.34	0.41	0.49	0.73	-3.63	0.50	0.49
CMC_3	1.02	-3.39	0.46	0.45	0.79	-2.02	0.45	0.42
CMC_4	0.33	-2.47	0.57	0.56	0.61	-3.47	0.56	0.56
CMC_5	0.26	-3.17	0.57	0.58	0.46	-2.73	0.48	0.50
CMC_6	-0.24	-2.39	0.52	0.56	0.24	-2.93	0.57	0.59
CMC_7	0.92	-2.58	0.55	0.54	2.01	-2.16	0.44	0.62
CMC_8	0.02	-2.34	0.50	0.54	1.75	-1.20	0.36	0.50
CMC_9	0.63	-2.48	0.55	0.58	For ICT, there were only 8 CMC items.			
<i>Means</i>	0.47	-2.70	0.51	0.53	0.87	-2.59	0.48	0.53

Note. The analyses for these results were undertaken using CMC items only

curred. Overall, the average discrimination showed moderate gains resulting in more differentiated measures for the NC scoring.

Differences between the two aggregation options were even more evident when comparing the reliability. For the Science domain, the NC scoring (EAP/PV reliability = 0.652, WLE reliability = 0.595) yielded higher reliability estimates than the All-or-Nothing scoring (EAP/PV reliability = 0.593, WLE reliability = 0.433). The reliability improved substantially for the NC scoring (EAP/PV reliability = 0.518, WLE reliability = 0.444) (especially for ICT) in comparison with the All-or-Nothing scoring (EAP/PV reliability = 0.444, WLE reliability = 0.150).

In order to evaluate the possible loss of information in the application of the All-or-Nothing scoring, the range of the abilities of persons within the categories that were collapsed in the dichotomous scoring was examined. For a reliable estimation of these abilities, the analyses were performed based on MC and CMC items. The range of person abilities for each CMC item was computed as the difference between the medium ability of subjects who were in the second-highest category and the medium ability of subjects in the lowest category (see Table 2). For example, regarding the first CMC item of the ICT test, which contained three categories, the range of person abilities within the base to the second categories was 0.67 logits, indicating that subjects reaching the second category had a higher overall ability by 0.67 logits on average than subjects who didn't solve any of the subtasks of the CMC item. In the dichotomous scoring, these categories within CMC items (for Item 1 in ICT category 0-2) were collapsed and scored with zero points.

Table 2 Range of the Abilities (in Logits) of Persons Who Answered Incorrectly or Only Partially Correctly

Item	Science		ICT	
	Number of categories	Range of abilities	Number of categories	Range of abilities
CMC_1	3	0.83	3	0.67
CMC_2	3	0.72	4	0.86
CMC_3	4	0.82	5	-0.16
CMC_4	5	0.51	5	0.47
CMC_5	4	1.00	3	0.80
CMC_6	3	0.47	5	0.74
CMC_7	4	0.57	6	1.02
CMC_8	4	0.79	4	1.00
CMC_9	4	0.90	-	-

For Science, the test consisted of nine CMC items, and persons who received no or only partial credit varied substantially in their general ability (computed across the other items in the test), with $M = 0.73$ logits ($SD = 0.18$) on average. The highest differences occurred for Item 5. Subjects who solved three out of the four subtasks correctly had a higher overall ability by about one logit than subjects who didn't solve any subtasks correctly for this item. However, the persons who differed considerably in their ability were treated equally in the NC scoring. Eight CMC items were included in the ICT test, and persons who were collapsed into one group in the dichotomous scoring also exhibited substantial variation in their overall estimated ability ($M = 0.68$, $SD = 0.38$), except for Item 3. This item had an unsatisfactory item fit, and the persons who didn't solve any of the subtasks correctly had a higher ability by 0.16 logits than persons who solved four fifths of the subtasks of the CMC item. In this case, the reversed range of abilities underlines the misfit of the item to the model.³ Overall, the analyses of the abilities' range indicate that persons who received no or only partial credit differed greatly in their general ability.

Taking together the impact of the two aggregation options on item difficulty, discrimination, test reliability, and person's range of abilities with no or partially correct answers, the results provide evidence for rather high gains in information about subjects' competencies using the NC scoring instead of the All-or-Nothing scoring.

4 Overview of Research on Weighting of Response Formats in NEPS Competence Tests

The question of how to appropriately weight different NEPS response formats in a IPL model was investigated in an elaborate study by Haberkorn et al. (2015), and the main findings of the study are presented in the following section. In order to examine the impact of different weighting schemes of CMC and MC items on the item parameters, Haberkorn et al. made analyses based on the same NEPS competence data of Science and ICT from the main study in G9 which was used for exploring the influence of aggregating CMC items. Since items with low item fit statistics were excluded from the final dataset (Schöps & Sass, 2013; Senkbeil & Ihme, 2012), the analyses of weighting were based on 9 CMC and 19 MC items in Science as well as 10 CMC and 17 MC items in ICT. Three different weighting procedures were compared by Haberkorn and her colleagues, and for each of the options, the categories of the CMC items were given partial credit. As a consequence, the degree of aggregation did not differ among the different weighting options. This allowed for disentangling item weighting from the aggregation procedure for the response categories. The implemented weighting options were as follows: The correctly solved MC items were always scored with one point. The CMC items a) were given a maximum score of one point to equal

3 Due to unsatisfactory item fit, this item was not included in the Scientific Use File.

Table 3 Example for Different Scoring Methods of a CMC Item With Six Categories

Categories of a CMC item with five subtasks	Three weighting options		
	(a) Maximum score is 1	(b) 0.5 points per correct subtask	(c) 1 point per correct subtask
0	0	0	0
1	0.2	0.5	1
2	0.4	1	2
3	0.6	1.5	3
4	0.8	2	4
5	1	2.5	5

their weight to the MC items, b) were scored by giving half points per category to reflect the reduced number of two response options within the subtasks instead of four response options in the MC items, and c) received one point per category, and the subtasks of the CMC items were thus weighted equally to the simple MC items. An example of the different scoring options used for a CMC item is depicted in Table 3.

Haberkorn et al. (2015) compared the weighted mean square (WMNSQ) and the respective *t*-value of the three scoring options in order to investigate the best a priori weighting for the two response formats of CMC and MC items. It is important to note that Haberkorn et al. used different statistical parameters for the evaluation of the weighting of item formats than for the evaluation of different aggregation options depending on the amount of information the parameters provided. The aggregation procedures, in particular, differed in their reliability and discrimination estimates but did not differ much in their WMNSQ estimates. The different weighting options also had different discrimination estimates, but the WMNSQ and corresponding *t*-value were more appropriate for an evaluation of the weighting options in order to find the most balanced fit for MC and CMC items within the Rasch model.

First, we present the main results for the Science domain found by Haberkorn et al. (2015). The impacts of the three weighting procedures for CMC items in relation to MC items (which were always scored with one point for a correct answer) are depicted in Figures 2 and 3: an equal weighting of MC and CMC items with a maximum score of one point, half points per subtask of CMC items, or one point per subtask for CMC items. Figure 2 includes means and standard deviations of the WMNSQ, separately computed across MC and CMC items, for the three different scoring options. Figure 3 depicts means and standard deviations of the *t*-value for the three different scoring options, separately computed across MC and CMC items.

As can be seen in these figures, an equal weighting of MC and CMC items, which meant that MC items as well as the polytomous CMC items were scored with a maxi-

Figure 2 Means and standard deviations of the WMNSQ for different item weightings in the domain of Science (Haber Korn et al., 2015)

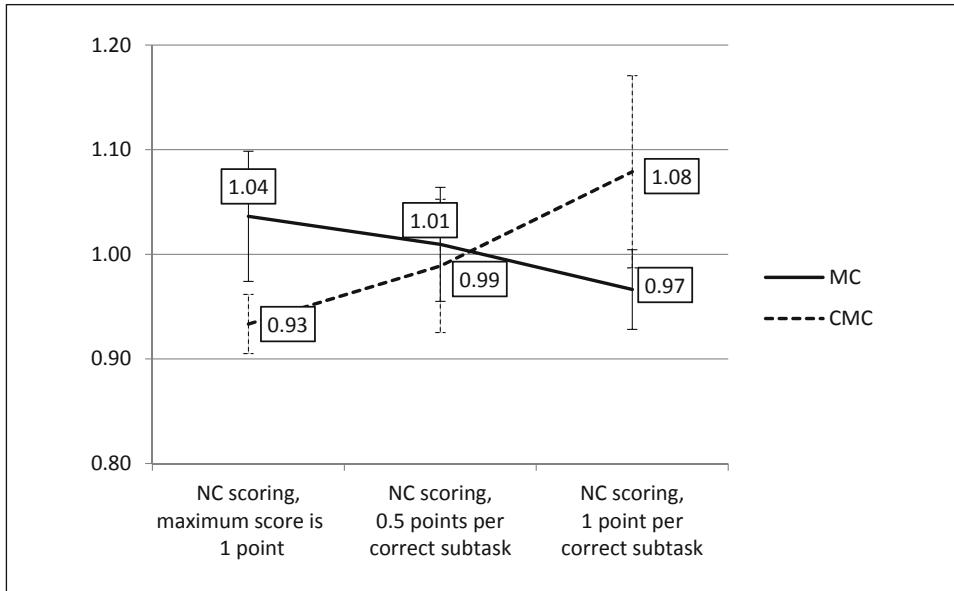
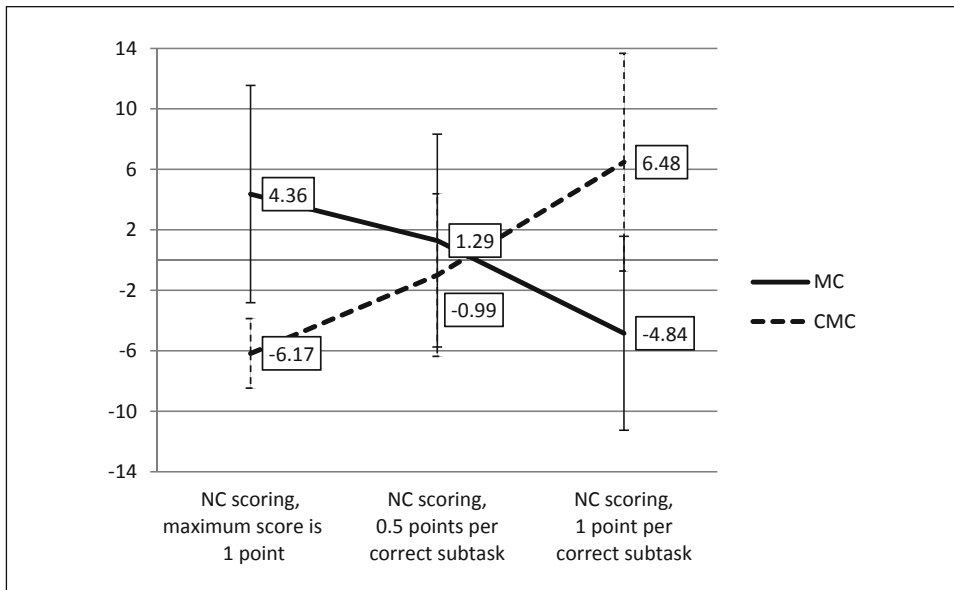


Figure 3 Means and standard deviations of the t-value of the WMNSQ for different item weightings in the domain of Science (Haber Korn et al., 2015)



imum of one point, resulted in an underfit for MC items and an overfit for CMC items. Both the WMNSQ (see Figure 2) and, more evident due to the rather large sample size, the t -value of the WMNSQ (see Figure 3) indicated that MC as well as CMC items did not fit the underlying model well. In contrast, the opposite was found to be true when each of the subtasks of CMC items was weighted equally to MC items and when correct responses to MC items as well as correctly solved subtasks of CMC items were consequently given one point in the scaling model. In this case, an overfit of MC items and a rather large underfit of CMC items emerged. A scoring of half points per category for the CMC items yielded the best item fit for the WMNSQ and the respective t -value. When the categories of the CMC items were given half of the weight of MC items, both MC and CMC items showed the most balanced fit.

Haberkorn et al. (2015) applied the same weighting procedures of CMC items in relation to MC items to the ICT data (see Table 4).

When looking at the WMNSQ and the respective t -value, the results of Science were replicated. An equal weighting of the MC items and the CMC items consisting of several subtasks caused an overfit of CMC items and a slight underfit of MC items. Conversely, with an equal weighting of the subtasks of CMC items to MC items, the CMC items showed a large underfit, and the MC items showed a slight overfit. Taking the fit of MC and CMC items together, the best fit of the weighted items to the model was given when each of the categories of CMC items was scored with half points. While a scoring of half points per category still resulted in a slight underfit of MC items in the Science domain, the same scoring option caused a quite optimal fit for both MC and CMC items for ICT (Haberkorn et al., 2015).

Haberkorn et al. (2015) also applied a restricted 2PL model in which loadings within response formats were set equal but were allowed to vary between response formats. By regarding the two discrimination indices for MC and CMC items, they received the empirical weight of the response formats. As expected, the values were close to 0.5. In addition to applying the different weighting approaches to NEPS com-

Table 4 Means and Standard Deviations (in Parentheses) of the WMNSQ and Corresponding t -Values for the Three Weighting Options in the Domain of ICT Literacy (Haberkorn et al., 2015)

Response format	Fit criterion	NC scoring, maximum score is 1	NC scoring, 0.5 points per correct subtask	NC scoring, 1 point per correct subtask
MC items	WMNSQ	1.02 (0.06)	1.00 (0.06)	0.97 (0.05)
	t -value	1.66 (6.75)	-0.06 (6.90)	-4.51 (6.87)
CMC items	WMNSQ	0.93 (0.04)	0.99 (0.03)	1.15 (0.05)
	t -value	-6.21 (3.30)	-0.26 (2.02)	11.41 (4.53)

Note. Correctly solved MC items were always scored with one point.

petence data, Haberkorn et al. studied the impact of the weighting options on fit indices in PISA competence tests. Their results replicated the findings of the NEPS research and demonstrated that weighting the subtasks of CMC items with half of the weight of MC items yielded a quite appropriate fit of MC and CMC items to the model.

5 Conclusion and Discussion

The aim of this chapter was to provide an overview of major research issues concerning the implementation of MC and CMC items in a Rasch model addressed in the NEPS. According to often-applied scoring procedures in competence assessments and based on theoretical deliberations, the impact of different degrees of aggregating response categories within polytomous CMC items was explored in the NEPS, and the appropriateness of different weighting schemes was investigated.

With regard to the aggregation options, the comparison of the All-or-Nothing scoring and the Number Correct scoring showed clear evidence of the discriminating effect of the NC scoring. To avoid a loss of information, CMC items should be scored as differentiated as possible. The application of a dichotomous scoring for CMC items may implicate the assumption that subjects answering no subtask correctly and subjects answering some subtasks of an item correctly do not differ in their ability. Indeed, the current investigation has documented that there is considerable variation in ability within these subjects. Thus, following the suggestions of other researchers (Si, 2002), NC scoring should be preferred over All-or-Nothing scoring to improve the accuracy of ability estimates. However, limitations in the application of NC scoring may arise due to low cell frequencies in certain categories. In this case, categories within CMC items may be collapsed in the scaling of the data in order to avoid estimation problems (OECD, 2009; Pohl & Carstensen, 2012, 2013).

The investigation of different weighting schemes for CMC items in relation to MC items carried out by Haberkorn et al. (2015) pointed consistently to the fact that a scoring of about half a point for the categories within CMC items while awarding one point per MC item matches the empirical data quite well. In contrast, the other weighting procedures performed substantially worse in the Science and ICT domains. Of course, the relative weight of MC and CMC items might differ with regard to other age groups, competence domains, or large-scale studies. Competence assessments that aim at assessing other abilities and skills using these item formats might obtain other suitable scoring schemes. In the development of a 1PL scaling model, it therefore seems crucial to empirically evaluate weights that are constituted theoretically a priori. As argued by Haberkorn et al. (2015), a combination of applying 2PL models in the development of a scaling model and using a priori weights in the final application of a 1PL model may hence serve as a promising procedure for competence assessments to implement theoretically constituted features and, simultaneously, enhance the statistical properties of the scaling model.

The analyses computed by Haberkorn et al. included the main item formats within NEPS competence tests; recommendations for weighting item formats are thus restricted to CMC and MC items. Further research on response formats applied in other large-scale studies, such as constructed response items, will be useful to extend weighting guidelines. Finally, studies on competence tests in other age groups, competence domains, and national as well as international studies will be of interest to expand upon the current understanding of the best way to comprise different response formats in a scaling model.

References

- Adams, R., & Wu, M. (2002). *PISA 2000 technical report*. Paris, France: OECD.
- Andrich, D. (1985). An elaboration of Guttman scaling with Rasch models for measurement. In N. Brandon-Tuma (Ed.), *Sociological methodology* (pp. 33–80). San Francisco, CA: Jossey-Bass.
- Ben-Simon, A., Budescu, D. V., & Nevo, B. (1997). A comparative study of measures of partial knowledge in multiple-choice tests. *Applied Psychological Measurement, 21*(1), 65–88.
- Blömeke, S., Kaiser, G., & Lehmann, R. (2010). *TEDS-M 2008—Professionelle Kompetenz und Lerngelegenheiten angehender Primarstufenlehrkräfte im internationalen Vergleich*. Münster: Waxmann.
- Downing, S. M., & Haladyna, T. M. (Eds.). (2006). *Handbook of test development*. Mahwah, NJ: L. Erlbaum.
- Downing, S. M. (2006). Selected-response item formats in test development. In S. M. Downing, & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 3–26). Mahwah, NJ: Erlbaum.
- Ferrara, S., Huynh, H., & Michaels, H. (1999). Contextual explanations of local dependence in item clusters in a large-scale hands-on science performance assessment. *Journal of Educational Measurement, 36*(1), 119–140.
- Gräfe, L. (2012). *How to deal with missing responses in competency tests? A comparison of data- and model-based IRT approaches* (Unpublished diploma thesis). Friedrich-Schiller-University Jena, Jena, Germany.
- Hahn, I., Schöps, K., Rönnebeck, S., Martensen, M., Hansen, S., Saß, S., ... Prenzel, M. (2013). Assessing scientific literacy over the lifespan—A description of the NEPS science framework and the test development. *Journal of Educational Research Online, 5*(2), 110–138.
- Haberkorn, K., Pohl, S., & Carstensen, C. (2015). *Incorporating different response formats of competence tests in an IRT model*. Manuscript submitted for publication.
- Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice, 23*(1), 17–27.
- Haladyna, T. M., & Rodriguez, M. C. (2013) *Developing and validating test items*. New York, NY: Routledge.

- Hsu, T. C. (1984). The merits of multiple-answer items as evaluated by using six scoring formulas. *Journal of Experimental Education*, 52(3), 152–158.
- Huynh, H. (1994). On equivalence between a partial credit item and a set of independent Rasch binary items. *Psychometrika*, 59(1), 111–119.
- Kline, T. (2005). *Psychological testing: A practical approach to design and evaluation*. Thousand Oaks, CA: Sage.
- Lukhele, R., & Sireci, S. G. (1995, April). *Using IRT to combine multiple-choice and free-response sections of a test on to a common scale using a priori weights*. Paper presented at the annual conference of the National Council on Measurement in Education, San Francisco, CA.
- Masters, G. N. (1982). A rasch model for partial credit scoring. *Psychometrika*, 47(2), 149–174.
- Neumann, I., Duchardt, C., Grüßing, M., Heinze, A., Knopp, E., & Ehmke, T. (2013). Modeling and assessing mathematical competence over the lifespan. *Journal of Educational Research Online*, 5(2), 80–109.
- OECD (2009). *PISA 2006 technical report*. Paris, France: OECD.
- Olson, J. F., Martin, M. O., & Mullis, I. V. S. (Eds.). (2008). *TIMSS 2007 technical report*. Chestnut Hill, MA: Boston College.
- Osterlind, S. J. (1998). *Constructing Test Items: Multiple-Choice, Constructed-Response, Performance, and Other Formats*. Dordrecht, Netherlands: Kluwer Academic.
- Penfield, R. D., Myers, N. D., & Wolfe, E. W. (2008). Methods for assessing item, step, and threshold invariance. Polytomous items following the partial credit model. *Educational and Psychological Measurement*, 68(5), 717–733.
- Pohl, S., & Carstensen, C. H. (2012). *NEPS technical report—Scaling the data of the competence tests*. (NEPS Working Paper No. 14). Bamberg: University of Bamberg, National Educational Panel Study.
- Pohl, S., & Carstensen, C. H. (2013). Scaling the competence tests in the National Educational Panel Study—Many questions, some answers, and further challenges. *Journal of Educational Research Online*, 5(2), 189–216.
- Pohl, S., Gräfe, L., & Rose, N. (2013). Dealing with omitted and not reached items in competence tests—Evaluating approaches accounting for missing responses in IRT models. *Educational and Psychological Measurement*, 74(3), 423–452.
- Rodriguez, M. (2002). Choosing an item format. In G. Tindal, & T. M. Haladyna (Eds.), *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation* (pp. 213–231). Mahwah, NJ: Erlbaum.
- Schöps K., & Saß, S. (2013). *NEPS technical report for science—Scaling results of Starting Cohort 4 in ninth grade*. (NEPS Working Paper No 23). Bamberg: University of Bamberg, National Educational Panel Study.
- Senkbeil, M. & Ihme, J. M. (2012). *NEPS technical report for computer literacy—Scaling results of Starting Cohort 4 in ninth grade*. (NEPS Working Paper No. 17). Bamberg: University of Bamberg, National Educational Panel Study.

- Senkbeil, M., Ihme, J. M., & Wittwer, J. (2013). The Test of Technological and Information Literacy (TILT) in the National Educational Panel Study: Development, empirical testing, and evidence for validity. *Journal of Educational Research Online*, 5(2), 139–161.
- Si, C. B. (2002). *Ability estimation under different item parameterization and scoring models* (Doctoral dissertation). Retrieved from http://digital.library.unt.edu/ark:/67531/metadc31116/m2/1/high_res_d/dissertation.pdf
- Stucky, B. D. (2009). *Item response theory for weighted summed scores* (Master's thesis). Retrieved from https://cdr.lib.unc.edu/indexablecontent?id=uuid:03c49891-0701-47b8-af13-9c1e5b60d52d&ds=DATA_FILE
- Sykes, R. C., & Hou, L. (2003). Weighting constructed-response items in IRT-based exams. *Applied Measurement in Education*, 16(4), 257–275.
- Wainer, H., Sireci, S. G., & Thissen, D. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement*, 28(3), 237–247.
- Weinert, S., Artelt, C., Prenzel, M., Senkbeil, M., Ehmke, T., & Carstensen C. H. (2011). Development of competencies across the life span. In H.-P. Blossfeld, H.-G. Roßbach, & J. von Maurice (Eds.), *Education as a lifelong process: The German National Educational Panel Study (NEPS)* (pp. 67–86). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Wongwiwatthanakul S., Bennett, D. E., & Popovich N. G. (2000). Assessing pharmacy student knowledge on multiple-choice examinations using partial-credit scoring of combined-response multiple-choice items. *American Journal of Pharmaceutical Education*, 64(1), 1–10.
- Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. (2007). *ACER ConQuest 2.0—Generalised item response modelling software*. Camberwell, Australia: ACER Press.
- Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, 30(3), 187–213.

About the authors

K. Haberkorn
University of Bamberg, Bamberg.
e-mail: kerstin.haberkorn@uni-bamberg.de

C. Carstensen
Leibniz Institute for Educational Trajectories (LifBi), Bamberg.
Chair of Department for Psychology and Methods of Educational Research,
University of Bamberg, Bamberg.

S. Pohl
Chair of Methods and Evaluation/Quality Assurance,
Free University Berlin, Berlin.

E. Wiegand
University of Mannheim, Mannheim.

V. Assessing Data Quality

Measurement of Preschool Quality Within the National Educational Panel Study— Results of a Methodological Study¹

Thomas Bäumer and Hans-Günther Roßbach

Abstract

It has been argued that for the assessment of educational quality—especially in preschool—observational methods are the silver bullet. However, in large-scale assessments like the National Educational Panel Study (NEPS), observations can hardly be conducted. Against this background, we carried out a study organized around the question of the extent to which preschool quality can be assessed using teacher questionnaires. Therefore, a standardized observation of 60 preschool groups, using the German versions of the Early Childhood Environment Rating Scale (ECERS; KES-R and KES-R-E)—a well-established rating instrument for (process) quality in preschool environments—was conducted. Moreover, teachers filled out a questionnaire on preschool quality from pilot studies of the NEPS. In this paper, we present main results from the comparative analyses of observations and surveys of preschool quality using regression analytical methods. It can be shown that on a global level, preschool quality can be reproduced quite well by means of questionnaire data. Conclusions concerning the questionnaire design of the NEPS main study in Starting Cohort 2—Kindergarten are drawn. Finally, some cautionary notes on the use of single indicators of preschool quality and on causal inferences are given.

1 We would like to express our gratitude and appreciation to Carina Pömp, who collected the data in the context of her diploma thesis. This paper discusses the main results of the study. We will deal with the rich data resource of this study in more detail in a forthcoming NEPS Working Paper.

1 Introduction

The measurement of the quality of learning environments is one of the most prominent but also one of the most difficult issues in educational research. It has been argued that a comprehensive view on educational quality necessitates a triangulation of the perspectives of teachers, students, and external observers. Every perspective has its own advantages as well as disadvantages for different aspects of quality and also in relation to outcomes that may depend on educational quality. In panel studies like the NEPS, external observations are often too costly and impose additional challenges on data dissemination. Therefore, we have to rely on reports from respondents, usually teachers and students. However, in preschool, only the teachers' perspective can be surveyed because children are too young to respond adequately to a questionnaire or interview. One has to account for respondent bias, especially in the case of preschool quality, because it is very likely that teachers will perceive the survey as an evaluation of their work. In the following section, we describe a methodological study that compares the observation of preschool quality with reports made by preschool teachers. First, we give a few notes on the concept of educational quality used within the NEPS.

2 Theoretical Background

At least since the "PISA Shock" in Germany, early education institutions have been brought to public attention. There is ample evidence that preschool attendance has long-term effects on the cognitive and social development of the child (cf. Anders, 2013). Furthermore, it can be shown that these effects are moderated by the quality of the preschool (cf. Roßbach, 2004, 2005; Roßbach, Kluczniok, & Kuger, 2008; Sylva, Melhuish, Sammons, & Taggart, 2004). Therefore, the assessment of the quality of the preschool groups attended by the children (target persons) of NEPS Starting Cohort 2—Kindergarten is given primary attention.

As has been argued by Clausen (2002), different perspectives in the evaluation of quality should be accounted for in educational settings. Teachers as well as students can perceive different aspects with different accurateness. An external observer might be the most reliable source of information but only has access to the educational situation for a limited time, and observation might additionally alter the situation that is being observed. In preschool, accounts by children cannot be used because they are too young to give reliable assessments. Therefore, observation is the method of choice in preschool settings. However, in large-scale assessments like the NEPS, with a huge amount of target persons and educational settings, observational studies are too costly and time-consuming to be able to be applied on a regular basis. Here, we have to rely on teachers' reports on the quality of their preschool groups. The assumption that teachers' perspectives might be distorted because they perceive the survey

setting as an evaluation of their own work within the preschool group motivated the present study.

Tietze et al. (1998) define the educational quality of preschool with the help of three elements that match the conceptualization of educational quality within the NEPS (cf. Bäumer, Preis, Roßbach, Stecher, & Klieme, 2011; Bäumer & Roßbach, 2012): *Structural quality* includes the framing conditions of an educational setting (like class size), *orientational quality* reflects the beliefs and opinions concerning the education of the actors within an educational setting, and *process quality* relates mainly to the interactions of the actors within an educational setting, for example, of teachers and children in a preschool group. Structural and orientational quality serve as input conditions for process quality that directly influence outcomes, whereas the former are proposed to mainly indirectly influence outcomes.

3 Method

3.1 Sample

The sample consists of 60 groups from 42 preschools in Upper Franconia. Recruitment was conducted under the following two conditions: (1) The age range of the attending children is between three and six years and (2) no special pedagogical conception (as used by Waldorf, for example) is followed.

3.2 Instruments

Observational Instruments

The observation was conducted using the *KES-R* (“Kindergarten-Skala—Revidierte Fassung”; Tietze, Schuster, Grenner, & Roßbach, 2005) and the *KES-R-E* (“Kindergarten-Skala-Erweiterung”; Roßbach & Tietze, in preparation), which are the German versions of the *ECERS-R* (“Early Childhood Environment Rating Scale. Revised Edition”; Harms, Clifford, & Cryer, 1998) and the *ECERS-E* (“Early Childhood Environment Rating Scale-Extension”; Sylva, Siraj-Blatchford, & Taggart, 2003), respectively. The *KES-R* assesses the global educational process quality of preschool groups. It consists of 43 items, each with a seven-point-rating scale ranging from 1 (“deficient”) to 7 (“excellent”). As a general guideline, scores of 1 to 3 denominate poor quality, and scores of 5 to 7 represent good quality. Scores of 3 to 5 are considered mediocre quality. The 43 items can be organized into seven subscales: *Space and Furnishings*, *Personal Care Routines*, *Language-Reasoning*, *Activities*, *Interactions*, *Program Structure*, and *Parents and Staff*. The last subscale was not used in this study, which thus used only 37 items for the observation. The *KES-R-E* adds more domain-specific aspects to the global assessment. It consists of four subscales: *Literacy*, *Mathematics*, *Science*

and *Environment* as well as *Diversity*. The 18 items are rated using the same scale as with the KES-R.

Teacher Questionnaire

The questionnaire used in this study was compiled of questions used in the first pilot study of NEPS Starting Cohort 2—Kindergarten. The questions can be subdivided into four broad domains: (1) *structural characteristics*: opening hours, closing times, group size, room count, and space and personnel (full time equivalents); (2) *compositional characteristics*: handicapped children, children with development disorders, children with migration background, age groups of children, children by gender, and amount of childcare per child; (3) *materials and activities*; and (4) *teachers' characteristics*: qualification, working hours and work schedule, further education, and supervision.

3.3 Procedure

The study was undertaken from November 2009 to February 2010. The observations took place during a three-to-four-hour session in the morning and were conducted by one or two trained observers. For half of the sample, the questionnaires were sent to the teachers prior to the observation, and for the other half, the questionnaire was handled over after the observation to control for the sequencing of the survey methods. The data of the ratings form and the questionnaires were manually keyed in twice by different typists to avoid input data error. Analyses were conducted using IBM SPSS Statistics 20.0.0.

4 Results

First, the descriptives of the observational and questionnaire data are shown. After that, the regression of the observed quality on teacher reports is presented.

4.1 Descriptives

KES-R Observational Data

In Table 1, descriptives of the scales and subscales of the KES-R and KES-R-E are summarized. All scales are computed as means of the corresponding items, and scale values range accordingly from 1 to 7. Most scale mean scores show a mediocre quality, with the exception of *Language Reasoning*, which is on average of good quality. Two scales show on average even poor quality of the 60 examined preschool groups: *Diversity* and *Personal Care Routines*. For diversity with a maximum score of 4.00,

Table 1 Descriptives of the KES-R and KES-R-E scales

Scale	<i>M</i>	<i>SD</i>	<i>min</i>	<i>max</i>
1. Space and Furnishings	3.59	0.84	2.13	5.88
2. Personal Care Routines	2.82	1.06	1.50	6.00
3. Language-Reasoning	5.17	1.07	3.00	7.00
4. Activities	3.96	0.79	2.50	6.11
5. Interactions	4.59	1.01	2.00	7.00
6. Program Structure	3.71	0.84	2.33	7.00
7. Literacy	3.89	0.95	1.67	6.17
8. Mathematics	3.43	1.08	1.22	5.67
9. Science and Environment	3.53	1.08	1.14	5.71
10. Diversity	2.39	0.97	1.00	4.00
11. KES-R	3.97	0.69	2.71	5.79
12. KES-R-E	3.31	0.93	1.72	5.30
13. KES-R/R-E Composite	3.64	0.75	2.53	5.46

no single preschool group reaches good quality. For the three general scales (*KES-R*, *KES-R-E* and *KES-R/R-E Composite*), the preschool groups score a mediocre quality on average. Correlations of the six *KES-R* subscales range from $r = .23$ and $r = .71$. Associations of the four *KES-R-E* subscales are generally higher, ranging from $r = .63$ to $r = .96$. Scale-subscale correlations range from $r = .64$ to $r = .81$ for the *KES-R* and from $r = .82$ to $r = .96$ for the *KES-R-E*. Finally, *KES-R* and *KES-R-E* scales are correlated by $r = .71$, thus forming a close association. Therefore, the *KES-R/R-E Composite* score is used for the regression analyses (see below). Correlations of the *KES-R/R-E Composite* to *KES-R* equal $r = .90$, and to *KES-R-E* $r = .95$. Associations with the 10 subscales range from $r = .52$ to $r = .92$ so that most information in the data is retained. Using the *KES-R/R-E Composite* score, most preschool groups show a mediocre quality ($n = 45$, 75%), and only four groups (7%) score above 5, thus showing a good quality. For 11 preschool groups (18%), the quality has to be termed poor.

NEPS Questionnaire Data

Description of the questionnaire data is ordered concerning the subdivision shown in section 3.2. Due to space limits, we only show variables that are used in the subsequent regression analyses. Selection was done by examining the bivariate correlations between all variables and due to content-related considerations. Moreover, the exact mathematical derivation of the variables from the items is not discussed in detail.

The following *structural characteristics* are considered:

- 1) *Opening Hours* (Monday to Friday): This is simply the sum of the daily opening hours. The score varies from 33 to 63 hours per week, with a mean of $M = 44.53$ ($SD = 4.50$; $n = 59$).
- 2) *Weekend Opening*: This is a dichotomous variable. The teachers of 4 of the 60 preschool groups (6.7 %) state that their preschool is open on the weekends.
- 3) *Vacation Closure*: Here, the weeks per year that the preschool is closed are specified. This variable ranges from 0 to 6 weeks, with a mean of $M = 3.95$ ($SD = 1.23$; $n = 58$).
- 4) *Group Size*: One teacher stated the group size to be 49 children. Because this clearly points to open group work, this group was not used in the regression analyses. The variable varies between 12 and 28 children per group, with a mean of $M = 23.53$ ($SD = 3.04$; $n = 59$).
- 5) *Room Number*: Most of the groups use one room ($n = 17$, 28.3 %) or two rooms ($n = 20$, 33.3 %), and only 12 groups use more rooms (20.0 %). There is no information available for 11 groups (18.3 %).
- 6) *Room Size*: The size of the available rooms varies from 28 to 431 sq. m, with a mean of $M = 78.93$ ($SD = 63.43$; $n = 45$). Please note the large standard deviation and number of missings.
- 7) *Working Time*: This variable consists of the full-time equivalent of the first and second teacher in the group. It varies from 86 % to 200 %, with a mean of $M = 167.10$ ($SD = 34.27$; $n = 58$).

Next, we display the *compositional characteristics*. In general, a bit more computation was necessary for these variables. Because the variable *Count of Children with Migration Background* did not vary sufficiently in the sample and is highly skewed ($M = 4.18$; $Md = 1$; $SD = 7.69$; $n = 60$), it was skipped in the analyses. 60 % of the preschool groups are attended by no or only one child with migration background.

- 1) *Count of Children with Disabilities*: Here, all children in a group with disabilities or developmental disorders were added. The variable is a composite of several items of the questionnaire. Most groups are not attended by these children ($n = 47$, 78.3 %). In the remaining groups, the count varies between 1 and 12. Therefore, the mean is relatively small, with $M = 0.58$ ($SD = 1.80$; $n = 60$).
- 2) *Average Age*: The average age of the children varies between groups from 43.5 months to 61.68 months, with a mean of $M = 55.11$ ($SD = 3.43$; $n = 54$).
- 3) *Age Variability*: Here, simply the standard deviation of the age in months of the children in the group was computed. It varies between 6 and 16 months, with a mean of $M = 11.88$ ($SD = 2.21$; $n = 54$). When average age and age variability are taken together, there is an indication that the preschool groups are aged-mixed.

- 4) *Average Hours of Care*: This variable was computed by summing up the hours of care for every child in a group divided by group size. The score ranges between 4.69 and 7.74 hours per day, with a mean of $M = 6.28$ ($SD = 0.82$; $n = 46$).
- 5) *Balanced Gender Ratio*: This variable is derived to reflect an equal count of boys and girls in the group because a scatterplot of the boy-girl ratio with *KES-R/R-E Composite* showed an inverted U-shaped curve with a maximum at 1 (corresponding to an equal count of boys and girls). The variable is computed by the absolute value of 1 minus the count of boys divided by count of girls. Thus, the larger the value, the more unbalanced the gender ratio in the group is. In contrast, a 0 reflects an equal count of boys and girls. Thus, a negative correlation coefficient reflects a positive association.

Concerning *materials and activities*, only a few variables can be used. For activities, an insufficient variability was realized in the study, leading to an adapted response scale in the NEPS main studies. In the following section, activities are excluded. Only three variables are treated:

- 1) *Materials*: The scale is a mean score of 14 items using a three-point response scale with values from 1 (materials are available for some children) to 3 (materials are available for all children). An indication that materials are not available at all is scored 0. The variable varies between 0.53 and 3.00, with a mean of $M = 1.52$ ($SD = 0.40$; $n = 60$).
- 2) *Visits (Museum)*: Teachers had to indicate how often such visits were offered to the children. The response scale ranges from 1 ("never") to 6 ("(nearly) daily"). In the sample, only values from 1 to 4 ("(nearly) monthly") were used. The mean equals $M = 1.71$ ($SD = 0.80$; $n = 58$).
- 3) *Visits (Theatre, Cinema, Concert)*: Using the same scale and realizing the same range, the mean of this variable equals $M = 2.39$ ($SD = 0.65$; $n = 57$).

Finally, the following *teacher characteristics* were considered:

- 1) *Years of Education*: 31 teachers (51.7%) left school after Grade 10, 24 teachers (40.0%) after Grade 12, and 5 teachers (8.3%) after Grade 13.
- 2) *Work Experience*: The total count of years in work was specified. The variable varies between 2 and 41 years, with a mean of $M = 17.65$ ($SD = 9.75$; $n = 57$).
- 3) *Contractual Working Hours*: The teachers reported contractual working hours between 19.5 and 40 hours a week. Mean equals $M = 34.59$ ($SD = 6.87$; $n = 60$).
- 4) *Overtime*: The difference between the factual and the contractual working hours was computed. It ranges from 0 to 11 hours per week, with a mean of $M = 2.40$ ($SD = 3.29$; $n = 56$).
- 5) *Hours of Work with Children*: The direct work with the children in the group varies between 0 and 38 hours per week, with a mean of $M = 24.85$ ($SD = 9.50$; $n = 56$).

- 6) *Hours of Work without Children*: This variable includes preparation time, team meetings, and other work. It ranges from 1 to 37 hours per week, with a mean of $M = 9.15$ ($SD = 6.66$; $n = 59$).
- 7) *Advanced Training*: This is simply a dichotomous variable. The item asks about additional certified training. 9 teachers (15%) affirmed this question, and 40 teachers (66.7%) negated it. 11 answers (18.3%) are missing.
- 8) *Number of Types of Further Education*: Teachers had to indicate how many courses of further education with different content they had attended in the last year. The answers vary between 0 and 11, with a mean of $M = 2.90$ ($SD = 2.05$; $n = 60$).
- 9) *Hours of Further Education*: Here, the total number of hours spent on further education in the last year was computed. This variable ranges from 0 to 171 hours a year, with a mean of $M = 28.70$ ($SD = 29.24$; $n = 60$).
- 10) *Supervision*: This dichotomous variable indicates the availability of supervision for the teachers. In the sample, only 7 teachers (11.7%) received supervision, whereas the majority did not ($n = 53$; 88.3%).

4.2 Regression Analyses

Because of the relatively small sample size, regression analyses are used in a more descriptive way. In the following section, separate regression analyses of the *KES-R/R-E Composite* are performed as criteria for the four domains of questionnaire data. Starting with the complete bunch of predictors of one domain, a stepwise procedure is employed to reduce the set to the most predictive characteristics. Instead of using an algorithm that is implemented in a statistical package, a selection of variables is done with the following steps: (1) collinearity diagnostics (using condition index, CI), (2) selection of variable(s) to be excluded (besides the current criteria, tolerance and variance inflation factor, correlation of predictors, and theoretical criteria are considered), and (3) regression with reduced set until the condition index falls under 30 (cf. Bühner, & Ziegler, 2009). After selecting the most predictive variables for each of the four domains, an overall regression series is conducted using all selected variables. For all analyses, only additive models are estimated, and no interactions between predictors are added.

Regression of KES-R/R-E Composite on structural characteristics

Table 2 shows the stepwise regression on the *structural characteristics*. The outcome variable is the *KES-R/R-E Composite*.

In the final step, two variables remain that explain 12% of the variation of the *KES-R/R-E Composite* (using the corrected explained variance, R^2_{corr}). *Working Time*, that is, the time spent by one or two teachers within the group, and *Opening Hours*, that is, hours the preschool can be attended on weekdays.

Table 2 Regression of KES-R/R-E Composite on structural characteristics

	β_1	β_2	β_3	β_4	β_5	β_6
Working Time	.44	.26	.26	.29	.28	.30
Opening Hours	.21	.30	.30	.17	.20	.18
Weekend Opening	.19	.17	.17	.14	.13	
Room Size	-.49	-.13	-.13	-.09		
Vacation Closure	.21	.27	.26			
Group Size	.02	.01				
Room Number	.46					
<i>R</i>	.54	.47	.47	.42	.41	.39
<i>R</i> ²	.29	.22	.22	.18	.18	.15
<i>R</i> ² _{corr}	.15	.10	.12	.09	.12	.12
<i>CI</i>	60.7	56.4	47.5	35.2	28.9	28.2

Regression of KES-R/R-E Composite on compositional characteristics

Table 3 shows the stepwise regression on the *compositional characteristics*. Within three steps, three variables are selected that explain 22% of the variance of the *KES-R/R-E Composite*. Whereas *Average Hours of Care* and *Balanced Gender Ratio* contribute positively to the overall quality of the preschool group, the more *Children with Disabilities* are taken care of, the lower the *KES-R/R-E Composite* is.

Table 3 Regression of KES-R/R-E Composite on compositional characteristics

	β_1	β_2	β_3
Average Hours of Care	.47	.47	.43
Balanced Gender Ratio	-.19	-.19	-.18
Count of Children with Disabilities	-.28	-.28	-.25
Average Age	-.14	-.14	
Age Variability	.00		
<i>R</i>	.54	.54	.52
<i>R</i> ²	.29	.29	.27
<i>R</i> ² _{corr}	.20	.22	.22
<i>CI</i>	52.5	46.4	18.9

Regression of KES-R/R-E Composite on materials and activities

Table 4 shows the stepwise regression on the *materials and activities* of the *KES-R/R-E Composite*. Using two of the three variables (*Materials* and *Visits (Museum)*), 4 % of the variance of the *KES-R/R-E Composite* can be explained.

Regression of KES-R/R-E Composite on teacher characteristics

Table 5 shows the stepwise regression on the *teacher characteristics* of the *KES-R/R-E Composite*. After a stepwise reduction of the set of 10 predictors, 5 variables remain that predict 12 % of the variance of the *KES-R/R-E Composite*. These teacher characteristics include *Hours of Further Education*, *Advanced Training*, *Years of Education*, *Hours of Work without Children*, and *Supervision*.

Overall regression of KES-R/R-E Composite

Finally, the *KES-R/R-E Composite* was regressed on the selected variables of the four domains. The same stepwise procedure as described above was applied to the 12 predictors. Table 6 shows the results.

In Step 8, five variables remain as the most sufficient predictors explaining 31 % of the variance of the *KES-R/R-E Composite*. Structural characteristics (*Opening Hours*), compositional characteristics (*Average Hours of Care*, *Count of Children with Disabilities*), and teacher characteristics (*Hours of Further Education*, *Supervision*) are included. Materials and activities do not reach a meaningful influence. All selected variables are of nearly equal importance (in relation to the standardized regression coefficient).

As shown in Figure 1, the overall quality of the preschool groups as measured via observation can be predicted quite well by the five selected variables of the teacher questionnaire. There are only a few cases that are severely over- or underestimated. Multiple correlation coefficient equals $R = .62$ (see Table 6, last column, as well as the adjusted regression line in Figure 1). Dotted lines in Figure 1 relate to the thresholds of the ECERS-Scales, indicating poor quality (below 3), good quality (above 5), and mediocre quality (between 3 and 5). Observed and predicted quality is mediocre in most cases.

Table 4 Regression of KES-R/R-E Composite on materials and activities

	β_1	β_2
Materials	.12	.17
Visits (Museum)	.16	.19
Visits (Theatre, Cinema, Concert)	.17	
<i>R</i>	.32	.27
<i>R</i> ²	.10	.08
<i>R</i> ² _{corr}	.05	.04
<i>CI</i>	11.1	9.4

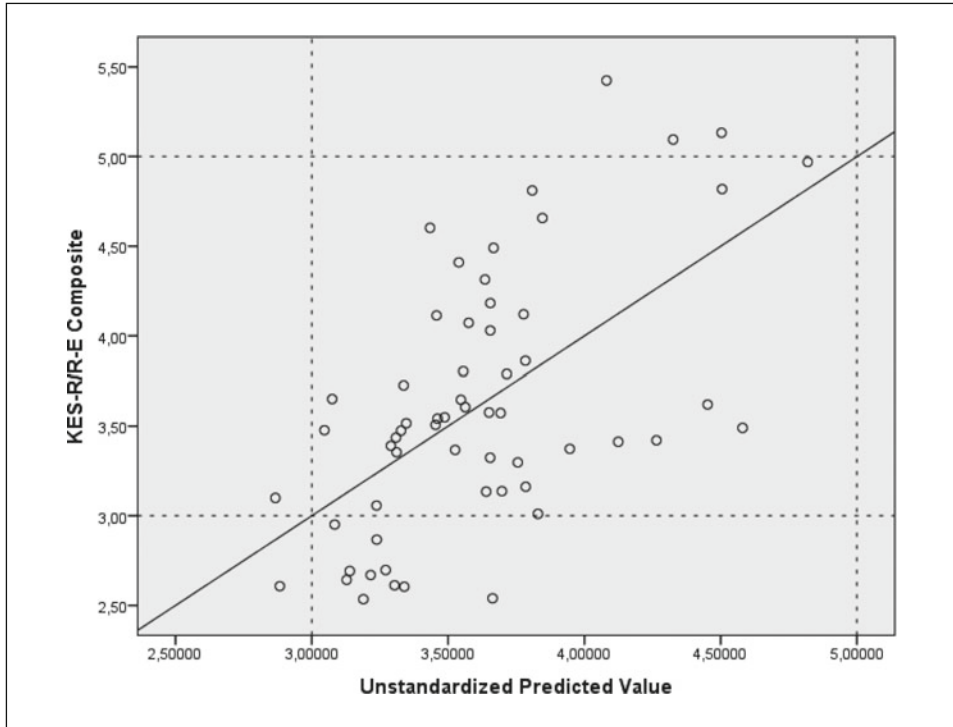
Table 5 Regression of KES-R/R-E Composite on teacher characteristics

	β_1	β_2	β_3	β_4	β_5	β_6
Hours of Further Education	.27	.27	.21	.24	.26	.26
Advanced Training	.17	.17	.18	.18	.19	.18
Years of Education	.13	.13	.17	.19	.17	.18
Hours of Work without Children	.05	.04	.02	.14	.14	.13
Supervision	.19	.19	.16	.17	.16	.16
Work Experience	-.03	-.03	-.03	-.09	-.10	
Hours of Work with Children	.03	.03	-.00	.10		
Contractual Working Hours	.24	.25	.25			
Overtime	-.15	-.15				
Number of Types of Further Education	.03					
<i>R</i>	.53	.53	.51	.48	.47	.46
<i>R</i> ²	.28	.28	.26	.23	.22	.21
<i>R</i> ² _{corr}	.06	.09	.10	.09	.10	.12
<i>CI</i>	29.4	27.8	26.5	21.9	18.3	16.1

Table 6 Overall regression of KES-R/R-E Composite

	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8
Opening Hours	.34	.34	.30	.30	.33	.31	.29	.24
Average Hours of Care	.26	.26	.29	.28	.32	.25	.26	.32
Count of Children with Disabilities	-.23	-.23	-.23	-.24	-.26	-.24	-.24	-.27
Hours of Further Education	.30	.30	.31	.32	.35	.31	.33	.32
Supervision	.21	.21	.22	.24	.27	.26	.26	.29
Advanced Training	.15	.15	.19	.21	.18	.15	.17	
Materials	.11	.11	.07	.08	.08	.08		
Visits (Museum)	-.15	-.15	-.18	-.20	-.18			
Working Time	.12	.12	.15	.17				
Years of Education	.11	.11	.13					
Balanced Gender Ratio	-.15	-.15						
Hours of Work without Children	.00							
<i>R</i>	.70	.70	.69	.68	.66	.64	.64	.62
<i>R</i> ²	.49	.49	.48	.46	.44	.41	.41	.39
<i>R</i> ² _{corr}	.25	.28	.28	.29	.28	.28	.30	.31
<i>CI</i>	60.0	57.5	55.1	50.5	46.9	42.8	35.3	33.3

Figure 1 Scatterplot of the unstandardized predicted value of the overall regression with five predictors of *KES-R/R-E Composite*



5 Discussion

Before discussing the results, some limitations of the study should be mentioned. Due to the small sample size, we did not use an algorithm to exploit the significance of regression coefficients for the selection of variables, but instead reduced the variable set with collinearity diagnostics. For sake of clarity, regression analyses were performed separately for the four domains of predictors. Additionally, interactions between predictors were not included. All this could lead to a suboptimal selection of variables. However, the aim of this paper is mainly the comparison of the observation and questioning of preschool quality. Therefore, the selection of the best variables of the teacher questionnaire for the prediction of observed quality is the main purpose. We will analyze the data more deeply in a forthcoming NEPS Working Paper. Nevertheless, it should be kept in mind that the study is basically explorative and is in need of replication of the results. Another limitation relates to the selection of the preschool groups, both concerning the regional context as well as the pedagogical approach. Further-

more, the selection of two groups in some preschools may lead to dependencies in the data that were not accounted for in the analyses. Generalization of the results should therefore be done very carefully. As for the observed quality (i. e., *KES-R/R-E Composite*) as the criteria of the regression analyses, it has to be noted that variability is limited. The upper as well as lower 1.5 points of the theoretical scale are missing, and most preschools score in the middle of the distribution. This may lead to restricted regression coefficients and an underestimation of the real associations. Moreover, the discrimination of the preschool groups by quality may be impaired due to the shortened variability.

Bearing these limitations in mind, we discuss the results in the following section, starting with the regression on the *structural characteristics*. *Working Time* and *Opening Hours* were the most predictive variables. Interpretation is straightforward: The more that time for education and care of children on both levels of preschool and for teachers' working hours is available, the more that educational quality is actualized. There is one limitation to this interpretation: Computing the variable *Working Time*, only the first and second teacher were included although the full-time equivalent of up to four teachers was accounted for in the questionnaire. This is due to the fact that the full-time equivalent of the third teacher did not correlate with the *KES-R/R-E Composite*, and the full-time equivalent of the fourth teacher was even negatively associated with preschool quality. We interpret this result with the psychological need of pre-school children for a definite attachment figure, which may be missing if more than two teachers share the group. Another possibility is that the need for a third or fourth teacher gives hints that there are problems within the group that also reduce the quality as a whole. Perhaps surprisingly, *Group Size* does not yield any effect.

Concerning *compositional characteristics*, three out of five variables contribute to the prediction of preschool quality. Whereas *Average Hours of Care* and *Balanced Gender Ratio* have a positive association with *KES-R/R-E Composite*, *Count of Children with Disabilities* correlates negatively. The positive effect of *Average Hours of Care* can be interpreted as before: the time of education and care (now on an individual level) for each child contributes to positive educational quality. The positive effect of *Balanced Gender Ratio* is interpreted as follows: An appropriate amount of diversity (in this case concerning gender) is associated with better quality. This is also true for *Count of Children with Migration Background*. Although this variable could not be used in the regression analyses because of its reduced variability, it shows a quadratic relation to the *KES-R/R-E Composite*. That is, a certain number of these children lead to better preschool quality than more or fewer children with a migration background in the group. The maximum of the regression line is at 30% of children with a migration background, which is in line with previous studies. Unfortunately, this interpretation does not hold for children with disabilities. The negative linear correlation shows that each additional child with disabilities in a group is followed by an impaired quality. This daunting result should in no way be interpreted causally as a direct effect of the presence of such children to the impairment of educational qual-

ity. Rather, we propose the following interpretation, which is in line with what has already been stated: A child with disabilities is certainly in need of more attention from the teacher than a child without disabilities. This reduces the total amount of education and care time for the group as a whole as well as for every (other) child, thus leading to an overall reduced evaluation of preschool quality. This could also be the case with children with a migration background. Once their count exceeds an appropriate ratio within a group, more attention of teachers is needed and less time is available for the rest of the children and the whole group. Additionally, this may generate the need for more teachers to care for the group, contributing to the negative correlation of the fourth teacher to preschool quality (see above).

Looking at *activities and materials*, the assessment of these variables was not successful in this study or the NEPS pilot study, leading to an adapted response scale in the NEPS main study. As a consequence, we cannot make any inferences on the significance of activities for (observed) preschool quality, although their theoretical importance is beyond doubt. Nevertheless, the availability of *Materials* and *Visits to Museums* could explain 4% of the *KES-R/R-E Composite*. This result can be interpreted quite easily: The more stimulation (related to education) the children receive, the higher the preschool quality is.

Not surprisingly for a teacher questionnaire, *teacher characteristics* provides the most variables. In the regression analysis, five out of ten variables add to the explained variance of 12%. The interpretation of the contribution of *Hours of Further Education*, *Advanced Training*, *Years of Education*, *Hours of Work without Children*, and *Supervision* is straightforward: The better the qualification of the teacher and the higher the effort for a better qualification by the teacher are, the higher the preschool quality is. This interpretation is in line with measures of professionalization of preschool teachers' who have been training in Germany within the last years.

The overall regression does not change anything for the aforementioned interpretation except that it selects the most significant predictors. These predictors come from the different domains of the teacher questionnaire (except *activities and materials* due to their impaired assessment) and are of nearly equal importance, thus pointing to evidence that there is no single indicator of preschool quality that could be asked for. The overall interpretation of the results refers to two major aspects: (1) the hours of education and care a preschool group and every single child receives and (2) the quality of teachers' education. The more time a well-trained teacher spends with the children he or she is responsible for, the better the educational quality is. Furthermore, there are hints that diversity in different aspects (e.g., gender distribution, children with migration background, and a variety of materials and activities) could contribute to the educational quality of a preschool group. The results of the overall regression also show that the observed global educational quality can be reproduced quite well with questionnaire data. There is no massive over- or underestimation, and the main purpose of this paper can thereby be viewed as having been fulfilled.

The results of this study, together with the results of the NEPS pilot study, have led to improvements of the questionnaire design of the NEPS main study. Response scales have been adapted, question texts have been altered, and in general, the most promising questions have been selected. Therefore, we are hopeful that the NEPS Scientific Use File of Starting Cohort 2—Kindergarten contains comprehensive data on preschool quality. The study clearly demonstrates that no single indicator of preschool quality can be expected from questionnaire data but that a vast number of variables should be considered that also have to be derived with more or less sophistication from the original questions. All variables in question are not indicators of quality per se, but they all contribute to quality to a certain extent. In this context, we wish to stress that causal inferences of these variables for preschool quality or for further outcomes like children's competencies or competence development should be drawn very carefully. For example, the negative relation of *Count of Children with Disabilities* to *KES-R/R-E Composite* does not mean that these children diminish the educational quality of the group or hinder other children in their development but rather shows that there are special demands that are not (yet) being met in the everyday work of the preschool group. All variables used in this study—along with a number of additional items—have been implemented in the NEPS main studies involving preschools. First descriptive results are presented by Linberg, Bäumer, and Roßbach (2013).

In a nutshell, it is possible to draw conclusions on educational quality using questionnaire data even from a single contributor who might be under suspicion of euphemizing the facts. However, it is important to keep in mind that all variables can only give hints about quality and should be derived and interpreted with caution. With this in mind, the NEPS provides a rich data resource for the analysis of the educational quality of preschool institutions and will contribute to highly valuable scientific insights in this issue.

References

- Anders, Y. (2013). Stichwort: Auswirkungen frühkindlicher, institutioneller Bildung und Betreuung. *Zeitschrift für Erziehungswissenschaft*, 16(2), 237–275.
- Bäumer, T., Preis, N., Roßbach, H.-G., Stecher, L., & Klieme, E. (2011). Education processes in life-course-specific learning environments. In H.-P. Blossfeld, H.-G. Roßbach, & J. von Maurice (Eds.), *Zeitschrift für Erziehungswissenschaft*, 14. *Education as a lifelong process: The German National Educational Panel Study (NEPS)* (pp. 87–101). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Bäumer, T., & Roßbach, H.-G. (2012). Die Familie macht's. Die Bedeutung der Familie für kindliche Bildungsprozesse. *DJI Impulse*, 100(4), 39–41.
- Bühner, M., & Ziegler, M. (2009). *Statistik für Psychologen und Sozialwissenschaftler*. München: Pearson Studium.

- Clausen, M. (2002). *Qualität von Unterricht—Eine Frage der Perspektive?* Münster: Waxmann.
- Harms, T., Clifford, R. M., & Cryer, D. (1998). *Early childhood environmental rating scale, revised edition (ECERS-R)*. New York: Teachers College Press.
- Linberg, T., Bäumer, T., & Roßbach, H.-G. (2013). Data on early child education and care learning environments in Germany. *International Journal of Child Care and Education Policy*, 7(1), 24–42.
- Roßbach, H.-G. (2004). Kognitiv anregende Lernumwelten im Kindergarten. In J. Baumert, D. Lenzen, R. Watermann, & U. Trautwein (Hrsg.), *Zeitschrift für Erziehungswissenschaft, 7. PISA und die Konsequenzen für die erziehungswissenschaftliche Forschung: The German National Education Panel Study (NEPS)* (S. 9–24). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Roßbach, H.-G. (2005). Effekte qualitativ guter Betreuung, Bildung und Erziehung im frühen Kindesalter auf Kinder und ihre Familien. In Sachverständigenkommission Zwölfter Kinder- und Jugendbericht (Ed.), *Bildung, Betreuung und Erziehung von Kindern unter sechs Jahren* (S. 55–174). München: Verlag Deutsches Jugendinstitut.
- Roßbach, H.-G., Klucznik, K., & Kuger, S. (2008). Auswirkungen eines Kindergartenbesuchs auf den kognitiv-leistungsbezogenen Entwicklungsstand von Kindern. In H.-G. Roßbach, & H.-P. Blossfeld (Eds.), *Zeitschrift für Erziehungswissenschaft, 11. Frühpädagogische Förderung in Institutionen: The German National Education Panel Study (NEPS)* (S. 139–158). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Roßbach, H.-G., & Tietze, W. (in prep). *Kindergarten-Skala Erweiterung (KES-E)*. Unpublished manuscript.
- Sylva, K., Melhuish, E., Sammons, P., & Taggart, B. (2004). *The Effective Provision of Pre-School Education (EPPE) Project: Final Report. A longitudinal study funded by the DfES 1997–2004*. London: University of London, Institute of Education.
- Sylva, K., Siraj-Blatchford, I., & Taggart, B. (2003). *Assessing Quality in the Early Years: Early Childhood Environment Rating Scale Extension (ECERS-E): Four curricular subscales*. Stoke on Trent: Trentham Books.
- Tietze, W., Meischner, T., Gänsfuß, R., Grenner, K., Schuster, K.-M., Völkel, P., & Roßbach, H.-G. (1998). *Wie gut sind unsere Kindergärten? Eine Untersuchung zur pädagogischen Qualität in deutschen Kindergärten*. Neuwied: Luchterhand.
- Tietze, W., Schuster, K.-M., Grenner, K., & Roßbach, H.-G. (2005). *Kindergarten-Skala Revidierte Fassung (KES-R)* (3. überarbeitete Aufl.). Weinheim: Beltz.

About the authors

T. Bäumer

Leibniz Institute for Educational Trajectories (LifBi), Bamberg.

e-mail: thomas.baeumer@lifbi.de

H.-G. Roßbach

Leibniz Institute for Educational Trajectories (LifBi), Bamberg.

Chair of Early Childhood Education, University of Bamberg, Bamberg.

Reading-Aloud Versus Self-Administered Student Questionnaires: An Experiment on Data Quality

Cornelia Gresch, Rolf Strietholt, Michael Kandera and Heike Solga

Abstract

A major finding from recent large-scale assessments on student achievement is that a remarkable proportion of students around the world are poor readers. This calls into question the quality of the data retrieved from self-administered background questionnaires. A better administration mode, especially for this student population, might be to have the administrator read the questionnaires out aloud, as is sometimes done in surveys at elementary schools. In order to provide empirical evidence on whether reading aloud helps improve data quality, we conducted an experimental study with 664 twelve-year-old students in lower secondary schools in Germany. One finding is that, unsurprisingly, reading questionnaires aloud increases survey time. Regarding data quality, however, item non-response rates decrease somewhat in the reading-aloud group, and filtering procedures also work better. This effect can be found regardless of students' status or reading speed. Regarding the acceptance of the mode, analyses on the role of migrant status and reading speed suggest that slow readers and migrant students particularly prefer being read the questionnaires aloud. Our study indicates that reading questionnaires aloud may be a meaningful administration mode not only in early primary school grades, but also at the beginning of secondary school. Data quality in studies involving at-risk students can particularly benefit from reading questionnaires aloud.

1 Introduction

In large-scale assessments of student achievement, students not only work on tests but also respond to background questionnaires. Whereas the questionnaires in studies targeting older students are mostly self-administered, the test administrator reads

out the questionnaires to the whole class in lower grades and in some special schools. The basic idea behind this procedure is that these students may have difficulties reading the questionnaires on their own because they lack sufficient reading proficiencies and cannot concentrate on all survey questions. However, a major finding from recent large-scale assessments of student achievement is that a remarkable proportion of students around the world are poor readers, not only in the early grades but also at the end of primary school and even at the end of secondary school (Mullis, Martin, Kennedy, & Foy, 2007; OECD, 2010). This finding raises concerns about the quality of data collected from secondary-school students because large-scale assessments are usually self-administered, meaning that these students read and fill out the background questionnaires on their own. The importance of language has been studied in the context of student assessments and survey research. However, even though survey experts have repeatedly emphasized the importance of wording and have suggested, for example, using simple syntax and familiar terms (e.g. Bradburn, Sudman, & Wansink, 2004), many questionnaires are still quite demanding. The field of test development faces similar challenges. Here, test developers have suggested making linguistically demanding tests like mathematical word problems more accessible by reading them aloud. Experimental studies indicate that poor readers, in particular, can benefit from this test administration practice (e.g., Meloy, Deville, & Frisbie, 2002; Randall & Engelhard, 2010; Wolf, Kim, Kao, & Rivera, 2009). In other words, students' test responses are not adversely affected by complex language when tests are read aloud to them. Consequently, these test results are more valid measures of the construct being assessed.

With respect to background questionnaires, however, we are not aware of any experimental research on how reading questionnaires aloud to students affects data quality. As a consequence, a number of key questions regarding the survey modes remain unanswered: What are the effects of survey modes on certain data-quality aspects, such as missing data, filter questions, and item responses? What are the consequences regarding interview time? Are there any substantial differences in the findings that question the comparability of surveys employing different modes? How do students and test administrators handle and accept the different survey modes? How does the survey mode affect certain groups of students, such as poor readers in higher grades, who are often referred to as "functional illiterates" or at-risk students? Even though a whole strand of research compares different modes of data collection in terms of different criteria within the field of survey methodology (e.g., de Leeuw, 2008; Dillman & Christian, 2005; Groves & Lyberg, 2010; Krosnick, 1999; Schwarz et al., 1991), we are not aware of any study that answers these questions in terms of large-scale assessments in which students are surveyed in clusters of classes or schools.

2 Purpose

The present study investigates how reading questionnaires aloud to students affects data quality as compared with using self-administered questionnaires. We present findings from an experimental study in which we randomized Grade-5 students into two treatment conditions depending on the administration of the background questionnaire. The main aim of this paper is to examine the effects of the different modes by focusing on four aspects of data quality: Patterns of missing values, filter questions, interview duration, and the level of acceptance by students and test administrators. Findings from this study will offer some guidance for choosing a suitable questionnaire administration mode in future surveys.

3 Research Design and Method

3.1 Data

This study is associated with a test development study that is part of the National Educational Panel Study (NEPS) in Germany (Blossfeld, Roßbach, & von Maurice, 2011).¹ About 50 schools from the lower secondary track (*Hauptschule*) in four German states were invited to participate. In most German states, tracking begins after Grade 4, and students in the *Hauptschule* track generally perform lowest in large-scale tests (Naumann, Artelt, Schneider, & Stanat, 2010). The four states represent different geographical parts of Germany, ranging from rural to urban areas. 27 schools were accepted to participate with a total of 664 fifth-grade students (323 girls and 341 boys with a mean age of 12.3 years, standard deviation of 0.9 years). Fieldwork began in November 2011 with a total duration of three months.

3.2 Procedures

At each school, we randomly assigned participants to one of the two conditions of questionnaire administration. Each group was tested and surveyed on two days. On both days, students first participated in achievement tests (of about 80 minutes). After a 15-minute break, the test administrator handed out the background questionnaires and introduced the various item formats before students began filling out the questionnaires.

1 The NEPS data collection is part of the Framework Programme for the Promotion of Empirical Educational Research, funded by the German Federal Ministry of Education and Research and supported by the Federal States.

On Day 1, the test administrator in the first group told the students to complete the questionnaire individually. The administrator in the second group read the questionnaires aloud to students item by item using a standardized script. On Day 2, the two conditions were reversed: The questionnaire was read out aloud to students in the first group, whereas students in the second group completed the questionnaires on their own. In this way, all students were exposed to both treatments (but in a different order), which enabled us to compare both modes for each student. The administrators were trained in an all-day workshop how to administer the survey and how to read out the questionnaire.

After 15 minutes, all students were requested to stop filling out the questionnaires. Thereafter, all students worked on an additional questionnaire about their personal preferences and experience with each of the two administration modes. This seemed necessary because students only had 20 minutes to fill out the questionnaire, and we wanted to ensure that each student answer these final rating questions, which were read aloud to all students in both groups.

3.3 Instruments

Questionnaire

The instruments are based on a short version of the NEPS questionnaire for fifth-graders at special schools (see Heydrich, Weinert, Nusser, Artelt, & Carstensen, 2013). They cover items on student background, grades, reading habits, and self-esteem. The questionnaire from Day 1 covered 19 questions (Day 2: 15 questions), scales, and different kinds of response sets, which added up to a total of 46 single items (Day 2: 52 items).

Survey on the mode (students/administrators)

At the end of the survey, all students were asked about their perception of the questionnaire, for instance, regarding any difficulties in understanding and their personal preferences regarding the survey modes. For example, we asked: "A questionnaire can be read out aloud by the test administrator, or it can be filled out independently. Which mode do you prefer?" The test administrators were also asked about their perceptions of the testing session. Here, we asked for personal preferences regarding the administration mode (reading aloud or self-administration) and for some statements on the mode currently used.

Reading speed

We used an NEPS reading speed test as an indicator of how well students read (NEPS, 2011). The test comprised 51 short statements (e.g., "mice can fly") that students had to read and decide whether they were correct or incorrect. We used the number of sentences assessed correctly within 2 minutes as a measure of reading speed. For the

current study, we distinguish between slow readers (lower quartile), regular readers (2nd and 3rd quartiles), and fast readers (upper quartile). Due to the nature of our sample (lower secondary-school students), many of the students had low reading competencies and also scored low on reading speed.

Timekeeper

In both conditions, the test administrator asked the students to mark the item they were currently working on in 5-minute intervals (Minutes 5, 10, and 15). Furthermore, the test administrators recorded the overall survey time. This provided us with information on survey time.

Migrant status

We distinguished between students without a migrant background (both parents born in Germany), the 2.5 generation (one parent born in Germany, the other born in a foreign country), the 2nd generation (student born in Germany, but both parents born in a foreign country), and the 1st generation (both student and parents born in a foreign country). Thirty-five students had to be excluded from the analysis due to item nonresponse on the migrant status variables.

4 Results

We focus on patterns of missing values and item-response patterns as key indicators of data quality and analyze how students answer filter questions. In addition, we differentiate by students' reading speed and migration background, investigating how these factors impact on data quality in each of the two different survey modes. As motivational factors for data quality, we compare the personal preferences of students and administrators. Finally, we present some findings on the time needed to complete the questionnaire in each of the two modes.

4.1 Item Nonresponse

Patterns of missing values are a good indicator of data quality. However, before analyzing these patterns, we need to make a general distinction. Missing values can have different sources: They can occur due to unit nonresponse (i. e., the respondent does not participate in the interview at all), dropout during the interview (e. g., the respondent runs out of time), or item nonresponse (e. g., due to the respondent's sloppiness or uncertainties while answering these questions). We focus on item nonresponse in particular, which is defined as missing values that do not result from dropout or filter questions. However, since we had a high dropout rate during the interview, we also present findings on this aspect in our analyses. Figure 1 displays the missing patterns

for the two conditions on the first day. The bars represent the proportion of missing values for each single item (bars), the dots in the upper graph display the proportion of students who dropped out at this question without answering any further questions, and the line in the upper graph shows the proportion of students who dropped out before this question.

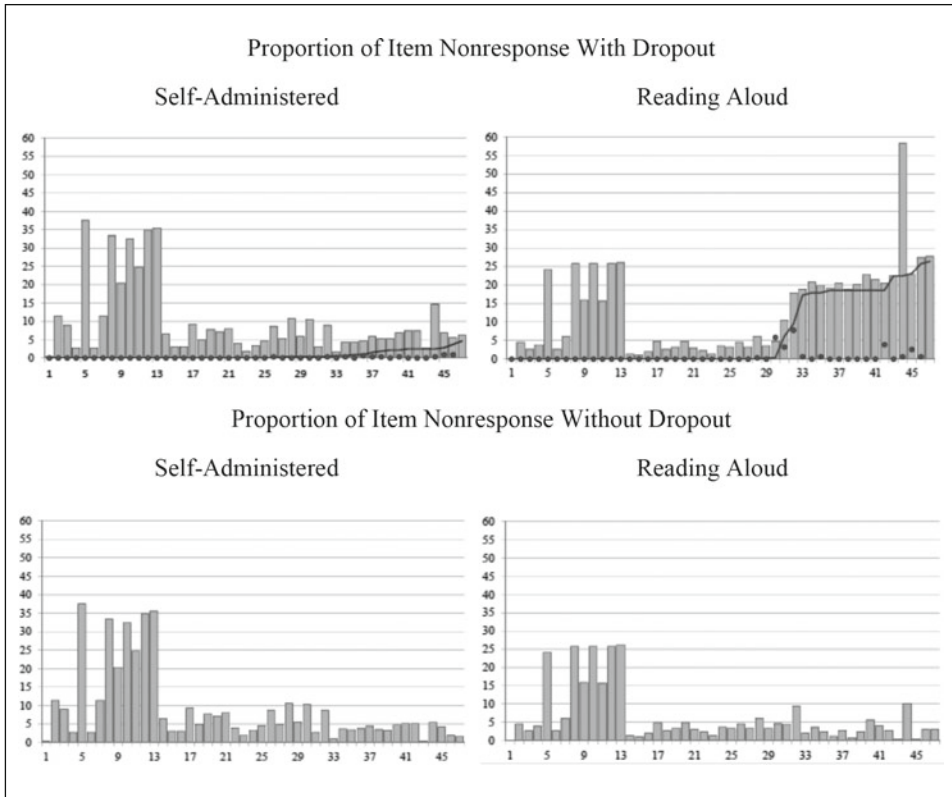
Figure 1 reveals two general “irregularities” in the sense of eye-catching peaks in the distribution of missing patterns. These can be found in both the self-administered and the read-aloud condition. First, we see that Questions 8 to 13 have a rather high proportion of missing values (between 20 % and 35 %) in both conditions. These are questions with multiple answering possibilities. Students were asked to provide information on the composition of their household, for example, if they lived together with their biological mother (yes/no), with their stepmother (yes/no), with siblings (yes/no), with their grandmother (yes/no), and so on. Since most students lived together with their biological parents and without grandparents, the high proportion of missing values results from students who did not tick off the “does not live in the household”-category for these additional people.

The second “irregularity” is that there were some filter questions that also led to systematic peaks (Questions 5 and 44). These questions were filtered to a very small proportion of students, who were also part of a highly selective group, for example, students who had moved to Germany (Question 5; 5 % of all students) or had repeated one year of school (Question 43; 10 % of all students). In the reading-aloud condition, the number of students who answered these questions was even lower than the number of dropouts at this point, resulting in a high nonresponse rate of 58 %. However, missing values in questions featuring multiple answers or filter questions are a common problem in student surveys. For a clear interpretation of students’ responses, it is therefore essential that they answer these questions as well as possible (see also Section 4.2).

In addition to these general irregularities, there is also a remarkable peak at the end of the questionnaire in the reading-aloud condition (upper graph). Later, when analyzing the interview time needed, we see that almost one third of all students in the reading-aloud condition were unable to finish the entire questionnaire due to time restrictions. This systematic “dropout” of entire classes can also be found in the general patterns of missing values. It begins at about 5 % of all students after Question 30 and increases to almost 10 % after Question 32. Reading aloud requires more time than the self-administration mode, and classes in the former condition were simply too slow to finish the questionnaire.

However, the systematic dropout of students does not tell us anything about differences in data quality between the two modes. Hence, for item nonresponse, this systematic dropout was taken into account by focusing only on those students who had reached the current question without having dropped out completely. The proportion of missing values among these students is presented in the two lower graphs in Figure 1. Here, the proportion of missing values is slightly lower in the reading-

Figure 1 Proportion of item nonresponse (bars) on Day 1, including student dropout (dots and lines) and excluding student dropout (in %, by survey mode)



aloud condition. On average, about 9.1 % of all items in the self-administered condition have missing values, whereas the corresponding proportion in the reading-aloud condition is about 6.4 %. We tested whether the data confirm our hypotheses that reading aloud decreases the proposition of missing data. A t-test on item level reveals that the observed difference is statistically significant ($N = 47$; $p < 0.05$, one-tailed). This difference might be underestimated.²

2 It is not reasonable to compare the average number of missing values on the respondent level due to the dropout of single respondents and entire classes. Therefore, we calculated the proportion of missing values for each item (if the student reached this question) and compared these average proportions for the two modes. However, the composition of the groups changed at the end of the questionnaire as a result of the dropout: In the self-administered mode, there were fewer slow readers; in the reading-aloud condition, entire classes (including good readers) were excluded. Against this background, the differences between the two modes might have turned out even stronger had all students completed the questionnaire.

A comparison with item nonresponse rates on Day 2 confirms this finding. Again, the average proportion of missing values in the self-administered mode is about 5.7 %, compared with about 2.5 % in the reading-aloud condition; the differences are significant ($N = 32$; $p < 0.001$, one-tailed). Moreover, the total proportion of missing values was lower on Day 2 in both conditions because questionnaires did not feature any of the multiple-answer questions or filter questions that produced the peaks of Day 1.³

4.2 Filter Questions

Item nonresponse is, however, only one aspect of data quality. As a second indicator of how accurately students fill out the questionnaire, we looked at filter questions and two common pitfalls. We examined whether students answered filter questions even though they were (1) *not* supposed to answer them and whether they (2) did not answer the filter questions although they *should* have. On the first day of testing, there were two filter questions. Early on in the questionnaire, students were asked about their country of birth. Afterwards, students were asked, “In case you were *not* born in Germany: How old were you when you moved to Germany?” This wording can be regarded as an implicit filter. The additional question is only to be answered by students who were born in a foreign country. At the end of the questionnaire, there was a second filter question. Students were asked if they had already had to repeat one or more grades in school, and if so, how many times this had happened. Due to the dropout of classes at the end of the interview in the reading-aloud condition, only a smaller proportion of students responded to this second filter. However, the data still provide an impression on the functioning of the filter. The findings are presented in Table 1. All in all, there is a tendency in favor of reading the questionnaire aloud. In this condition, a smaller proportion of students skipped this question even though they should have answered it ($p = .07$). Likewise, there were fewer students who answered this question even though they were born in Germany ($p < .05$).

The second filter confirms this finding, with one exception: There were slightly more students in the reading-aloud condition who failed to answer the question although they were filtered to it. However, the number of cases is diminishingly small, and the differences are not statistically significant ($p = .6$).

3 For the second day of testing, we had to exclude a series of filter questions for immigrant students from the analysis because there was no definite way to identify the respondents who were supposed to answer them.

Table 1 Functioning of the Filter Question by Mode (Absolute Numbers, Percentage in Parentheses)

	Total	Self-administered	Reading aloud
Filter 1: Migration age if not born in Germany			
Students supposed to pass through the filter	40	23	17
Students who <u>failed</u> to answer the question	10 (25%)	8 (35%)	2 (12%)
Total of students who answered filter question	117	73	44
Filter question answered inadequately	73 (62%)	53 (73%)	20 (46%)
Filter 2: Number of repeated grades			
Students supposed to pass through the filter	131	73	58
Students who <u>failed</u> to answer the question	8 (6%)	3 (4%)	5 (9%)
Total of students who answered filter question	139	82	57
Filter question answered inadequately	17 (12%)	13 (16%)	4 (7%)

4.3 The Role of Migrant Status and Reading Speed

We also tested the effects of the two survey modes on item nonresponse with regard to migrant status and reading speed. The correlation between the two is low (see Table 2).

Table 2 Connection Between Migrant Status and Reading Speed (N = 602^a)

Migrant status	Reading speed			N
	Slow readers (Lower quartile)	regular readers (2nd and 3rd quartiles)	Fast readers (Upper quartile)	
No migrant background	19	59	22	421
2.5 generation	40	40	20	35
2nd generation	37	42	20	83
1st generation	21	59	21	63
Total	23	56	22	602

Note. Row percentage of total.

^a Due to item non-response (35 students did not answer the questions on migrant status), the total number of cases differs slightly from the other analysis.

Table 3 Average Proportion of Missing Values (Item Nonresponse) in the Self-administered Mode and the Reading-Aloud Condition (in %)

	Test Day 1 (n _{items} = 47)			Test Day 2 (n _{items} = 32)		
	Self-administered	Reading aloud	Difference ^a	Self-administered	Reading aloud	Difference ^a
All students	9	7	2	6	3	3
Immigrant state						
No immigrant background	8	6	2	5	2	3
2.5th Generation	11	9	2	6	2	4
2nd Generation	10	7	3	8	5	3
1st Generation	10	9	1	5	3	2
Reading speed						
Slow readers	9	8	1	8	5	3
Regular readers	9	7	2	6	2	4
Fast readers	10	5	5	4	1	3

Note. ^a Positive difference: pro reading aloud; negative difference: pro self-administration.

Hence, the two survey modes indicate different reading ability dimensions (German skills and reading speed). Due to the small number of cases, we only analyze the patterns of item nonresponse and not of the filters. The average proportion of item nonresponse is presented in Table 3.

In both conditions, students with a migrant background had a higher proportion of missing values than those without a migrant background. These differences are significant on a .05 percent level for the second generation. However, the differences within the groups remain quite stable in both conditions, although the proportion of item nonresponse decreased when the questionnaire was read out aloud.

The patterns for the different reading-speed groups on the first day of testing in the self-administered condition did not confirm our expectations, though. We had expected fast readers to perform better than slow ones, but instead, both groups offered a similarly poor performance, with the proportion of missing values even slightly higher for fast readers than for slow readers. The differences are not statistically significant, but since we had expected the opposite effect, this finding merits further consideration. Detailed analyses have revealed that multiple-answer questions and filter questions, in particular, produced a remarkably higher item nonresponse rate in the fast-reading group. One explanation might be that fast readers read less precisely

than regular readers and tend to overlook questions more often if they are not that prominent. The items analyzed on the *second day* included neither filter questions nor multiple-answer questions. Here, in the absence of “complex” questions, the pattern occurs as expected: Fast readers had more valid responses than regular or slower readers, as was the case in the reading-aloud condition on Day 1. Overall, all groups showed lower item nonresponse rates under the reading-aloud condition, but the largest difference can be found for fast readers.

4.4 Acceptance of the Different Modes

In addition to item nonresponse and missing values as objective criteria of data quality, we also studied students’ subjective level of acceptance of the two modes. We asked students and administrators which survey mode they preferred. At the end of each questionnaire, we included a few questions for students to provide a general assessment of the questionnaire. Among other things, we asked for an overall judgment of the survey mode they liked best. The distributions of student preferences on the second day of testing (i. e., after they had experienced both modes) are presented in Table 4.

Overall, there were no differences in the average preferences on the second day in general or regarding the different modes the students had experienced. About half of the students preferred the reading-aloud condition, whereas the other half favored the self-administered mode. But which students preferred which mode? An in-depth look at the results reveals that it was migrant students and slow readers, in particular, who said they preferred the reading-aloud condition. For example, about 62 % of migrants who moved to Germany with their parents (2nd generation) indicated that they preferred having the questions read aloud to them, and these differences are statistically significant ($p < .05$). There are similar findings with respect to students’ reading speed: 67 % of slow readers preferred the reading-aloud mode, whereas fast readers favored the self-administered mode (62 %). These observed differences are statistically significant ($p < 0.05$).

At the end of the reading-aloud survey, we also asked students about the extent to which they agreed with the statement that they felt disturbed by administrators reading the questionnaire aloud. As can be seen in Table 5, about 73 % of all students disagree with the statement that reading aloud disturbed them. This general tendency can also be found among students with different migrant backgrounds and different reading speeds. However, there are further differences within these groups with regard to the perceived disturbance: The second generation still feels less disturbed than students without a migrant background, and slow readers feel less disturbed than fast ones ($p < .05$).

And how did the test administrators assess the two modes? The study involved a total of 19 test administrators, who carried out between two and 12 sessions. All test

Table 4 Student Preferences Regarding Survey Mode (Reading-Aloud vs. Self-Administration, Test Day 2, in %)

	Student preferences		Sign.	N
	Self-administered	Reading-aloud		
Total	50	50		637
Survey mode on the second day of testing				
Questionnaire read aloud	51	50		325
Questionnaire self-administered	50	50		312
Migrant status				
No migrant background	54	46		421
2.5 generation	60	40	*	63
2nd generation	39	62	*	83
1st generation	32	68		35
Reading speed				
Slow readers	33	67	**	143
Regular readers	52	48		354
Fast readers	62	38	*	140

Note. *** = $p < .001$; ** = $p < .01$; * $p < .05$.

administrators experienced both survey modes because the mode was switched for each test group on the second day. After each test session, the test administrators were asked about their personal preferences (reading-aloud or self-administered). The results show a preference for the self-administered survey mode in 62 % of all sessions. In addition, most of the test administrators had stable preferences for one of the two modes: about 10 out of 19 test administrators did not change their preferences at all during the testing period, and only one test administrator changed his mind more than once. If we look only at the last session of each test administrator (i. e., after they had fully experienced both survey modes), an even higher percentage of administrators (68 %) preferred the self-administered survey.

We also included some items on test administrators' subjective views regarding the advantages each of the two modes held for students. Here, we did not find strong differences between the modes. After the reading-aloud sessions, most test administrators (75 %) agreed that reading aloud was good for the students. After the self-administered sessions, most test administrators (76 %) agreed that self-administering

Table 5 Disturbance Perceived by Students as a Result of Administrators Reading Aloud the Questionnaire (in %)

	Strongly disagree	Somewhat disagree	Somewhat agree	Strongly agree	<i>N</i>
Total	57	16	10	17	637
Migrant status					
No migrant background	55	16	12	18	421
2.5 generation	57	15	8	20	63
2nd generation	70	18	6	6	83
1st generation	66	6	7	20	35
Reading speed					
Slow readers	64	15	9	12	143
Regular readers	57	16	11	16	354
Fast readers	51	16	10	23	140

the survey was good for the students. Furthermore, there is a strong correlation between administrators' personal preferences and their opinion as to what constitutes the "best practice" for students.

4.5 Duration of the Interview

Finally, we present findings on the time needed for the interview. Even though we do not consider the duration of an interview to be a quality indicator, it is an important piece of information when planning a study. It is important to keep in mind that administrators were trained to allow the class 15 minutes for filling out the regular questionnaire, then to stop, and to use the remaining 5 minutes for the final personal assessment items. In the reading-aloud condition, a couple of administrators were not able to finish the regular part of the questionnaire. As can be seen in Table 6, almost 30% of students in the reading-aloud condition on both days were not able to complete the entire questionnaire, whereas the corresponding proportion in the self-administered mode is lower than 10%.

The information on interview duration is only available at the class level and is based on the administrators' records. To receive more detailed information at the respondent level, students were asked to mark the item they were working on every 5 minutes. This allowed us to compare how many items students were able to complete after 5, 10, and 15 minutes. The findings for the first day of testing are present-

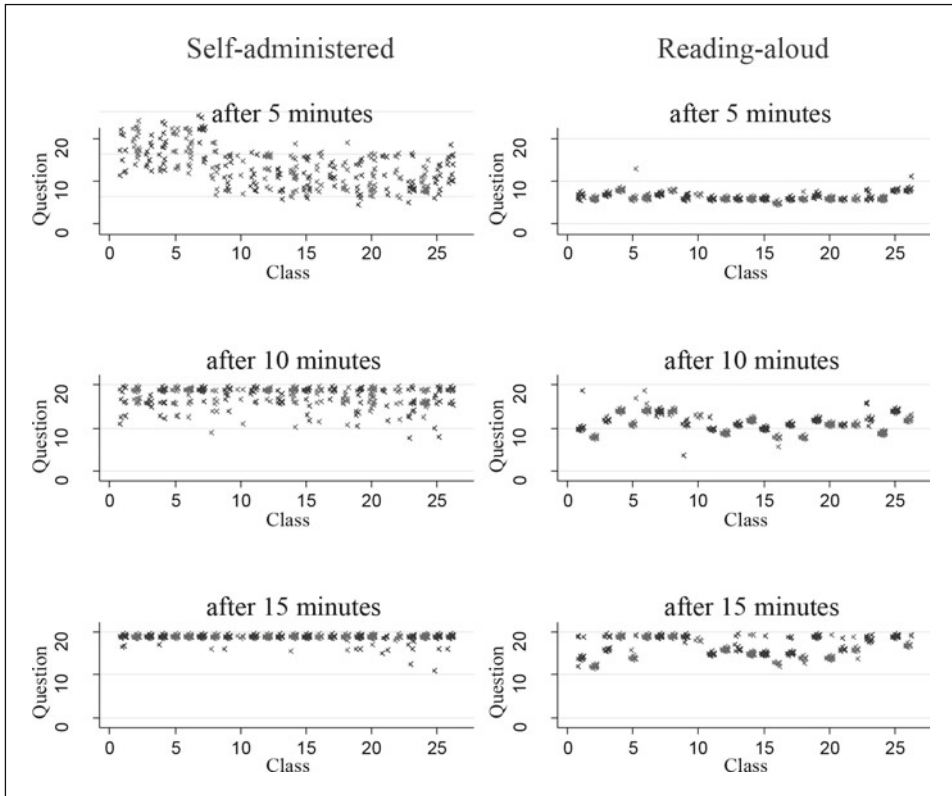
Table 6 Percentage of Students who Completed the Questionnaire in the Given Time (by Survey Mode)

	Test day 1		Test day 2	
	Self- admin.	Reading aloud	Self- admin.	Reading aloud
Completed questionnaires (in %)	95	72	90	71

ed in Figure 2. The y-axis represents the number of questions completed by students, and the x-axis represents the different classes. Each student is represented by an *x*.

The speed at which students went through the questionnaire is shown in the left graphs for the self-administered mode and in the right graphs for the reading-aloud mode. As can be seen, there is strong variance in the self-administered mode. After the first 5 minutes, some students were still answering the fifth question while others

Figure 2 Questions completed after 5, 10, and 15 minutes, by survey mode (Test Day 1)



in the same class had already finished the last question (No. 19). Five minutes later, most students had reached the last third of the questionnaire; after 15 minutes, only a few students were behind schedule.

In the reading-aloud condition—, all students fill out the questionnaire together without much variance within the classes, as was expected. As a result, all students needed more time compared with the other condition. After the first 5 minutes, most students had reached Question 7 or 8; after 10 minutes, some classes were still working on Question 8 while others had already reached Question 15. After 15 minutes, not even half of all students had reached the last question. These findings correspond with the high percentage of administrators who finished the interview in the reading aloud condition without finishing the entire questionnaire (Table 6).

5 Summary and Conclusion

In this paper, we presented initial findings on how reading a questionnaire aloud affects data quality compared with having respondents self-administer the survey. The analyses were based on an experiment in lower secondary schools in Germany, which are disproportionately attended by poor readers. In the reading-aloud condition, we could find higher data quality in terms of item nonresponse, multiple-answer questions, and filter questions. Further analysis on the role of migrant status and reading speed shows that the administration mode does not particularly benefit immigrants or slow readers: Among this lower secondary school population, all students benefit from reading questionnaires aloud. Slow readers and students with a migrant background perform worse in general, but they profit in similar ways from having the questionnaire read aloud to them.

Regarding students' subjective preferences for one mode or the other, there are differential findings: Migrant students and slow readers prefer the reading-aloud mode more than other students do. The majority of students do not feel disturbed by administrators reading out the questionnaire aloud; again, this is more the case for slow readers and immigrants. Interestingly, there are differences in students' preferences even though the modes' effect on performance is the same for all students.

Finally, the majority of test administrators preferred the self-administered mode, which is understandable since this mode means less effort for them. Nevertheless, they did not see any problems with the reading-aloud condition for the students and gave this mode a good rating as well. Despite these rather positive effects of reading the questionnaire aloud, reading the questions out takes more time. In case of time restrictions, this can lead to a higher dropout rate at the end of the questionnaire.

Finally, we want to point towards some limitations of our study and our analysis. First, we only focused on certain aspects of data quality: missing values and filter questions. However, the survey mode might also affect the quality of the answers given. For example, students might understand or interpret questions differently,

which is likely to result in different answers. Analyses on this aspect still remain to be carried out. Therefore, it is important to keep in mind that for this paper, we only investigated some of the factors that might generate differences in data quality depending on the survey mode (self-administered vs. reading aloud). These different sources of data-quality issues might end up in larger differences between the two modes than those presented in the paper. Second, the vast majority of our testing population consisted of poor or less-proficient readers. Therefore, our findings should not be generalized to apply to higher secondary tracks without further analysis.

Which conclusions can be drawn from these findings for future surveys? Our analyses indicate that filter questions and questions with multiple answers are complex questions that lead to high error rates in the response patterns and that these errors can be reduced by reading the questionnaire aloud. Against this background, the study indicates that reading questionnaires aloud may also be the preferred administration mode for studies of higher grade levels that pay attention to the difficulties of disadvantaged students.

However, given that reading questions aloud to respondents takes more time, survey designers should always weigh all the pros and cons when deciding on the survey mode. As shown above, the drawback of having a higher percentage of questionnaires completed in the self-administered mode is that data quality tends to be somewhat inferior. Thus, the question for interviews with a fixed time frame is: Should more questions be covered at the expense of data quality, or should better data quality and a higher motivation for survey participation among low-achieving youth be achieved at the expense of covering fewer questions?

References

- Blossfeld, H.-P., Roßbach, H.-G., & von Maurice, J. (Eds.). (2011). Education as a lifelong process: The German National Educational Panel Study (NEPS) [Special issue]. *Zeitschrift für Erziehungswissenschaft*, 14. Wiesbaden: VS Verlag für Sozialwissenschaften.
- Bradburn, N. M., Sudman, S., & Wansink, B. (2004). *Asking questions: The definitive guide to questionnaire design* (Vol. 2). San Francisco: Jossey Bass.
- De Leeuw, E. D. (2008). Choosing the method of data collection. In E. D. de Leeuw, J. J. Hox, & D. A. Dillman (Eds.), *International handbook of survey methodology* (pp. 113–135). New York: Psychology Press Taylor & Francis Group.
- Dillman, D. A., & Christian, L. M. (2005). Survey mode as a source of instability in responses across surveys. *Field Methods*, 17, 30–52.
- Groves, R. M., & Lyberg, L. (2010). Total Survey Error. *Public Opinion Quarterly*, 74(5), 849–879.
- Heydrich, J., Weinert, S., Nusser, L., Artelt, C., & Carstensen, C. H. (2013). Including students with special educational needs into large-scale assessments of competencies:

- Challenges and approaches within the German National Educational Panel Study (NEPS). *Journal for Educational Research Online*, 5(2), 217–240.
- Krosnick, J. A. (1999). Survey research. *Annual Review Psychology*, 50, 537–567.
- Meloy, L. L., Deville, C., & Frisbie, D. A. (2002). The effect of a read aloud accommodation on test scores of students with and without a learning disability in reading. *Remedial and Special Education*, 23(4), 248–255. doi: 10.1177/07419325020230040801
- Mullis, I. V. S., Martin, M. O., Kennedy, A. M., & Foy, P. (2007). *PIRLS 2006 international report: IEA's progress in international reading literacy study in primary school in 40 countries*. Chestnut Hill, MA: Boston College.
- Naumann, J., Artelt, C., Schneider, W., & Stanat, P. (2010). Lesekompetenz von PISA 2000 bis PISA 2009. In E. Klieme, C. Artelt, J. Hartig, N. Jude, O. Köller, M. Prenzel, ... P. Stanat (Eds.), *PISA 2009. Bilanz nach einem Jahrzehnt [PISA 2009. Review after one decade]* (pp. 23–71). Münster: Waxmann.
- NEPS. (2011). *Starting Cohort 3. Main Study 2010/11 (A28). Students, 5th Grade, Regular Schools: Information on the Competence Test*. Retrieved from https://www.neps-data.de/Portals/0/NEPS/Datenzentrum/Forschungsdaten/SC3/1-0-0/C_A28_en.pdf
- OECD. (2010). *PISA 2009 results: What students know and can do. Student performance in reading, mathematics and science* (Vol. 1). Paris: OECD Publishing.
- Randall, J., & Engelhard, G. (2010). Performance of students with and without disabilities under modified conditions: Using resource guides and read-aloud test modifications on a high-stakes reading test. *The Journal of Special Education*, 44(2), 79–93. doi: 10.1177/0022466908331045
- Schwarz, N., Strack, F., Hippler, H.-J., & Bishop, G. (1991). The impact of administration mode on response effects in survey measurement. *Applied Cognitive Psychology*, 5(3), 193–212.
- Wolf, M. K., Kim, J., Kao, J. C., & Rivera, N. M. (2009). *Examining the effectiveness and validity of glossary and read-aloud accommodations for English language learners in a math assessment* (CRESST Report No. 766). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Acknowledgement

We want to thank Franziska Matthes and Georg Camehl for their excellent assistance in data preparation and analysis, as well as Lena Nusser and Sabine Weinert for the opportunity to connect our project with their developmental study.

About the authors

C. Gresch

Institute for Educational Quality Improvement (IQB), Berlin.

e-mail: cornelia.gresch@iqb.hu-berlin.de

M. Kanders

Institute for School Development Research (IFS),

TU Dortmund University, Dortmund.

H. Solga

Social Science Research Center Berlin (WZB), Berlin.

R. Strietholt

Institute for School Development Research (IFS),

TU Dortmund University, Dortmund.

Quality Assurance in the Context of Data Collection

Franziska Fellenberg, Heiko Sibberns, Birgit Jesske and Doris Hess

Abstract

The NEPS maintains a number of specific requirements of effective quality assurance in the context of data-collection. These requirements concern the recruitment and training of test administrators and interviewers as well as their continuous supervision and the monitoring of fieldwork. Challenges to the NEPS are based on the multitude of different instruments, the heterogeneity of respondents, and the complex logistics and variability of data-collection settings. The NEPS has assigned two data-collection institutes with expertise in differing core areas of empirical social research. The IEA Data Processing and Research Center (IEA DPC) focuses mainly on tests with children and adolescents in the general education system. The Institute for Applied Social Sciences (infas) has its focal point in interviews and tests with individuals in private households. Both institutes have approved concepts of recruiting and training their data-collection staff as well as reliable strategies of monitoring fieldwork. These overall concepts and strategies are realized by measures that are adapted to the specific requirements of the NEPS substudies. The NEPS works in close exchange and cooperation with both institutes 1) in the phase of recruiting and training the test administrators and interviewers and 2) in the phase of supervision and field monitoring during the current data collection. During the first phase, the NEPS sets the standards and specified content for the training courses—and also partly for the recruitment—and takes an active part in the trainings to some extent. In the current field phase, the NEPS controls the data collection by means of standardized reports from the data-collection institutes and visitations of surveys in the field and the call center as well as via a random selection of recorded interviews.

1 Introduction

The quality of quantitative empirical data highly depends on a standardized, high-quality data-collection process with exceptionally skilled staff. This is especially relevant for the NEPS because the heterogeneity of respondents, data-collecting settings, and survey modes places the highest demands on the assigned data-collection institutes. NEPS data are based on interviews and tests with individuals of different ages and education, partly in a group context—often in educational institutions—and partly with single persons by phone or at home. The administration of telephone interviews, paper-and-pencil questionnaires, and competence tests in a paper-pencil or computer-based format requires survey staff with exceedingly specialized abilities and skills.

There are two fundamental components of quality assurance in the fieldwork: the recruiting and training of the test administrators and interviewers on the one hand and the supervision of test administrators and interviewers as well as the monitoring of the whole fieldwork on the other hand. The NEPS has assigned two external data-collection institutes that have excellent expertise in the field of empirical social research. The IEA Data Processing and Research Center (IEA DPC) is responsible for tests and interviews with children and adolescents in the general education system. The scope of duties from the Institute for Applied Social Sciences (infas) includes 1) interviews with mothers of infants up to the age of 3 as well as tests with their children; 2) interviews and tests with (generally) young adults in vocational schools, universities, and universities of applied sciences; and 3) interviews and tests with adult persons between 23 and 72 years old. The interviews with the target persons are conducted partly face-to-face and are partly self-administered.

2 Recruitment and Training of Test Administrators and Interviewers

2.1 Requirements for Test Administrators and Interviewers in the NEPS

Empirical social and educational research has to deal with several requirements in the context of the recruiting and training of the data-collection staff in order to safeguard high data quality. Elaborated recruiting and training strategies and an appropriate payment, in particular, are necessary to set up a competent team of test administrators and interviewers. Particularly important is avoiding test-administrator and interviewer effects on response behavior, test performance, respondent cooperation, and non-response (e. g., Lipps and Pollien, 2011; Lüdtke, Robitzsch, Trautwein, Kreuter & Ihme, 2007; Pickery, Lossveldt & Carton, 2001; Schräpler, 2004). The NEPS has to deal with specific demands in the context of data-collection, that is, 1) the multitude of survey instruments, 2) the different age cohorts, 3) the partly complex logis-

tics of the surveys, and 4) the different settings of data collection. Due to these special characteristics of the NEPS, the freelance working test administrators and interviewers are faced with very specific tasks that vary over the single inquiries with different samples and at different times. These single inquiries are called substudies below.

First, the test administrators and interviewers must be familiar with the survey instruments to ensure the error-free implementation of the NEPS guidelines within the scope of data-collection. Particularly important in this context is that the test administrators or interviewers of both institutes must often deal with interviews or questionnaires on the one hand and competence tests on the other hand. Therefore, the data-collection staff must be aware of the different demands that are required in the administration of an interview and in the administration of a competence test. In particular, the combination of both requires the test administrator or interviewer to change from being responsive to the target person to applying strict rules of standardization in the test administration.

Second, the test administrators or interviewers must be able to deal with the respondents of different age cohorts in an adequate way. As a result, interaction with children younger than 3 has different requirements than the interaction with Kindergarten children, adolescents, or adults. Interviewing and testing young adults reveal other possible problems than interviewing and testing older adults. The test-administrators and interviewers are confronted with different behaviors as a consequence of a target person's age cohort. They must therefore be carefully prepared for the interaction with the specific cohort and possible problems before they go into the field.

Third, the administration of CATI and CAPI as well as competence tests on paper or in a computer-based format includes the handling of (partly very) complex logistics. While interviews per telephone or face-to-face mainly require dealing with the software free of errors, the logistics for competence tests are often clearly more complex. For example, the competence tests for the Kindergarten children contain a lot of material because of the pictorial representation of the tasks and the response categories. The material must be carried undamaged and must first and foremost be presented in the correct order and duration as well as in an appropriate way so that the child is able to edit the tasks (e. g., the size of a chair and table has to be appropriate). Tests with infants younger than 3 need the correct installation of laptops for picture presentations, the setup of digital video cameras, and the accurate handling of test sequences with toys and other objects, including their disinfection.

The fourth requirement arises from diverse data-collection settings. With regard to the settings of the NEPS surveys, there is task sharing between the institutes. Surveys in an institutional setting are mainly administered by IEA DPC, and individual surveys are mainly administered by infas. However, there are also overlaps in this task sharing, for example, the surveys in universities and universities of applied sciences are conducted by infas. The demands on the data-collection in Kindergartens and schools include 1) the appropriate interaction with contact persons and invigilators and 2) the safeguarding of the standardized data-collection in the given context.

Similar demands are aimed at the test administrators of the surveys in universities and universities of applied sciences. The needs of individual surveys differ between telephone interviews and face-to-face interviews because of the different communication channels. It is crucial to motivate the target person and present the questions and answer categories in an understandable way in order to get valid information. Within both settings, an important challenge lies in possible disruptions and interferences in the domestic environment of the target person. As a consequence, the preparation and training of the test administrators and interviewers must be very specific to the demands of the different substudies.

In the following section, the recruitment and training concepts of the IEA DPC and infas are described in detail, as is the part of the NEPS in this phase of fieldwork. Concerning the terms used by data-collection institutes, there is a difference in the description of the persons who conduct tests and interviews. The IEA DPC refers to test administrators and infas refers to interviewers because the core areas of both data-collection institutes are different.

2.2 Recruitment and Training Procedures of the IEA Data Processing and Research Center (IEA DPC)

Overall recruitment and training concept (IEA DPC)

In general, test administrators are recruited from universities via job advertisements. These job advertisements contain a short job description, a profile of the desired candidate, and a list of the expected demands. Irrespective of the area of work, test administrators should have a background in education or psychology. They must be reliable and should be at least in their 3rd semester. They should also have communicative and social skills in order to cope with the complex test situation in which test specifications need to be followed exactly on one hand and interruptions need to be handled adequately and with care and tact on the other hand. Test administrators should be fluent in German and should be available throughout the testing window. Depending on the location of the testing, a car may also be required. Test administrators are not employed by the IEA DPC as regular staff, but act as freelancers on behalf of the IEA DPC.

Training the test administrators focuses on all aspects related to the testing situation. Prior to practical instruction tailored to the needs of the study, test administrators are informed about the general structure of the NEPS, its main stakeholders, and the expected outcome of the panel. After this broad overview, test administrators are familiarized with the particular substudy, its objective, its scope, and its size. The flow of information and the material as well as communication protocols are of special importance during the training as test administrators make their appointments with schools directly. Therefore, they must know exactly when they should contact the schools in order to arrange a testing session, how and where to communicate this

information in order to receive the necessary testing materials, how to control and confirm the receipt of material, and what to do in case of a disruption.

During the testing session, the correct identification of test takers, the correct allocation of testing material, the compliance with data-protection regulations, the correct completion of lists and protocols, as well as the handling of unforeseen situations are of paramount importance and need thorough preparation. Therefore, importance is attached to hands-on training and practice sessions, which include a number of incidents that need to be dealt with adequately. Subsequent to the introduction to the testing, test administrators learn how to assess outgoing material for completeness, what to do in case of missing instruments, and where and how to send the material. Test administrators are also informed that quality-control observers may make unannounced visits to a testing session and report results to the study center. After testing has been conducted in a school, the school coordinator is asked to complete a quality control fax with sections related to the test administrators' behavior. The fact that this may happen is also brought to the attention of test administrators. All the above information is written down in detail in test-administrator manuals. Finally, future test administrators receive all relevant contractual details.

Surveys with preschool children (IEA DPC)

The testing of preschool children has to meet requirements that differ from testing in school settings. Most importantly, the testing is organized as a face-to-face test rather than a group test. The test administrators make appointments with the Kindergarten and, contingent on parental permission, conduct the test with child. Since the test takes place in the Kindergarten during regular opening hours, care has to be taken in the provision of a quiet room and a setting that avoids disturbance from other groups of children. Testing in schools, on the other hand, follows the time schedule in place as closely as possible and only has to deal with disturbances during breaks.

After controlling formal requirements like parental permission and completing relevant forms, the starting point in preschool testing is always a familiarization of the child with the test situation. Depending on the child's reaction and his or her developmental stage, this can take some time and may also involve Kindergarten staff if necessary, for instance, if the child is very shy and asks for, or needs, a person he or she has confidence in.

Since children of this age can neither read nor write, icons, pictures, or symbols are presented to the children, and they must point to the correct answer. In some cases, a spoken response is desired. One part of the test resembles a game to be played on a board. Here, the children have to solve several tasks that are read and explained to them or are shown on cards, avoiding any positive or negative influence on the child as much as possible. Once the child has reacted to a stimulus, the test administrator has to record the reaction she has observed according to a pre-specified coding scheme. This is also different from the school situation, where test administrators lead a testing session but have a much less prominent role than in the preschool situation.

These conditions of testing preschool children also require special test-administrator recruitment procedures and training. One prerequisite of the NEPS is that all test administrators be female. All need a background as described in the section *Overall recruitment and training concept (IEA DPC)*.

For test-administrator training, a 'train the trainer' model has been adopted: Future trainers go through an intensive training program that deviates from the usual training procedures. After this training, they are in a position to competently select and train future preschool test administrators themselves.

For trainers as well as test administrators, general information regarding the NEPS follows the regular presentations. However, emphasis is put on the challenges related to the testing of preschool children: Different developmental stages, language barriers, concentration deficits, and shyness have to be addressed. Prior to the training, problems that have to be appraised by the future test administrators are illustrated through a video. The same video is shown after the training in order to get an indication of the learning effect.

Surveys with school children (IEA DPC)

Since there is no direct one-on-one interaction between the tested child and the test administrator in school surveys, the role of the test administrator is much more restricted to the formal adherence of test protocols. Tests are self-administered, and the children have to work through the test on their own. Depending on the content of the test and the age of the children, support from the test administrator may be permitted; the test administrator may read the test to the children or be allowed to answer a child's comprehension questions. However, most tests for school children are constructed in ways that allow no intervention by the test administrator. Questionnaires, on the other hand, may be completed with the support of test administrators. During the training, future test administrators are taught how to address possible questions that the children may have.

As already described, much more emphasis is put on the lines of communication and the more formal aspects of the test. All procedures are described in detail in a test-administrator manual. This manual is accompanied by a script that gives exact instructions as to what test administrators have to say and to do.

The tests are accompanied by listing and tracking forms that need to be completed prior to, during, and after the testing. Testing materials also need to be prepared. The correct completion of lists and the correct assembly of all material are one of the major tasks during the training.

Prior to testing, the permission of parents needs to be documented. Without written consent, no student is allowed to take tests or questionnaires. Therefore, the documentation of the correct completion of approval forms is emphasized during training. In the next step, the testing material has to be allocated to the students. Since only the school has a list with names and corresponding identification numbers and all printed test material only has the imprinted identification number, test administra-

tors have to follow a procedure with a temporary allocation of names to test booklets by use of sticky labels that can be removed after testing. This procedure has to be applied to all materials that need to be distributed to the children. Since permutations can occur when this procedure is applied, test administrators receive hands-on training where they have to allocate test material but also have to deal with permutations: These need to be recorded in the tracking form and test-administrator protocol of the tests. Test administrators are also given instructions regarding the correct placement of the children and the expected distribution procedure.

During the tests, the instructions need to be read, the participation status has to be recorded, time needs to be kept, and any interruptions or questions need to be documented. This is also simulated in the hands-on training session mentioned above. In particular, the treatment of unforeseen or disruptive events is trained in some detail, such as the treatment of late arrivals or early leavers, discipline problems, and interruptions including fire alarms.

After the tests, test administrators must ensure that all forms are completed and the testing material is fully collected and returned to the IEA DPC. The related procedures are also part of the test administrator training.

2.3 Recruitment and Training Procedures of the Institute for Applied Social Sciences (infas)

Overall recruitment and training concept (infas)

In recruiting interviewers, project-specific requirements (e.g., general qualification, training, foreign language skills) are taken into account as well as their availability for the time periods required and their proximity to sample communities. The conclusive choice of interviewers and their assignment to projects requires a comprehensive qualification program, which begins with a compulsory basic-training event for both face-to-face and telephone interviewers.

Thorough training and qualification activities for interviewers are an essential requirement for imparting the set of standard rules with regard to conducting interviews. These activities primarily aim at providing all interviewers with rules for standardized behavior in interview situations and committing them to comply with the rules.

The main focal points of the basic training event are the general features of both contacting and interviewing. *Contacting*, on one hand, includes the issues on how to prepare for establishing contact, how to communicate throughout the process of contacting, how to avoid refusals, and how to document contact results. *Interviewing*, on the other hand, refers to the standard rules for reading out the questions, probing, data entry, and handling specific questionnaire conventions.

The interviewers' successful participation in the basic training is checked with a short test. Telephone interviewers usually begin working in a coaching area, where

they are closely supervised during their first interviews, as soon as they have completed the basic training. For a first assignment to a project, participation in a project-specific training event is indispensable.

Study-specific training concept (infas)

While basic training deals with cross-project standards of interviewing and the collaboration between interviewer and survey agency in general, project-specific trainings refer to specific surveys and address the interviewers who are to conduct these particular interviews.

During project-specific trainings, interviewers get an overview of the study design (research objectives, sample, target group, etc.), a refreshment of contacting strategies (precontact letter, Refusal Avoidance Training, etc.), as well as information about the survey itself. Further central elements in the training concept are role plays and interview exercises, in particular.

As supplement to training events, all interviewers are handed out a study-specific manual, which works as a reference book and documents all necessary specific characteristics of a survey and requirements for the realization of interviews.

For example, the following description refers to a study-specific training concept for NEPS Starting Cohort 1. Interviews that are conducted within NEPS Starting Cohort 1 comprise both a personal interview and developmental psychological tests. The interviews are conducted with one of the baby's parents within the parental household. The tests are videotaped by the interviewer for later analysis.

Solely women are allowed to be interviewers in this project. The concept of the class-room-training divides the training event into two separate blocks, each two days long. The first block provides the interviewers with basic information about the project and information about hygiene standards. The interviewers are also introduced to the correct conducting of the tests (so-called games). About two weeks after the first training block, a second, obligatory follow-up training takes place. Due to the games' complexity and technical specifications, the games and possible difficulties in conducting them are trained in intensive exercises by groups of 2 interviewers.

Both training blocks are connected through the preparation of demo videos by each interviewer. On the basis of these demo videos, typical mistakes in performance can be regulated before the beginning of the field time. Furthermore, the final decision about assigning the interviewers to the fieldwork is made.

2.4 Monitoring of Recruiting and Training Processes by the NEPS

The degree to which the NEPS takes active part in the recruiting and training processes depends on the requirements of the particular substudy. The NEPS develops written instructions for the test administrators and interviewers—in an elaborated and a compacted form—in collaboration with the data-collection institutes for all substud-

ies. These instructions for the interviewers and test administrators are an important part of the briefing procedures. Furthermore, the NEPS visits training courses of both institutes for all cohorts at random.

In substudies with infants and Kindergarten children, the NEPS takes an active part in the selection and training of the test administrators. Before the potential test administrators are allowed to administrate tests in surveys with these target groups, they must run through an elaborate selection procedure (as described above). First, they must successfully complete a training course. The NEPS plays a significant role in the framework of such training, either in terms of a train-the-trainer course (for the IEA DPC) or by conducting important parts of the interviewer trainings (for infas). All these courses include the production of a test video by the participants that shows the accomplishments of the survey in as real a manner as possible. These videos are evaluated by members of the NEPS. Only those persons who show an error-free test administration in the video are accepted into the specific substudy.

The NEPS participates in the conception and realization of training units but not in the selection of the test administrators and interviewers in surveys that are implemented in schools, universities, universities of applied sciences, or in a domestic context. As a result, the NEPS compiles slides for the training courses and gives the staff of the data-collection institutes a briefing if necessary. As a rule, the slides of the NEPS contain basic information about the whole NEPS, the specific substudy including the characteristics of the sample, and the content and structure of the instruments. While the collaboration of the IEA DPC and the NEPS is limited to the preparation of the training courses, the collaboration with infas usually includes the active participation of NEPS staff members in the trainings. The latter includes theoretical instructions as well as practical group work.

3 Supervision and Monitoring of the Fieldwork

3.1 Requirements for the Supervision of Test Administrators and Interviewers and Monitoring of the Fieldwork

The supervision of test administrators and interviewers and the permanent monitoring of the fieldwork are essential for the quality assurance in the field. A major requirement is the monitoring of an enormous number of substudies that run partly parallel. A given institute is responsible for the supervision of its test administrators and interviewers in the fieldwork. In respect to the field monitoring as a whole, there is a close collaboration between both data-collection institutes and the NEPS. One aim of the field monitoring is to interfere within the current fieldwork if severe problems occur, for example, if the targeted sample size is not achieved. Another important aspect of the monitoring within pilot studies is to get relevant information to improve the survey processes in the main studies, for example, the choice of incentives.

The data-collection institutes have their own proven methods for the supervision of the test administrators and interviewers. The NEPS visits selected studies in order to inspect the fieldwork. Monitoring a multitude of test-administrators and interviewers in an extensive number of surveys that include the diverse requirements described above is a great challenge. If mistakes and deficits are noticed, it becomes necessary to take appropriate measures, which mostly involve additional trainings or instructions for the respective individuals or, in severe cases, the suspension of the test administrator or interviewer.

The monitoring of a survey's progress is another necessity for all parties involved. The field monitoring is based on three elements. The first element is objective data, such as response rate, the distribution of sample characteristics, and the duration of the surveys. The data-collection institutes monitor and document the current status of the field continuously and give the information to the NEPS. The second element refers to individual feedback from the test administrators and interviewers, and the third element applies the visits of the surveys by the NEPS staff. If the objective data or the individual observations show aberrations, the data-collection institutes and the NEPS reflect upon the problem together and take measures if necessary.

In the following section, the IEA DPC and infas present their supervision and monitoring concepts in detail. Finally, the foundation pillars of the monitoring in the current field phase from the NEPS are described.

3.2 Supervision and Field Monitoring by the IEA Data Processing and Research Center (IEA DPC)

Supervision of test administrators in Kindergartens and schools (IEA DPC)

Independent of the testing situation, test administrators have to report back to the IEA DPC on a regular basis. Once they have made contact with the schools, arranged for a testing date, received test materials from the printer, and when the testing is completed, they must inform the institute. There are also critical incidents when immediate contact with supervisors at the institute is mandatory: If the number of students willing to take the test is below a pre-defined rate, if the printed material is incomplete or faulty, or if the preparation of the tests in the schools was inadequate, for example, if students were not informed or parental permission was not collected. If test administrators do not report to the institute, they will be contacted by their supervisors.

Direct supervision is carried out on a random basis. Randomly selected testing sessions are observed by the NEPS staff or by the IEA DPC staff.

Finally, the quality of the returned material is judged by staff at the institute. The criteria are the correct order of materials, the completeness of forms, and the completeness of all testing materials. In case of deficiencies, payments to test administrators are reduced or completely retained.

Monitoring strategies in the fieldwork (IEA DPC)

Fieldwork is monitored in various ways. Test administrators report participation rates via email immediately after a test has been conducted. Unexpected occurrences are also reported, for example, too much or insufficient testing time or comprehension problems in parts of the questionnaires. Participation rates can be only preliminary but offer a first impression of the willingness to participate. This may lead to changes in recruitment procedures. Reports on other matters may lead to adjustments in test-administration procedures.

Secondary monitoring tools comprise the test administrator protocols and questionnaires. Test administrators should report on the cooperation with schools, testing conditions, deviations from procedures, and any occurrences that may lead to biased results.

Finally, school coordinators are asked to complete a quality questionnaire. Cooperation with the institute as well as with test administrators is of interest here. The appearance and behavior of the test administrator towards principals, teachers, and students is also judged.

3.3 Supervision and Field Monitoring by the Institute for Applied Social Sciences (infas)

Monitoring strategies in fieldwork (infas)

Continuous monitoring is essential to check the interviewers' compliance with the standards. Monitoring focuses on performance and quality with regard to contacting addresses of the gross sample as well as on the collected survey data. These two aspects are most important for interviewer monitoring. Regarding the contacting of addresses, it has to be guaranteed that the interviewers only work on the given addresses and not select other target persons on their own. Concerning the collected data, interviewers are monitored to check that they comply with standard rules and do not influence the target persons' answers.

At infas, monitoring covers a variety of information and data sources. These include statistics on response rates, cooperation rates, refusals, and contact attempts, which are collected as percentages, quotas, and averages and include results from interviewer monitoring via mail and telephone as well as evaluations derived from observed (watched) or recorded interviews.

Evaluations from monitoring processes in ongoing surveys are not only used to detect deviant interviewer behavior. Monitoring always implies a direct intervention in the process of data-collection. This can either mean that interviewers need to be detracted from projects or that they will receive a refresher training.

Study-specific supervision and monitoring (infas)

In the context of NEPS Starting Cohort 1 surveys, comprehensive study-specific monitoring standards have to be defined. The video recordings are always at the center of quality-management processes. For each game that a parent consents to and that is “played,” one video file is submitted by the interviewer. Before the videos are forwarded for coding and analysis, they go through a thorough monitoring process at infas.

For the purposes of quality control, a rating scale was implemented. By means of this scale, which consists of a certain set of criteria, the general quality of conducting the games is continually supervised. A so-called “Blitzrating” of eight variables helps to rate the technical video quality (e. g., field size, chosen image section, lighting conditions). In addition to this Blitzrating, a more detailed rating is carried out to monitor the correct game procedures and interviewer performances in a subsample of about 30 % of all videos. Apart from any inaccuracies in the experiment setup, mistakes in conducting the games as well as interferences from outside the game situation are documented and evaluated.

Evaluations based on these ratings are used for further supervision. In the event that difficulties occur in complying with the set of performance standards, interviewers receive immediate individual feedback with directions for a correct game procedure. In addition, further training sessions are also advised. With help of this two-part rating process, insecure behavior in interviewer performance as well as in conducting the games can be detected immediately. Performances are strengthened by means of a regular feedback for all interviewers in order to prevent the repetition of any inaccuracies.

In addition to this feedback, all interviewers in this project receive a monthly comprehensive review of their video-rating results.

3.4 Monitoring of the Fieldwork by the NEPS

The monitoring of the fieldwork by the NEPS is based on two pillars, namely frequent reports from the institutes on the one hand and visitations to the field on the other hand. These two pillars are completed by interview recordings that infas sends to the NEPS. The data-collection institutes deliver so-called reportings to the NEPS at regular intervals. These reportings contain the current status of a survey, such as response rates (see above). When a substudy is completed, the data-collection institutes send methodological reports to the NEPS that contain comprehensive information about the field progress of the whole substudy. These reports are especially relevant for the evaluation of pilot studies and for the preparation of a main study as well as being an important information source for data users.

The other pillar of the fieldwork monitoring by the NEPS is the visitation of the surveys at random. The visits to the field are unannounced as a rule. If studies in private households are visited, the permission of the target person must be obtained be-

forehand. The main criterion for the evaluation is that the test administrators and interviewers accomplish the survey as written down in the instructions. Furthermore, important aspects for the evaluation include the appearance of the test administrators and interviewers and their behavior towards the target persons and third persons, such as school coordinators. Another relevant criterion for the evaluation of the visits is the process of the study as a whole. Any detected deficits caused by the premises on which the study takes place, by inaccurate material, or by the planned field progress of the substudy are relevant because the NEPS can interfere immediately or modify the survey in the main study.

The evaluation of the telephone interviews is carried out in two ways. First, the NEPS visits infas' telephone studio regularly and examines the current interviews locally. Second, infas sends a cross-section of interview recordings from several surveys to the NEPS. The main criterion for the evaluation is the adherence to the instructions for interviewers and face-to-face interviews. Other relevant aspects for the appraisal of the interviews are verbal issues, such as speed and articulation, and contact with the target person, such as civility and friendliness.

These elaborate supervision and monitoring strategies of the data-collection institutes and the NEPS ensure the continuous and close control of field work. If improvements or interventions are necessary, the NEPS and the respective institute work together closely so that effective measures can be implemented.

References

- Lipps, O., & Pollien, A. (2011). Effects of interviewer experience on components of nonresponse in the European Social Survey. *Sociological Methods & Research*, 23, 156–172. doi: 10.1177/1525822X10387770
- Lüdtke, O., Robitzsch, A., Trautwein, U., Kreuter, F., & Ihme, J.-M. (2007). Are there tests administrator effects in large-scale educational assessments? Using cross-classified multilevel analysis to probe for effects on mathematics achievement and sample attrition. *Methodology*, 3(4), 149–159.
- Pickery, J., Lossveldt, G., & Carton, A. (2001). The effects of interviewer and respondent characteristics on response behavior in panel surveys: A multilevel approach. *Sociological Methods & Research*, 29, 509–523. doi: 10.1177/0049124101029004004
- Schräpler, J.-P. (2004). Respondent behavior in panel studies: A case study for income nonresponse by means of the German Socio-Economic Panel (SOEP). *Sociological Methods & Research*, 33, 118–156. doi: 10.1177/0049124103262689

Table 1 Measures for Quality Assurance in Field Work

	DPC	infas	NEPS
Requirements for test administrators and interviewers	<ol style="list-style-type: none"> 1) Background in education or psychology 2) At least in the 3rd semester 3) Communicative and social skills 4) Fluent in German 5) Availability and mobility 	<ol style="list-style-type: none"> 1) Project-specific requirements, such as general qualifications, foreign-language skills 2) Proximity to sample communities 3) Availability and mobility 4) Successful completion of qualification program 	<ol style="list-style-type: none"> 1) No overall requirements 2) Substudy-specific requirements, such as the sex of the test administrators and interviewers 2) Especially in substudies with young children: successful completion of a training course checked by video recordings
Embodiment of training courses	<ol style="list-style-type: none"> 1) Overall contents, such as the contact with Kindergartens and schools, identification of test takers, data-protection regulations, completion of lists and protocols 2) Survey-specific contents mentioned age specific aspects as well as the setting for <ol style="list-style-type: none"> a) preschool children and b) school children 	<ol style="list-style-type: none"> 1) Overall content, mainly contacting persons and conducting interviews 2) Survey-specific contents for several target groups, such as study design, refreshment of contact strategies, information about the substudy with specific role plays and exercises 3) Check of competencies by tests and special coaching in the first phase of telephone interviews 	<ol style="list-style-type: none"> 1) Participation in the preparation of training courses, mainly with respect to overall information about the NEPS and substudy-specific information concerning the questionnaires and tests 2) Taking an active part in trainings for substudies with infants and Kindergarten children and partly in trainings for substudies with adults; DPC: "train-the-trainer" workshops; infas: specific sections within the training courses
Information sources for supervision and monitoring	<ol style="list-style-type: none"> 1) Participation rates, unexpected occurrences 2) Direct examination of returned material 3) Test-administrator protocols and questionnaires and direct supervision at random 4) School-coordinator quality questionnaire 	<ol style="list-style-type: none"> 1) Response rates, refusals, and contact attempts 2) Ratings from observed or recorded interviews 3) Response patterns and distributions for selected variables 	<ol style="list-style-type: none"> 1) Reports from the institutes with statistics of the current status of the field work 2) Visitations in the field 3) Recorded interviews

About the authors

F. Fellenberg

Leibniz Institute for Educational Trajectories (LifBi),
Wilhelmsplatz 3, 96047 Bamberg, Germany.
e-mail: franziska.fellenberg@lifbi.de

D. Hess

Institute for Applied Social Sciences (infas),
Department of Social Research,
Friedrich-Wilhelm-Str. 18, 53113 Bonn, Germany.

B. Jesske

Institute for Applied Social Sciences (infas),
Department of Social Research,
Friedrich-Wilhelm-Str. 18, 53113 Bonn, Germany.

H. Sibberns

IEA Data Processing and Research Center (DPC),
Überseering 27, 22297 Hamburg, Germany.

VI. Data management, Coding, Dissemination, and User Support

Data Dissemination, Documentation, and User Support

Jan Skopek, Knut Wenzig, Daniel Bela, Tobias Koberg,
Manuel Munz and Daniel Fuß

Abstract

A major goal of the NEPS is to provide scientific use data to the international research community. For this purpose, the NEPS has set up a Research Data Center (RDC) that offers a comprehensive portfolio of services, allowing researchers to gain access to and work with NEPS data effectively. The RDC's support concept combines well-known approaches with innovative means of data documentation, data dissemination, and user support. Important building blocks of our dissemination strategy include the provision of (a) user-friendly and edited scientific use data, (b) flexible data access to the scientific community, (c) sufficient, easy-to-obtain, and clearly arranged documentation of NEPS surveys and data, and (d) extensive user support fostering good scientific practices and high-quality educational research. To achieve the highest standards in publishing panel data, the RDC has established a powerful infrastructure for data management, data dissemination, data documentation, and user support. In this chapter, we discuss the core elements of our concept.

1 Introduction

Implementing a multi-cohort sequence design, the National Educational Panel Study (NEPS) maintains six simultaneous panel studies (starting cohorts) for collecting comprehensive longitudinal data on educational careers and competence development representative of Germany. More than 60,000 target persons as well as additional context persons, such as parents, teachers, and school principals, take part in this large-scale social science endeavor. The NEPS approaches its starting cohorts with at least one wave per year. Moreover, the NEPS has conducted additional state-specific studies in Baden-Württemberg and Thuringia.

A major goal of the NEPS is to provide scientific use data to the international research community. For this purpose, the NEPS set up a Research Data Center (RDC) that offers a comprehensive portfolio of services, allowing researchers to gain access to and work with the NEPS data effectively. The RDC's support concept combines well-known approaches with innovative means of data documentation, data dissemination, and user support. Important building blocks of our dissemination strategy include the provision of (a) user-friendly and edited scientific use data, (b) flexible data access to the scientific community, (c) sufficient, easy-to-obtain, and clearly arranged documentation on the NEPS surveys and data, and (d) extensive user support fostering good scientific practices and high-quality educational research. To achieve the highest standards in publishing panel data, the RDC has established a powerful infrastructure for data management, data dissemination, data documentation, and user support. In this chapter, we discuss the core elements of our concept.

2 Preparation of Scientific Use Files

Consistent with the multi-cohort sequence design of the NEPS, the RDC releases six separate lines of Scientific Use Files (SUF). We expand these lines with new data in the course of annual panel waves. Two additional Scientific Use Files are published for the school-reform studies that are conducted in Thuringia and in Baden-Württemberg. In sum, the NEPS publishes 8 Scientific Use Files that provide comprehensive longitudinal data as a result of the first 5 years of the project. In the following section, we describe some important aspects that are relevant to the publication of Scientific Use Files.

Release Strategy

We aim to provide scientific use data to the scientific community no later than 18 months after the end of fieldwork. Thus, we publish data in the form of releases identified by a unique version number with three digits. As a result, researchers working with NEPS data are able to refer precisely to a specific version of the data. We even preserve and archive older versions of data, a measure that ensures that research can replicate results. Furthermore, the version number also discloses the number of included panel waves as well as major and minor data updates.

Citable Data Using DOIs

Traceability of the research process is a significant issue for ensuring good scientific practice. In this regard, the need for citable datasets has been increasing in recent years. Therefore, for each data release, we assign a unique digital object identifier (DOI) that is registered at *da|ra*¹ (Wenzig, 2012). Using the DOI enables researchers

1 *da|ra* is the registration agency for social and economic data run by GESIS and ZBW. See <http://www.da-ra.de/en/home/>.

to cite NEPS data in a very easy and precise way. The DOI also refers to a landing page at the NEPS web portal that provides basic metadata relating to the data and describing methods of data access.

Scientific Use Data

In order to provide high-quality panel data that are ready for scientific use, thorough preparation and editing of input data are essential. Consequently, before releasing any datasets to the research community, we conduct a set of preparation steps, including anonymization, data cleaning, editing, coding, variable generation, data enrichment, data linkage, and quality testing. Throughout the whole preparation process, the design of user-friendly data structures ensuring a maximum of data usability represents a central concern. Therefore, we work intensively on conceptual questions of how to design convenient data structures of Scientific Use Files. For instance, we offer integrated data structures in long format when reasonable and possible. Furthermore, while keeping the grade of detail high in the original panel data, we additionally generate more user-friendly data files that provide derived variables and data in a rather condensed and analytical form.

Large-Scale Editing of Longitudinal Data

The dissemination of huge collections of empirical data embedded in the complex panel design of the NEPS makes collaborative and systematic preparation of data indispensable (see Bela in this volume). The workflow of data editing mainly incorporates editors from the RDC, but experts in substantial topics across other units in the NEPS also participate in specific coding or edition tasks (e.g., the preparation and scoring of competence data by psychometricians, the coding of countries and languages of origin by experts on migration issues). To achieve high-quality standards in a multi-editor environment, we have built up a collaborative infrastructure and committed all coworkers to principles guiding the data-edition process. In addition to leaving raw data unchanged and organizing the edition process in intermediate steps, a fundamental principle is keeping the data edition traceable and replicable. We strive to achieve this goal through completely syntax-based procedures using the software Stata² as a standard technology. A distributed version-control program enables us to keep track of who changed what in syntax files—a technique commonly used in decentralized software development. As a result, the whole process of preparing Scientific Use Files remains traceable at any point in time because it is documented in high precision. For instance, the release of the first SUF for Starting Cohort 6 (NEPS SC6:1.0.0) represented a major collaborative effort. Up to nine data editors participated in the data preparation process. In sum, editors wrote more than 150,000 lines of syntax code and documented almost 900 code revisions.

2 <http://www.stata.com/>

Integrated Datasets

Due to annual or semiannual data collection sweeps along the different cohorts, the handling of panel-wave data (like data merging or harmonization of variables) would soon turn into a serious and exhausting task for researchers, even for a small number of waves. Thus, in order to minimize the research users' data-management efforts that are necessary to work with NEPS data, we produce integrated longitudinal datasets in long format where reasonable. This also includes the integration and harmonization of the data structure across different waves. For example, we usually harmonize variables over waves and cohorts. In particular, life-course data (episodes and spells) collected retro- and prospectively over waves are integrated and harmonized, which is a complex and time-consuming task. As a result, we significantly reduce the burden of data management on the researchers' side. Beyond that, we provide integrated tracking data for each cohort (via files called "*CohortProfile*") that allow for tracking the respondents' participation in the panel samples over waves.

Coding of Occupations, Branches, Courses, and Regions

The coding of open inputs in questionnaires is crucial for improving the quality and utility of data. A routine task in data editing is the recoding of textual responses to residual categories. However, substantial effort goes into the coding of occupations, which requires a significant amount of occupational knowledge and methodological know-how. The coding is mostly done manually, but automatic procedures (e. g., string comparison with keyword lists) that support the coding process have also been developed (see Munz, Wenzig, & Bela in this volume). We apply a set of quality measures in order to optimize the coding outcomes.

Occupations and vocational trainings are coded into the German classification of occupations ("Klassifikation der Berufe," KldB-2010, cf. Paulus, Schweitzer, & Wiemer, 2010). Furthermore, occupational codes according to the KldB-88 and the International Standard Classification of Occupations (ISCO; cf. ILO 2008) from 1988 (ISCO-88) as well as from 2008 (ISCO-08) are provided. Branches are coded according to the classification of branches by official statistics ("Klassifikation der Wirtschaftszweige," WZ08, cf. DESTATIS, 2008). The content of further education courses, such as those collected in Starting Cohort 6, are coded according to the catalog of competencies provided by the Federal Employment Agency ("Kompetenzkatalog der Bundesagentur"), and course identification numbers ("Kurskennziffern") of the catalog are assigned. Moreover, data on locations (like the location of employment or vocational trainings, current place of residence, place of birth) are coded to districts ("Kreise") using official district numbers ("Kreiskennziffer").

Scales and Generated Variables

To improve the usability of datasets, we generate useful additional variables. We deliver additional variables that measure socioeconomic status and occupational prestige, such as the Magnitude-Prestige scale (Wegener, 1985), the SIOPS-88 and

SIOPS-08/Treiman scale (Treiman, 1977), the International Socioeconomic Index of Occupational Status (ISEI-88 and ISEI-08, Ganzeboom, de Graaf, Treiman, & de Leeuw, 1992; Ganzeboom, 2010), the EGP classes (Erikson, Goldthorpe, & Proto-carero, 1979), and the occupational class scheme from Blossfeld (Blossfeld, 1985; Schimpl-Neimanns, 2003) by default. To classify educational attainment, we generated variables for CASMIN and ISCED-97. Relying on the threefold NUTS³ hierarchy of regional clusters, we provide variables containing administrative regions (“Regierungsbezirke,” NUTS Level 2) and federal states (“Bundesländer,” NUTS Level 1) from the coded districts (NUTS Level 3).⁴ Several other variables are generated that measure more specific issues, such as migration background, or that are needed for technical reasons.

Generated Files

We also provide generated data files that offer the user more simple data structures. These files consist of generated variables to a large extent. In particular, in enhancing the usability of the life-course data of Starting Cohort 6, we provide complex life-course data of an adult population in an easy-to-understand and condensed representation. For instance, we generated a data file titled “*Education*,” which provides simple-to-use data on educational transitions across an individual’s life course already coded in CASMIN and ISCED-97 (see Skopek & Munz in this volume). Another generated file contains all transitions of individuals in marital status reconstructed from comprehensive data on partnership biographies. Further biographical information are compiled in a user-friendly way in specific datasets on “*Further Education*,” “*Vocational Training*,” “*Employment*,” “*Military/Civilian Service*,” “*Children*,” and so on. We plan to extend the provision of generated files in the future in collaboration with researchers. In principle, users of the NEPS could also develop and provide generated files that might become published on the website or even part of the Scientific Use File in future data versions.

Weights

The method group of the NEPS prepares three basic types of weights that enable enhanced data analyses. For each first wave data release, (1) design weights, (2) nonresponse adjusted weights, and, if appropriate, (3) post-stratification weights are available. In particular, design weights are important since they account for the unequal selection probabilities in the sampling. A complete revision of implemented sampling strategies can be found in Aßmann et al. (2011). For some starting cohorts, additional replication weights complete the portfolio (Zinn, 2013).

3 NUTS stands for “nomenclature des unités territoriales statistiques,” that is, territorial units for statistics.

4 Currently, we are not allowed to provide NUTS Level 2 or Level 3 variables in conjunction with NEPS data collected in the context of schools or higher-education institutions. However, a federal state identifier (NUTS Level 3) will be available soon for these starting cohorts.

Imputations

Enriching the analytical potential of longitudinal data, the NEPS provides multiple imputations for selected variables containing missing data. Currently, researchers can rely on imputations for income in the context of Starting Cohort 6. The preparation of further files containing plausible values from multiple imputation models (MI files) is planned for the future.

Record Linkage With Administrative Data

Linkage of data—particularly of administrative data—is an important source for improving the quality of survey data. Record linkage with process data is a multifaceted venture in methodological, technical, and data-protection terms. However, conceptual and technical questions have already been clarified together with experts from the Institute for Employment Research (IAB) in Nuremberg. Data are linked on the basis of the respondents' explicit consent. The linkage is established at an individual level by means of probability matching relying on address data as well as on individuals' basic socio-demographic traits. The NEPS has already achieved a linkage of administrative data from the Federal Employment Agency (Bundesagentur für Arbeit) with the data of all published waves of Starting Cohort 6.

Regional data

The NEPS provides regional and macro-level data that can be merged easily with panel data. Fine-grained regional data up to street-section level are available for all cohorts (data are linked at the address level). Regarding these high-resolution regional data, we offer databases from two leading commercial providers of geodata in Germany: *infas Geodaten (now Nexiga)*⁵ and *microm*.⁶ Additionally, based on demand, we provide the service of matching users' own data on regional indicators (e.g., from the German Federal Statistical Office) with NEPS data. For data-protection reasons, analyses relying on linked regional data are restricted to on-site or remote data usage.

Metadata Enrichment

We utilize a database containing structured metadata (see below) to add rich metadata to the Scientific Use Files. Variable and value labels are added, edited, and checked for correct assignment. We translate these metadata into English, allowing international researchers to work with NEPS data comfortably. We additionally extend data files in Stata format by attaching the corresponding question in the survey instrument to the variables. Using a special Stata command, *infoquery* (Bela, 2013)—which is part of the Stata toolset *NEPStools* provided online by the RDC—users can immediately check how questions of certain variables were phrased (in German as well as translated to English) on the Stata's output console. Although this innovative feature

5 infas Geodaten/Nexiga, see <http://www.nexiga.com/>.

6 microm consumer marketing, see <http://www.microm-online.de/>.

is restricted to Stata, it has turned out to be very helpful for a broad range of users because it brings rich codebook information directly into the data. The same is true for *nepsmiss*, another Stata tool that automatically recodes all of the numeric missing values from the NEPS SUFs (-97, -98, etc.) into Stata's extended missing codes (.a, .b, etc.).

3 Data Dissemination

To grant access to the NEPS data, we have established a threefold infrastructure for data access. Research data is distributed by (a) secure download from the NEPS website, (b) an innovative remote-access technology (*RemoteNEPS*), and (c) on-site access. We designed these modes of data access not only to support the full range of users' interests and to maximize data utility, but also to comply with high standards of data confidentiality. While all three access modes provide scientific use data with a common data structure, they differ with regard to their degree of data anonymization.

Disclosure Control and Data Anonymization

To ensure the best possible confidentiality protection of individuals and individual microdata, the NEPS complies with strict national and international standards. First, the disseminated data have to be de facto anonymous data. This implies that we coarsen or cut off identifiable information to minimize the risk of statistical disclosure. String variables relating to openly asked questions are thoroughly checked. Second, the use of data is strictly confidential and for statistical purposes only. Therefore, access to NEPS data is granted exclusively to researchers (i. e., members of the scientific community) who sign a contract with the NEPS. The NEPS has made a large effort regarding legal regulations to keep the data with as much explanatory power as possible. If data modification is necessary, we only employ non-perturbative methods. Our concept of data dissemination distinguishes between three hierarchical levels of data sensitivity. While having the same dataset structure, data files available "on-site" provide more sensitive information than files available with "remote access," which, in turn, contain more information than the "download" versions of data files.

Data Contracts and User Management

Accessing data presupposes a signed contract that contains several data-use agreements. These stipulations require that the applicant handle the data in a secure and confidential manner. In particular, applicants commit themselves to strict data-protection guidelines that forbid any attempt at re-identification, passing along any data without permission, or using the data for purposes other than the specified research objective. The contract defines serious penalties if these stipulations are violated (e. g., high monetary penalty, proscription, exclusion from further data usage). Only re-

searchers who are members of a scientific institution (university or research institute) are eligible applicants. Researchers must provide a brief description of their project based on NEPS data and have to specify the expected duration of usage as well as further participants in the project. Since the NEPS grants data access for scientific purposes only, we check contract proposals strictly for scientific intentions. Approved data-use agreements are published on the NEPS's homepage together with the projects' description provided by the researchers. Contract documents are available in German as well as in English and are freely accessible on the NEPS website. Finally, after a contract has been approved by the NEPS, the researcher receives an NEPS login consisting of a username and password combination.

Secure Data Download

Researchers with a valid contract are able to download all available Scientific Use Files from the NEPS homepage via a secure SSL connection after login. A lot of additional documentation material necessary for using the datasets is provided.

RemoteNEPS: Access to Data in an Innovative and Secure Research Environment

The data access option *RemoteNEPS* represents a real remote infrastructure that was established by the RDC in addition to well-known on-site data usage and physical distribution of SUF (see Skopek, Koberg, & Blossfeld in this volume). *RemoteNEPS* provides safe and powerful access to sensitive NEPS data in an online research environment equipped with common statistical software packages and tools. Researchers can use *RemoteNEPS* with their NEPS login and an additional biometrical authentication.

On-Site: Acquiring the Greatest Detail of Data in a Physically Controlled Environment

The analysis of highly sensitive microdata is only possible on-site in a controlled physical environment at the Leibniz Institute for Educational Trajectories in Bamberg. On-site usage pertains mainly to the analysis of fine-grained regional data in combination with survey data as well as very sensitive items. The secure site prevents any copying or removing of sensitive data from the premises of the NEPS. All input and output devices are locked down, and the computers are not connected to the Internet or any other local area network. RDC staff is allowed to monitor any work performed on the data at all times. Any access to printers is controlled, and outputs process a review before they are provided (output control).

4 Documentation

Comprehensive and accessible documentation is crucial for good scientific research. Thus, the RDC has established a documentation structure relying on an integrated approach to metadata management. English documentation as well as powerful meta-

data services and tools are provided. All information is available on the research-data web portal maintained by the RDC.

Integrated Management of Metadata

A majority of the numerous NEPS substudies usually involve several instruments (i. e., survey questionnaires or competence tests) that typically define dozens of questions and items as well as filtering or interviewer instructions. Many of these items are repeatedly deployed for collecting panel data not only within one cohort, but also over different cohorts. As a consequence, an extraordinary abundance of metadata have to be administered and documented. Additionally, both metadata and the resulting documentation material have to be accessible in the English language since the NEPS aspires to deliver data to an international scientific community of educational researchers. In effect, metadata management is a crucial and nontrivial task at the NEPS.

The NEPS's metadata strategy strives for a structured approach to documentation (see Wenzig et al. in this volume). Only a structured documentation enables us to efficiently link, de-duplicate, reuse, and present all metadata. For this purpose, the RDC has developed a relational SQL database in collaboration with the German Institute for International Educational Research (DIPF) that enables the storing and linking of diverse metadata on studies, instruments, items, datasets, and variables in a systematic, powerful, and highly consistent fashion. As a crucial feature, metadata entities can be cross-referenced, for example, questionnaire items can be linked to datasets, allowing a dynamic documentation that directly leads from a dataset variable to the corresponding question in a questionnaire. In addition, the reuse of metadata enhances data quality because it allows for the tracking of inadvertent changes in variables across panel waves and starting cohorts. The structured documentation also enables an efficient translation of metadata as one has to translate reused elements only once.

As a result of a systematic approach to metadata management, researchers working with NEPS data enjoy a high documentation utility. Since we maintain metadata centrally in a database, corrections and extensions become effective in all derived documentation material, including multilingual codebooks, survey instruments, and dataset labels, in a synchronous and consistent manner. As described above, we even enrich the data files by using the meta-database.

Bilingual Metadata

To facilitate the international use of NEPS data, we consequently translate metadata of survey instruments (e.g., questions, answer schemes) and datasets (variable and value labels) into English. Since the translation of survey instruments is a complex and difficult task, we outsource this to professional translation agencies. We also rely on bilingual metadata for providing bilingual variable- and value labels in datasets.

NEPSplorer: An Efficient Tool for Searching the Meta-Database

In collaboration with the software-development unit at the NEPS and DIPF, the RDC has developed and published the online service *NEPSplorer*, which provides an efficient metadata service that enables the researcher to interactively explore, conveniently search, and quickly retrieve metadata. Like a search engine for the NEPS, it offers a full-text search for all documented metadata of survey instruments and Scientific Use Files. Users can search for and browse any items and variables of interest. For each item, information on question phrases, corresponding variables, answer categories, interview instructions, concepts, keywords, and many other things is available. The tool also displays cross-links between items in survey instruments and variables in Scientific Use Files. Users can store items of interest in a watch list and print an overview of these items. Furthermore, descriptions of surveys and starting cohorts are available. We have optimized the usability of the service by employing a modern asynchronous web frontend that possesses minimal response times.

Data Manuals, Codebooks, and Technical Reports

Apart from the SUF, the RDC prepares enhanced written documentation that is available for download. Most importantly, Scientific Use Files are equipped with *Data Manuals* (in English). The idea behind providing these manuals is to reduce usage hurdles by offering a succinct and user-friendly introduction to the data. For example, our data manuals describe the surveys, the file structure of datasets, content of data files, and the logic of file merging (for an example, see Skopek, 2013). Furthermore, the manuals include exemplary Stata and SPSS syntax that introduce typical data-management operations, such as merging files, handling spell data, and using weights while accounting for sample stratification.

In addition, we provide codebooks and a set of technical reports relating to the Scientific Use Files. The latter include methods reports that document the sampling and fieldwork process, weighting reports, anonymization reports, and data reports. Moreover, there are further supplements, including how-to guides (e.g., working with regional data) and interviewer manuals.

Finally, we also offer so-called *semantic data structure files*. These are data files of a Scientific Use File that have been emptied and thus contain variables and metadata (variable and value labels) but no data rows. These semantic files provided in Stata and SPSS format allow researchers to easily and intuitively explore data files without accessing real data and before signing a data-use agreement with the NEPS.

5 User Support

Currently (as of February 2015), more than 800 researchers have concluded a data contract with the NEPS. Up to 25 new valid contracts arrive per month. Hence, the NEPS is facing a rising demand for data. To facilitate proper usage of the NEPS data,

the RDC offers extensive user support. At its heart, the support program provides training courses including a comprehensive portfolio of theoretical, methodological, and technical topics relevant for working with NEPS data. We hold courses on a regular basis—about 8 to 10 per year—at our training facilities in Bamberg. Moreover, to proliferate the NEPS data, we occasionally provide courses off-site, which mostly take place abroad. User-training sessions usually come as two-day courses. While the first day provides a general overview of the NEPS study, the structure of datasets, the terms and conditions of data usage, and issues of privacy and data protection, the second day includes in-depth presentations, extended exercises, and hands-on data sessions.

To ensure continuous user support, the RDC additionally maintains an email hotline as well as a telephone hotline. The email hotline is supported by an electronic ticket system that facilitates an efficient internal workflow. The phone hotline is available at a separate phone number from Monday to Friday. Our hotline support provides a very high degree of individualized support. Nevertheless, we also provide on-demand, hands-on support in methodological and technical terms for data users (e.g., supporting syntax development, revising syntax, etc.).

6 Conclusion

A major mission of the NEPS is to provide high-quality scientific use data to the international research community. The NEPS has been successful in setting up a Research Data Center (RDC) that is capable of offering a comprehensive portfolio of services, allowing researchers to get access to and work with NEPS data effectively and with minimal constraints. The aim of this chapter was to provide an overview of the powerful infrastructure for data management, data dissemination, data documentation, and user support that has been implemented by the NEPS Research Data Center (RDC). Our strategy's cornerstones embrace the provision of (a) user-friendly and pre-edited scientific use data, (b) flexible data access for scholars, (c) clearly arranged documentation, and (d) comprehensive user support. As a result, a series of highly innovative approaches, instruments, and tools have been developed thanks to a young and highly motivated team that strives to achieve the highest standards in publishing panel data as well as to an adequate funding situation in the NEPS. These ingredients are important for promoting good scientific practices and high-quality educational research.

Finally, it should be noted that even if data management, data dissemination, and user support are crucial issues in social science's survey projects, they are also the issues that are most understated and underestimated in practice. Consequently, history has shown that many data-collection projects face difficulties or even fail to publish a consistent, usable, and well-documented database in a reasonable amount of time. Hence, researchers should be aware of this at the time of proposal writing.

References

- Aßmann, C., Steinhauer, H. W., Kiesl, H., Koch, S., Schönberger, B., Müller-Kuller, A., ... Blossfeld, H.-P. (2011). Sampling designs of the National Educational Panel Study: Challenges and solutions. In H.-P. Blossfeld, H.-G. Roßbach, & J. von Maurice (Eds.), *Zeitschrift für Erziehungswissenschaft, 14. Education as a lifelong process: The German National Educational Panel Study (NEPS)* (pp. 51–65). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Bela, D. (2013, July). *Provide, enrich and make accessible. Using Stata's capabilities for dissemination NEPS Scientific Use Data*. Presented at the meeting of German Stata Users Group, Potsdam.
- Blossfeld, H.-P. (1985). *Bildungsexpansion und Berufschancen: Empirische Analysen zur Lage der Berufsanfänger in der Bundesrepublik*. Frankfurt, New York: Campus.
- DESTATIS (2008). *Klassifikation der Wirtschaftszweige. Mit Erläuterungen*. Wiesbaden: Statistisches Bundesamt.
- Treiman, D. J. (1977). *Occupational prestige in comparative perspective*. New York: Academic Press.
- Erikson, R., Goldthorpe, J. H., & Portocarero, L. (1979). Intergenerational class mobility in three Western European societies: England, France and Sweden. *British Journal of Sociology, 30*(4), pp. 341–415.
- Ganzeboom, H. (2010). *Questions and answers about ISEI-08*. Retrieved from <http://home.fsw.vu.nl/hbg.ganzeboom/isco08/qa-isei-08.htm>
- Ganzeboom, H., de Graaf, P. M., Treiman, D. J., & de Leeuw, J. (1992). A standard international socio-economic index of occupational status. *Social Science Research, 21*(1), pp. 1–56.
- ILO (2008). *International Standard Classification of Occupations. Structure, group definitions and correspondence tables*. Geneva: International Labour Office.
- Skopek, J. (2013). *Starting Cohort 6: Adult Education and Lifelong Learning*. (NEPS Technical Report No. 3.0.1). Retrieved from <https://www.neps-data.de/de-de/datenzentrum/forschungsdaten/startkohorteerwachsene.aspx>
- Paulus, W., Schweitzer, R., & Wiemer, S. (2010). *Klassifikation der Berufe 2010. Entwicklung und Ergebnis* (Methodenbericht der Statistik der Bundesagentur für Arbeit). Nürnberg: Bundesagentur für Arbeit.
- Schimpl-Neimanns, B. (2003). *Umsetzung der Berufsklassifikation von Blossfeld auf die Mikrozensus 1973–1998* (ZUMA-Methodenbericht 2003/10). Mannheim: ZUMA.
- Wegener, B. (1985). Gibt es Sozialprestige? *Zeitschrift für Soziologie, 14*(3), pp. 209–235.
- Wenzig, K. (2012). *NEPS-Daten mit DOIs referenzieren*. (RatSWD Working Paper Series No. 202). Retrieved from <http://EconPapers.repec.org/RePEc:rsrw:rswwps:rswwps202>
- Zinn, S. (2013). *Replication weights for the cohort samples of students in grade 5 and 9 in the National Educational Panel Study*. (NEPS Working Paper No. 27). Bamberg: University of Bamberg, National Educational Panel Study.

About the authors

D. Bela
Leibniz Institute for Educational Trajectories (LifBi),
Bamberg, Germany.

D. Fuß
Leibniz Institute for Educational Trajectories (LifBi),
Bamberg, Germany.

T. Koberg
Leibniz Institute for Educational Trajectories (LifBi),
Bamberg, Germany.

M. Munz
University of Erlangen-Nuremberg,
Erlangen, Germany.

J. Skopek
European University Institute (EUI),
Florence, Italy.
e-mail: Jan.Skopek@EUI.eu

K. Wenzig
German Institute for Economic Research (DIW Berlin),
Berlin, Germany.

RemoteNEPS—An Innovative Research Environment

Jan Skopek, Tobias Koberg and Hans-Peter Blossfeld

Abstract

This chapter provides an introduction to conceptual, technical, and workflow issues of the National Educational Panel Study's remote-data-access solution. We illustrate that *RemoteNEPS* is capable not only of providing safe and highly controlled access to sensitive individual-level data, but moreover of offering an innovative, user-oriented, and very powerful research facility for analyzing rich and complex NEPS data. We present crucial conceptual aspects in the design of the NEPS and reveal how we put these aspects into practice. As we show, using *RemoteNEPS* is very simple. Contracted researchers need nothing more than web access and a (properly configured) standard web browser. It must be noted that running a remote-access site in this dimension is costly; however, the NEPS provides this service at no charge to its users. Importantly, our discussion on *RemoteNEPS* is not merely a conceptual blueprint; rather, it documents a system that has been in productive use for almost four years and that serves more than 200 users. *RemoteNEPS* is undoubtedly blazing the trail to the effective employment of remote access in the German context of research-data dissemination. Still, there are manifold ways in which *RemoteNEPS* could be improved in the future. We discuss the most promising aspects in our conclusion.

1 Introduction

Scientific data providers like research data centers often refrain from including sensitive microdata on individuals (or aggregated units like firms) in distributable Scientific Use Files due to data disclosure concerns. As a result, research-data centers often “bunker” important research data at the data custodian's location and provide access on-site only (if at all). This might harm the scope of scientific usage tremendously by

imposing a costly burden on researchers (especially from abroad). Sometimes, this might even inhibit use of valuable data, leading to an underutilization of research data, which was expensive to gather. The provision of remote-execution services and job-submission systems has partly resolved these deficits (Alda & Rohrbach-Schmidt, 2010; Frick, Goebel, Engelmann, & Rahmann, 2010). However, in most cases, *remote execution* represents a rather indirect and inconvenient method of data access involving the permanent manual intervention of service staff, who perform output control, thereby putting strain on users and research-data centers alike (Lane, Heus, & Mulcahy, 2008; Brand & Zwick, 2009). Not surprisingly, in recent years, *remote-access* services that allow direct but secure access to data have been discussed as being the most promising future approach for accessing sensitive microdata (Lane et al., 2008). In context of official microdata, several remote-access procedures are available in several European countries, such as Denmark, France, Sweden, Luxembourg, and the Netherlands (Reuter & Museux, 2010), of which the MONA (microdata online access) system of Statistics Sweden (Söderberg, 2005) seems to be the most developed (Brand & Zwick, 2009). Recently, the initiative “data without boundaries” began. This initiative is funded by the European Commission and tries to enhance scientific access to official microdata in Europe by connecting research-data centers that use remote-access systems (Silbermann, Bender & Hundepool, 2011). However, except for some first approaches,¹ data dissemination via remote access is still in its infancy in Germany with regard to official microdata and survey data.

By employing a real *remote-access* system for granting access to sensitive research data, the National Educational Panel Study (NEPS) takes on a pioneering role in Germany. Our remote-access solution, called *RemoteNEPS*, is a key component in the strategy of data dissemination by the NEPS (Barkow et al., 2011) and has been in productive use since August 2011. Merely being equipped with a current web browser, researchers can directly and visually access sensitive but still anonymized NEPS data in a fully equipped remote desktop environment. Importantly, *RemoteNEPS* is safe, powerful, and embedded in the overall legal framework of the NEPS.

The aim of this chapter is to describe key issues of *RemoteNEPS*. We begin with a conceptual overview on the design of *RemoteNEPS*, followed by a technical overview. In a second part, by referring to the functional features of remote access, we describe typical workflow elements using *RemoteNEPS* as an effective research environment for data-analytic projects. Finally, we provide a brief outlook by discussing future paths of development.

1 It is notable that the prototype “Morpheus” was recently implemented by the State Statistical Institute of Berlin-Brandenburg (Höhne & Höninger, 2012).

2 Conception and Technical Principles

2.1 Motivating a Remote-Access System

The National Educational Panel Study (NEPS) collects rich longitudinal data on educational trajectories and competence development over the whole life span. The core mission of the NEPS is to provide its data to the scientific community with a minimum of limiting boundaries and a maximum of data usability. However, these data partly contain quite sensitive information on respondents and institutions, such as characteristics of schools, regional information combined with occupational data, and fine-grained information on migration background. Additionally, NEPS data are subject to strict standards of data protection, and the NEPS has established measures of disclosure control at high levels. Hence, with regard to disclosure control, this gives rise to a restricted and regulated data access, at least for some parts of the NEPS data. Apparently, tension arises between data usability on the one hand and compliance with high data-protection standards on the other hand. While researchers should have the highest level of detail and differentiation possible for their empirical analyses, some information has to be restricted due to data-protection concerns. The NEPS resolves this conflict to by providing a threefold access to Scientific Use Files (SUF) of different sensitivity: (a) physical distribution of data files via web download, (b) remote access, and (c) on-site data access at the local data-security site of the NEPS.

While downloadable SUF are applicable for a plethora of scientific investigations, compared with other methods of access, their informational content is the most restricted in terms of data anonymization (see the chapter by Koberg in this volume). However, datasets accessible on-site provide data in the highest resolution, but a researcher has the burden of traveling to the NEPS at Bamberg. In order to improve accessibility to a wide array of data, the NEPS introduced its own remote-access solution that bridges the gap between downloads and on-site access. Our remote access solution, *RemoteNEPS*, balances privacy and access to data based on a portfolio approach to data protection (Lane & Schur, 2009).

When designing *RemoteNEPS*, we had two basic goals in mind: First, we wanted *RemoteNEPS* to provide contemporary data access that is safe enough for disseminating even sensitive microdata. Beyond that, by exploiting common features of remote access technologies, we also wanted *RemoteNEPS* to provide researchers with a powerful toolbox furnished with recent statistical packages for analyzing NEPS data. Hence, *RemoteNEPS* is not only secure, but it is also a usable research site. We elaborate on selected highlights of this concept in the following section.

2.2 Safe Remote Access to Sensitive Data

RemoteNEPS is safe in the sense that it represents a highly controlled method of accessing microdata. A crucial idea behind remote access in general is that data do not physically leave the data custodian's site. In most applications, users visually work in a (remote) window-based desktop—either from their own computer or in a controlled setting in a trusted center—but do not download any data files to their local PC. *RemoteNEPS* adheres to this idea by providing a web-based application that connects to a remote desktop server running at the NEPS. Researchers can effectively access data only remotely, not locally. However, by starting a remote desktop session on their PC, they can work with data as if it were on their PC. At the same time, the data do not physically leave the secure site of the NEPS.

Technically, using *RemoteNEPS* solely requires web access and a current web browser to establish a remote connection. Similar to the MONA system in Sweden, *RemoteNEPS* implements a terminal server solution based on Microsoft remote-desktop technology (cf. Reuter & Museux, 2010). A secure socket layer encryption (via HTTPS) makes this client-server connection safe. We employ a Java client that runs within all modern web browsers, independent of the operating system. Hence, in addition to a browser with a Java plugin, activated cookies, and an Internet connection, the user does not need any special hard- or software environment; users can access *RemoteNEPS* independent of whether or not they have a Mac OS-, Windows-, Android-, or Linux-driven computer. Importantly, the configuration of *RemoteNEPS* assures that no data- and file exchange can take place between the user's local desktop and the remote desktop. This suppresses any copy-and paste-features between local and remote desktops.

Compared with the remote execution or job-submission systems mentioned above, a major advantage of remote access is that it requires significantly less output control. Users get immediate screen outputs when working with the data and do not have to wait for approval and the provision of job outputs. Annoying waiting times for receiving a command output are thereby omitted, which increases the user experience. Hence, unnecessary overhead for the service staff as well as for the users is minimized. The quality of scientific results might eventually be enhanced as researchers—facing significantly fewer restrictions—are encouraged to explore the data more precisely, that is, to undertake alternative analyses (e.g., to check the robustness of results).

2.3 Biometrical Authentication

There is a common consent in the interpretations of data-protection laws that only registered persons (i.e., by user contract) may receive microdata for a specific purpose. Guaranteeing this is practically impossible when distributing data by deliver-

ing physical files (e.g., data delivery on DVD or by downloadable files). However, remote access replaces the physical distribution of files by providing a remote connection to data. Still, a remote-access system has to make sure that only authorized persons, namely contracted researchers, access sensitive microdata remotely. In this regard, a simple login based on a username and password combination would not be satisfactory since people could share account data even more conveniently than data files. For this exact reason, remote-access systems sometimes rely on additional hardware tokens or biometrically regulated access, such as fingerprint systems (Reuter & Museux, 2010). While these procedures assure that only privileged persons may access the data, they usually involve significant additional expenses for purchasing and distributing appropriate hardware and software devices (e.g., fingerprint scanners).

RemoteNEPS also relies upon biometrical authentication. However, when designing *RemoteNEPS*, we opted for a method that is safe on the one hand and that opposes a minimum of hurdles to the research users on the other hand. Specifically, we decided to employ a lean biometrical procedure that recognizes users by the way they type. Keystroke biometrics, which is based on the premise that individual key-striking behavior is as unique as handwriting, has become quite a popular approach for securing critical enterprise applications (Banerjee & Woodard, 2012). We utilized this technology in developing *RemoteNEPS*.² Before connecting to a remote-data desktop, users have to authenticate themselves by (1) a username, (2) a password, and (3) their keystroke behavior.

Like other biometric access procedures, keystroke biometrics require the system to learn a person's idiosyncratic way of typing in a first step. Hence, in a so-called enrolment step, a user has to type a predefined sentence around 8 times under the supervision of authorized NEPS staff. During this phase, the biometric software traces and stores the user's very individual typing profile. Later on, the software compares the user's typing behavior upon login with the stored profile. The biometric application allows some variation in typing; thus, the keystroke biometrics work on all standard keyboards and even on most non-standard keyboards (except for virtual-touch keyboards, which can be found, e.g., on tablets). Users can usually enroll in the program at NEPS user-training events. A typical enrolment process does not take longer than 5 minutes. Importantly, as typing behavior might change over time, the biometric systems account for this with every new login. Data of the typing profiles are stored at the NEPS separate from other personal data on our data users.

2 We purchased a solution from the company KeyTrac/TM3 Software GmbH in Regensburg (Germany).

2.4 User-Rights Management System

In addition to a safe and controlled entrance to the remote desktop, our concept also involves controlled access to remote resources. In other words, we have to manage *who* may access which kind of *data* and *service* for how long. To do this, we have implemented a user-rights management system that guards and documents exactly what access privileges our research users currently have and had in the past. This pertains to access to all of the NEPS services in general as well as to the remote access service in particular.

2.5 Powerful Research Environment

Up to now, we have outlined *RemoteNEPS* as a safe and rather comfortable method of accessing sensitive data when compared with other solutions. However, we argue that our remote-access system offers more than that. A second goal of *RemoteNEPS* is to provide easy and convenient data access. Users should not have to deal with cryptic installation procedures, but rather experience an out-of-the-box accessibility without much of their own configuration. Within the remote desktop, we equip users with an array of useful tools for analyzing NEPS data. These tools include recent and widespread statistical packages for data analysis. Currently, we provide and maintain recent versions of Stata, SPSS, and R.³ Moreover, office suites like MS Office and Libre Office, powerful text editors like Notepad++, and PDF tools are available to the user. This enables users to prepare publishable tables, figures, and documents within the data environment. We also included open-source software for version control, which gives users the possibility to develop syntax in a systematic, traceable, and collaborative way. Version control is especially useful when users want to share remote projects with other users and work together in a collaborative workspace. Furthermore, research users can take full advantage of high-performance servers' capacity for running computationally intensive jobs. Taken together, *RemoteNEPS* provides a fully equipped and powerful research environment directly out of the box without any further software installation needed on the user's part.

2.6 Inputs and Outputs

RemoteNEPS is equipped such that researchers can arrange the whole data analysis within the remote environment. However, they eventually need to export their results

3 To optimize user experience and utility, we provide an array of up-to-date user-written Stata commands and R packages in addition to the core installation. If something is missing, it is installed and configured by the NEPS staff upon request.

(like tables or figures) to prepare a publication. Alternatively, researchers might have developed preliminary syntax using the downloadable data files and wish to re-run their analysis using variables that are only available in the remote version of the data. That is, they might want to put results into *RemoteNEPS*. Getting work into and out of *RemoteNEPS* is quite easy via a personal web interface available to the *RemoteNEPS* user after login on the NEPS pages. The NEPS Data Center service staff archives and checks inputs and outputs to ensure the integrity of data. Once approved, outputs are delivered to the researcher in a personalized download section on the webpage in a timely manner, and inputs are delivered within the remote-data session.

2.7 Legal Framework

Researchers can use *RemoteNEPS* based on three preconditions. First, only contracted researchers may use NEPS data. In a standard agreement of data usage, a researcher declares the final, scientific purpose and duration of his or her data use. Additionally, when using remote access, the researcher must fill out a contract supplement containing special stipulations relating to the usage of *RemoteNEPS*. With this supplement, users commit themselves to utilizing the remote access exclusively for their own purposes in a closed environment and not in public as well as to refraining from using equipment for image recording (photo and video cameras) and producing screenshots when working with *RemoteNEPS*. Furthermore, the training informs the researcher about the fact that the NEPS stores outputs and inputs and reserves the right to carry out privacy checks.

Secondly, a researcher has to participate in one user training provided by the NEPS Data Center. NEPS user trainings give special instructions regarding how to use *RemoteNEPS* properly, how to respect data privacy, how to avoid misuse, and what the legal, contractual, and professional consequences of data misuse are.

Third, before being able to use remote access, researchers have to be enrolled in the biometric authentication and identification service connected to *RemoteNEPS*. During the supervised enrolment step, the system records a user's typing profile automatically. Recording these data essentially requires a signed consent by the users, which is provided to the users in advance. This is usually done when a researcher is visiting the user training.

2.8 Technical Specification

Maintaining a usable remote-access application for a multitude of data users sets high demands on the remote server's hardware equipment. As a benchmark, we aligned the *RemoteNEPS* system to be capable of serving 50 users simultaneously. To do this, our server system running the remote terminal server possesses 72 CPU cores with

144 GHz of power. Additionally, the system currently has 1,344 GB of RAM. At best, each user may enjoy the advantages of 8 CPU cores with 2 GHz each and 64 GB of RAM. However, this hardware back-end is used to support several virtualized services; hence, how many resources are assigned to a user's data session depends on the overall system load.

2.9 Licensing

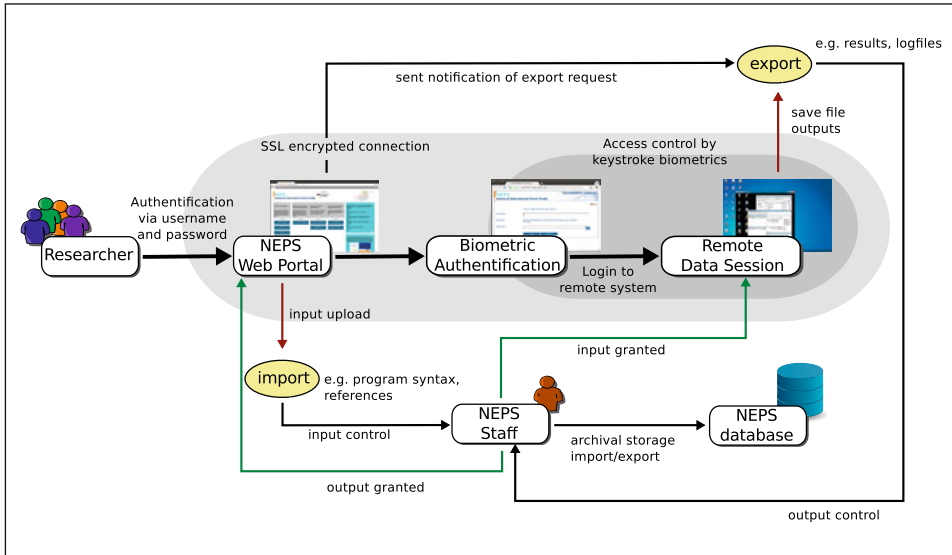
Since we provide software in the remote environment, licensing issues have to be considered. In general, our model of fifty users must simultaneously account for covering concurrent users with adequate software licenses. This is not only true for visible applications like Stata, SPSS, and Microsoft Office, but also for the underlying services like Microsoft terminal server, MS SQL database server, and the operating system. While we receive a slight academic discount from the software companies and distributors, licensing a remote-access system like *RemoteNEPS* demands a serious monetary investment and costs on a regular basis. Of course, keeping commercial packages like Stata and SPSS up to date requires the calculation of licenses for new versions in the overall budget. Nevertheless, the NEPS offers access via *RemoteNEPS* without charge.

3 Workflow

In the following section, we illustrate a workflow of data analysis using *RemoteNEPS*. In doing so, we presuppose a user who has a valid contract with the NEPS, has participated in the NEPS user training, and was enrolled in keystroke biometrics. Figure 1 supports our discussion by illustrating different steps in working with *RemoteNEPS* from a user-centric perspective. Figure 2 provides an overview of the technical workflow showing which system components are involved in the process.

3.1 Login to the Remote System

To begin remote access, an enrolled user simply has to open a web browser and request the URL address of *RemoteNEPS* (<https://remote.neps-data.de>). Alternatively, the user can find a hyperlink on the NEPS webpage. Afterwards, the user arrives at the login screen (Steps 2 and 3 in Figure 2). The following authentication process involves three verifications: First, the system asks for the username each NEPS user receives after signing a valid contract of use. The system looks up the username in an active directory application that centrally stores NEPS user data (Step 4). Second, the system prompts the user to provide a keystroke sample by typing a preconfigured sen-

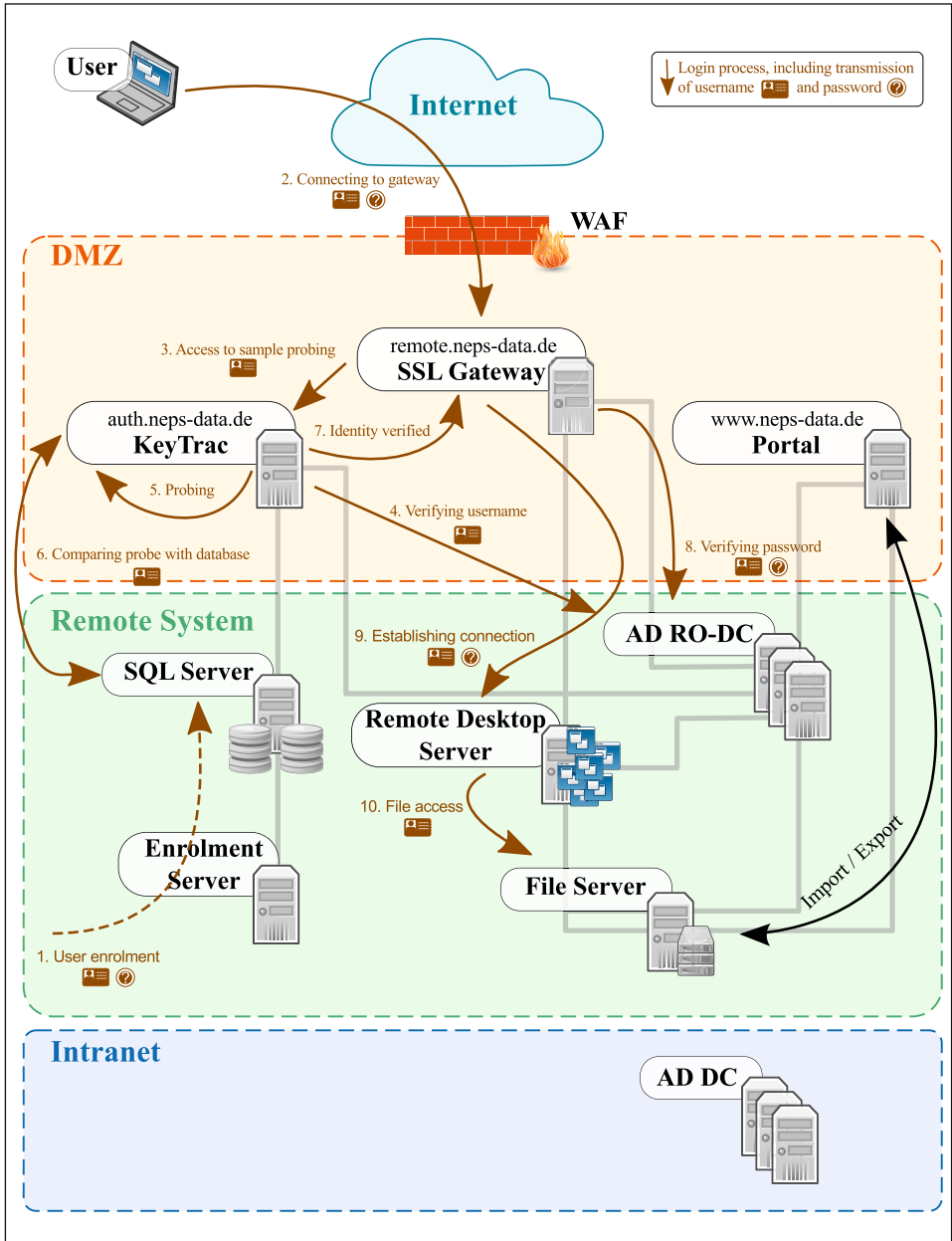
Figure 1 User Workflow of *RemoteNEPS*

tence in the enrolment into a text input (Step 5).⁴ It is important to note that the system recognizes users by the way they type rather than by the exact letters they type, as is true for password protection. Technically, recording is achieved by client-based JavaScript. Specific characteristics of the keystroke sample, such as intervals between keystrokes and the duration of key pushing, are compared with the user's keystroke profile previously stored at biometric enrolment in an SQL database (Step 6). If the grade of comparison is beyond a defined threshold, that is, if comparison produces a satisfying outcome, the biometric authentication was successful and the user is personally identified (Step 7). Moreover, the system adds the current sample to the user's profile to adjust for changes in typing behavior. If the keystroke sample deviates from the profile too strongly, the system displays an error prompting the user to re-type the sentence.

After successfully recognizing the user by his or her typing, the system asks for a password in the last step (Step 8 in Figure 2). Passwords for NEPS services (like *RemoteNEPS* and secure-file download on the webpage) are unique (i. e., one password serves all services), are also stored in the centralized active directory, and, most importantly, have to be strong. Thus, prior to the first use of any NEPS service, a user has to change the initial password he or she received after signing a valid contract. New passwords have to comply with rules that make passwords strong; if a password is too

4 We employ the phrase "National Educational Panel Study: Education as a Lifelong Process" as a standard sentence.

Figure 2 Technical Workflow of RemoteNEPS



- 1) *User's home directory*. Only the user has access rights to this. The home directory is useful for organizing syntax files.
- 2) *Project tree*. If more than one person is working on a research project, a project folder is available for storing joint documents or command files.
- 3) *Exchange folder*. This represents the interface for imports and exports (see below). This folder is partially personalized, that is, each user account has its own exchange folder. Access is only possible for the owning user and staff of the NEPS Data Center.
- 4) *Data inventory*. All published NEPS Scientific Use Files are available in a data drive. This includes not only recent versions of scientific use data, but also all previous versions. Furthermore, not only are the remote-access data versions of data files available, but so, too, are the downloadable versions. Beyond the data, the inventory provides extra tools (e.g., Stata ado files) and documentation. All valid users are able to read the content of the data inventory.

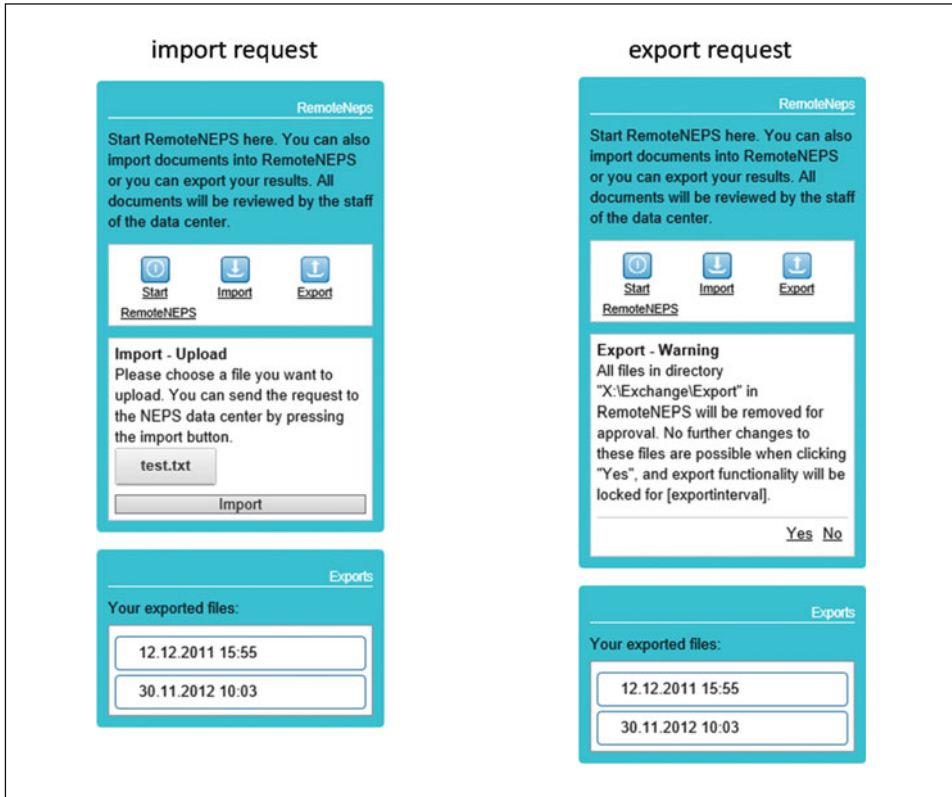
To enable a convenient working environment, several software applications are available that can be called by clicking on application icons on the desktop. Users can find icons for opening up Stata, SPSS, R (with GUI), office applications like Excel, a text editor, and the version-control software BAZAAR. Invoking (system) functions or programs beyond the scope of the provided software is not feasible. Thus, users cannot (accidentally) misconfigure their working environment and render it poorly usable or even inaccessible.

3.3 Import and Export

Interaction between the user's desktop and the remote system is limited to mouse and keyboard input and screen outputs. For imports and exports, we provide a web interface with file upload that enables users to submit import requests (see left panel in Figure 4). The system sorts import requests into a waiting queue, which is successively processed by experienced service staff of the NEPS Data Center ("input control," see Figure 1). Once checked and approved to meet formal standards (e.g., after undergoing a virus scan and checking for data-security specifications), the input is granted and provided in the user's exchange folder on the remote desktop. The system simultaneously notifies the user via email.

The treatment of export requests works vice versa. Users put results that they want to export into the export folder on the remote desktop. By clicking on the export button in the remote-access web interface (see right panel in Figure 4), the user generates an output request that is immediately sent to the export queue ("output control," see Figure 1). Again, staff members control and approve the output is compliant with data-security regulations, particularly that it is confidential (e.g., no export of individual-level data). After success, we (automatically) notify the user by email, who

Figure 4 Web Interface for Import and Export Requests



then finds the requested output in a personalized download section on the web portal (see “Your Exported Files” in Figure 4). To avoid permanent output requests, we configured a minimum time interval of 7 days between two requests.

4 Conclusion

This chapter provided an introduction to conceptual, technical, and workflow issues of the National Educational Panel Study’s remote-data access solution. We illustrated that *RemoteNEPS* is capable not only of providing safe and highly controlled access to sensitive individual-level data, but moreover of offering an innovative, user-oriented, and very powerful research facility for analyzing rich and complex NEPS data. We presented conceptual cornerstones in the design of the NEPS as well as how we put these designs into practice. As we have shown, using *RemoteNEPS* is very simple. Contracted researchers need nothing more than web access and a (properly config-

ured) standard web browser. Of course, running a remote-access site in this dimension is costly; however, the NEPS provides this service at no charge to its users. It is important to emphasize that our discussion of *RemoteNEPS* is not merely a conceptual blueprint; rather, it documents a system that has been in productive use for almost four years and that already serves more than 200 users, which is approximately 40 % of the total number of researchers who currently have a valid data contract with the NEPS (as of February, 2015). *RemoteNEPS* is undoubtedly blazing the trail to the effective employment of remote access in the German context of research-data dissemination.

Although the current version of *RemoteNEPS* builds on state-of-the-art technology, there are many possibilities for further enhancements. We now wish to point out the possible improvements that we have reflected on most.

First, our users are currently mostly on their own during their sessions. We provide full support for requests by phone and email, but when there is a need to assist in syntax development or to interpret software errors, no direct access to the user's session is currently possible. This sometimes leads to difficulties in solving blocking problems since the service staff cannot directly see what the user is seeing. Therefore, features for controlled session sharing would provide the service staff with an efficient means to solve these issues.

Second, although remote access radically reduces administration overhead input- and output control compared with alternative remote execution, all inputs and outputs still have to be reviewed manually. As user numbers and the system load increase, the number of outputs and inputs to be reviewed naturally increases, as well. We need supporting software to handle this effort at some upper limit. Although similar software has been developed and used in other research-data centers, a solution for the NEPS would require a heavily customized configuration.

Our users may cooperate in a scientific project. *RemoteNEPS* accounts for this by providing a shared project folder in the remote session. Additionally, we provide standard-version control software that researchers may use for collaborative-syntax development. For the future, even more collaborative features with which to furnish the remote desktop could be developed.

Another future improvement pertains to the Java-based client. Since Java is (still) quite common, a Java web-browser plugin implies only marginal demands of the user's environment. This fact notwithstanding, for the future, we will consider alternative technologies to establish the remote access connection. Indeed, the newly introduced HTML5 has features that reset the demand of the user's client to an up-to-date browser with no need for an installed Java plugin. We are currently testing and evaluating such a solution.

Finally, a major extension would be to provide not only remote desktops, but also remote virtual servers. We currently run a kind of "one-size-fits-all" solution by allocating desktops to different users. However, these desktops run on the same operating-system environment, thereby significantly restricting the possibilities for individ-

ual adjustments of the whole environment. Nevertheless, we could provide our own (virtual) remote server for each user by means of server virtualization, thereby allowing for a very high level of customization and flexibility. For example, users could choose between different operating systems (e. g., Windows or a Linux distribution). Moreover, users could configure the operating system directly, thereby creating an environment perfectly suited to their needs (e. g., they could install special-purpose software, software-development kits, or compilers) without affecting other users. The MONA system at Statistic Sweden, clearly an international benchmark for remote access, has already opted for this strategy.

References

- Alda, H., & Rohrbach-Schmidt, D. (2010). New data and services for vocational education and training research—Research Data Centre of the Federal Institute of Vocational Education and Training (BIBB-FDZ). *Schmollers Jahrbuch*, 130(2), 253–267.
- Banerjee, S. P., & Woodard, D. L. (2012). Biometric authentication and identification using Keystroke Dynamics: A survey. *Journal of Pattern Recognition Research*, 7(1), 116–139.
- Barkow, I., Leopold, T., Raab, M., Schiller, D., Wenzig, K., Blossfeld, & H.-P., Rittberger, M. (2011). RemoteNEPS: Data dissemination in a collaborative workspace. In H.-P. Blossfeld, H.-G. Roßbach, & J. von Maurice (Eds.), *Zeitschrift für Erziehungswissenschaft*, 14. *Education as a lifelong process—The German National Educational Panel Study (NEPS)* (pp. 315–325). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Brandt, M., & Zwick, M. (2009). infinitE—Eine informationelle Infrastruktur für das E-Science Age. Verbesserung des Mikrodatenzugangs durch “Remote-Access”. *Wirtschaft und Statistik*, 7, 670–675. Wiesbaden: Statistisches Bundesamt.
- Frick, J. R., Goebel, J., Engelmann, M., & Rahmann U. (2010). The Research Data Center (RDC) of the German Socio-Economic Panel (SOEP). *Schmollers Jahrbuch*, 130(3), 393–401.
- Höhne, J., & Höniger, J. (2012). *Morpheus—Remote access to micro data with a quality measure*. (RatSWD Working Paper Series No. 203). Berlin: German Council for Social and Economic Data.
- Lane, J., Heus, P., & Mulcahy, T. (2008). Data access in a cyber world: Making use of cyberinfrastructure. *Transactions on Data Privacy*, 1(1), 2–16.
- Lane, J., & Schur, C. (2009). *Balancing access to data and privacy: A review of the issues and approaches for the future*. (RatSWD Working Paper Series No. 113). Berlin: German Council for Social and Economic Data.
- Reuter, W. H., & Museux, J.-M. (2010). Establishing an infrastructure for remote access to microdata at Eurostat. In J. Domingo-Ferrer, & E. Magkos, (Eds.), *Privacy in statistical databases, lecture notes in computer science* (Vol. 6344, pp. 249–257). Heidelberg: Springer.

Silberman, R., Bender, S., & Hundepool, A. (2011). The need for networks on data access—data without boundaries project and the workshop on data access. *Joint UNECE/Eurostat work session on statistical data confidentiality*. (Working Paper No. 35). Taragona, Spain.

Söderberg, L.-J. (2005). MONA—Microdata ON-line access at statistics Sweden. *Joint UNECE/Eurostat work session on statistical data confidentiality*. Geneva, Switzerland.

About the authors

H.-P. Blossfeld
European University Institute (EUI), Florence, Italy.

T. Koberg
Leibniz Institute for Educational Trajectories (LifBi),
Bamberg, Germany.
e-mail: tobias.koberg@lifbi.de

J. Skopek
European University Institute (EUI), Florence, Italy.
e-mail: Jan.Skopek@EUI.eu

Management of Metadata: An Integrated Approach to Structured Documentation

Knut Wenzig, Christian Matyas, Daniel Bela, Ingo Barkow
and Marc Rittberger

Abstract

This chapter discusses the core elements of the metadata strategy striving for a structured approach to documentation, which is necessary to efficiently link, de-duplicate, re-use, and present all information in two languages for the six panels and two school-reform studies of the National Educational Panel Study (NEPS). For this purpose, the NEPS Data Center has implemented a relational database that enables storing and linking diverse metadata on studies, instruments, items, datasets, and variables in a systematic, powerful, and highly consistent fashion. As a crucial feature, metadata entities can be cross-referenced; for example, questionnaire items can be linked to datasets, allowing for a dynamic documentation that leads directly from a dataset variable to the corresponding question in a questionnaire. In addition, the re-use of metadata enhances data quality because it allows for the tracking of inadvertent changes in variables across panel waves and starting cohorts. The structured documentation also enables an efficient translation of metadata because re-used elements only need to be translated once. As a result of the systematic approach to metadata management, the documentation utility for researchers working with NEPS data is optimized. Since metadata are centrally maintained in a database, corrections and extensions become effective in all derived documentation material, such as multilingual codebooks, survey instruments, and dataset labels, in a synchronous and consistent way. In addition, the NEPSplorer offers a powerful metadata online service that provides the researcher with an interactive exploration of, efficient search for, and fast retrieval of metadata. Finally, the established metadata system offers a solid basis for further highly innovative developments in the field of metadata management.

1 Metadata Services and Underlying Concepts

The National Educational Panel Study's (NEPS) metadata structure currently allows bilingual documentation of survey instruments (e. g., questionnaires) and Scientific Use Files (SUFs). They form an integral part of data edition (cf. Bela 2016 in this volume) and innovative retrieval facilities. This chapter provides an insight into the design and current usage of NEPS metadata. The main goal was to build a repository of all metadata needed to conduct such a complex project as the NEPS. We use this repository as a single source of information for every metadata product, such as codebooks, the website, and even metadata that are shipped with Stata and SPSS data files. This keeps redundancy at a minimum level and makes the re-use of information possible in the first place. In order to achieve this goal, the NEPS uses an abstract data layer comparable with the standard of the Data Documentation Initiative (DDI)¹ but implemented as an Entity-Relationship-Model hosted by a productive relational database system. In our case, we use the Microsoft SQL Server, which is able to respond quickly enough even to a very complex search query that is needed to support our metadata management system and also the public retrieval interfaces. Figure 1 gives a broad overview of the architecture.

1.1 Services

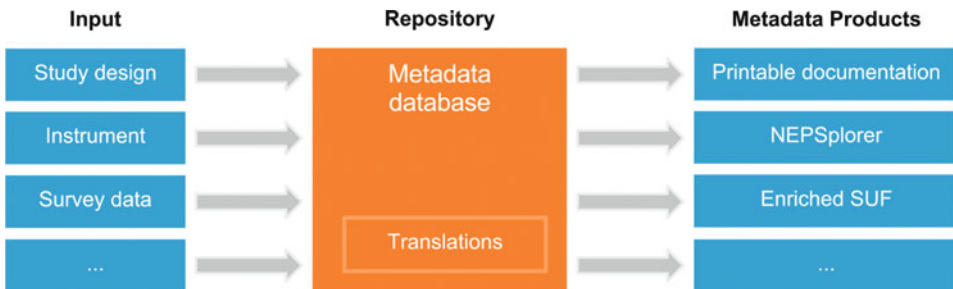
The first milestone of the NEPS metadata project was to gather the survey instruments, including questionnaires and corresponding codesheets (PAPI) and programming masters (CATI), in a structured way, as described in Section 2.1. A translation process (cf. Section 2.3) was established upon this base that allows for the re-use of work done by external translation agencies. These survey instruments are now bilingually available as generated versions for data users.

Furthermore, the Scientific Use Files (SUFs) are nearly² completely mapped in the metadata. (cf. Section 2.2) This information is used during the edition process and therefore plays an important role in this area: The variable names can be adjusted for an international scientific community, the sequence of the variables in the datasets is controlled by the metadata, and variables that are not included in the metadata will not be delivered. The English versions of the datasets are also produced by relying on the translated metadata.

There is a modeled relation between the variables in the survey instruments and the SUFs. With this information, the variables in Stata datasets can be enriched with the wording of the corresponding question (cf. Section 3.1).

1 <http://www.ddialliance.org/>

2 The missing values are not included in the metadata, which describe an SUF.

Figure 1 Concept of the metadata management

On this basis, bilingual codebooks accompanying the SUFs are produced that show the questions and frequencies for all variables. The online service *NEPSplorer* delivers a user interface to the NEPS metadata, which opens up the whole data offering with a few mouse clicks (cf. Section 3.3).

These innovations would not be possible without the interdisciplinary collaboration of specialists from the fields of information technology and social sciences. These specialists have built up a pronounced vertically integrated infrastructure that is now indispensable for delivering NEPS data to the scientific community within tight schedules.

1.2 Vertical Integration: Re-Use as a Paradigm

Through the structured collection of metadata of instruments (cf. Section 2.1) and datasets (cf. Section 2.2), content in many places and structures can itself be re-used. On the one hand, resources can be saved if repeated questions don't have to be re-entered and thus do not need to be translated again. On the other hand, consistency is ensured because before the design of new questions, a powerful database is available to research already-used verbalization.

Re-use takes place in the following points in particular:

- Label sets of variables, known as schemes, are recorded only once and re-used where applicable. This also applies to their translations.
- Where possible, already-recorded questions are re-used in other questionnaires.
- Translated variables and value labels of programming masters and coding sheets are re-used in the datasets as applicable.
- From the viewpoint of the database, the structure of survey instruments (e.g., questionnaires) and SUFs (data files) is almost identical. The translation process for the labels in the datasets could be applied immediately.

- Because the structure and metadata is stored in related tables with little redundancy (normalized), these data can be represented quite differently and re-used as generated CATI programming masters or PAPI questionnaires, as codebooks, within the NEPSplorer, and within datasets. A correction in a variable label found in a programming master must thus be made once and will appear in datasets and NEPSplorer. It is also impossible for a text to change in the various modes of representation because manual intervention is not necessary.

Structuring and re-use also challenge all stakeholders: Explicit modeling is required, and in most cases, exceptions to the structured model cannot be made.

2 Core Areas of the Metadata Structure

There are three areas that can be described as the core of the NEPS metadata. The questionnaires, the Scientific Use Files (SUF), and the metadata gathered during the translation process.

2.1 Representing the Questionnaires

Within NEPS, surveys are almost exclusively³ performed with paper questionnaires (PAPI) or are computer-assisted (CATI/CAPI). These survey instruments that are created by item developers are the basis for data collection done by external survey institutes.

In addition to the wording of questions, variable names and labels complete the specification of the dataset, which is delivered by the survey institutes after field work. In this respect, paper questionnaires consist of two separate documents: the printer's copy of the questionnaire with the original lay-out on one hand and the coding sheet with variable names, variable labels, value labels, and the assignment of numerical values to responses on the other hand. In the case of CATI interviews, both sources of information are integrated into the programming master, and the surveys are programmed and later on performed based on this information. The original documents for this purpose are created with popular office software in a multi-stage editing process in which nearly the whole NEPS consortium is involved.⁴

3 The conducted online surveys are based on programming masters that are very similar to those of CATI interviews.

4 At the beginning, the development process of the survey instruments should have had support from the metadata services. The impressive concept of collecting meta-information at the source has been dropped for now due to the complexity of this case.

Basically, the two survey modes do not differ in their requirements for data storage. The following outline exists from the outermost layer moving inwards, and this outline is also reflected in Figure 3:

- a survey instrument with a title and an introductory and concluding text,
- a hierarchical structure (e. g., chapter or modules) with a title and an introductory and concluding text for each element,
- questions, and
- question number and filter information—as properties of the question, when used in this instrument.

In this case, a question is represented in the database in three forms:

- a question text (in general, possibly several pairs of condition and question text),
- an interviewer instruction for various survey modes, and
- one or more variables with a variable name, a variable label, a subsidiary question, a response specification (see below), and if necessary (usually at CATI interviews), missing values (e. g., not specified, do not know) and schema extensions (which are more pairs of values and value labels).

A free text or a numerical answer can serve as a response specification as well as a response scheme. The latter consists of an ordered list of pairs of a numeric value and a text, the value label. The database is thereby structured so that response schemes may (and should) be re-used within various variables. Additionally, questions can be re-used in different instruments.

Figure 2 shows a sample question that is used in a CATI programming master and in a PAPI questionnaire. The following stands out:

- A question can be re-used in various survey instruments and modes.
- After metadata gathering, it is possible to select display formats that show different survey modes.
- Elements such as question numbers and filters are features that the question receives if it is used in a particular instrument. They may be different for the same question from instrument to instrument.
- The interviewer instruction is stored mode-specific.
- In this representation, the variable name and label are suppressed, and the numerical values corresponding to the answers are not shown. Another representation displays this information to assist during data analysis.

The example in Figure 2 shows that for the use of a question in an instrument, additional properties (question number and filter) can be stored. Moreover, the variable

Figure 2 The same question used in a CATI programming master (top) and a PAPI questionnaire (bottom)

22203	<pre> --va: etzeit --fn: 22203 --vb: Temporary employment --fr: Did you work as a temporary or contract worker then? --in: <<Also at a personnel recruitment agency.>> --we: 1: Yes 2: No BUTTONS: Refused (-97), Don't know (-98) --af: IF 22203 = 1 GOTO 23300 --end-- </pre>
12 Did you work as a temporary or contract worker then?	
Yes	<input type="checkbox"/>
No	<input type="checkbox"/>
Refused	<input type="checkbox"/>
Don't know	<input type="checkbox"/>
"Yes": Please proceed with question 15	

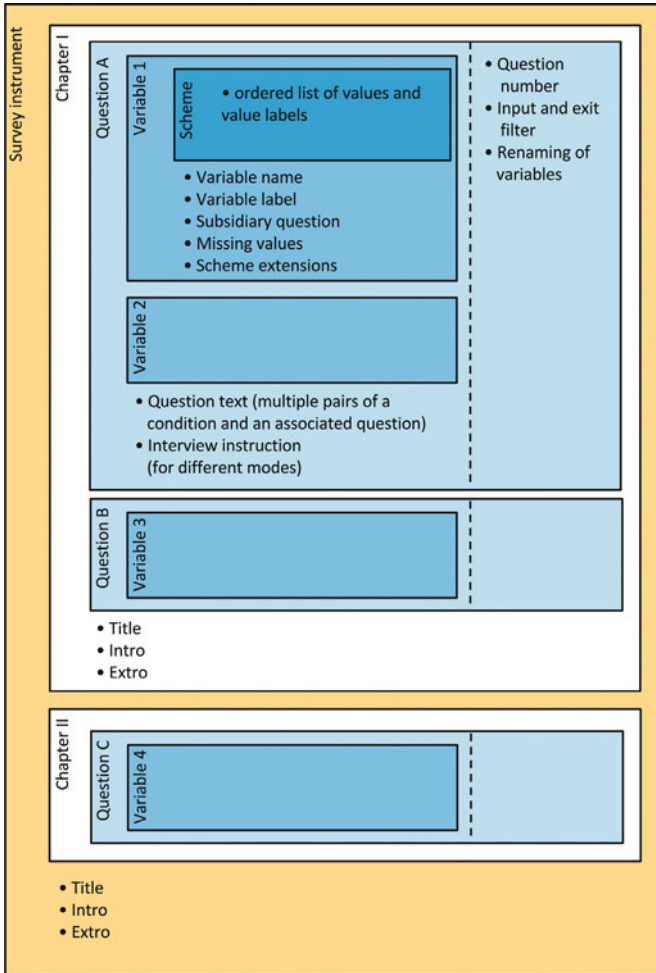
name can be overridden for use in a specific questionnaire: When two questions differ only in the variable names, they need to be entered once.

If the questions were entered and used in a questionnaire, a report with this questionnaire can be generated. Thus, the most basic requirement is satisfied: The input of the split and structured information can be output in a form that is very close to the raw material.

The output is produced by reports that are provided with the help of SQL Server Reporting Services from Microsoft. With this functionality, both the CATI and the PAPI instruments are modeled reasonably well for documentation purposes. While the mere reproduction of the original templates facilitates the monitoring of inputs, in principle, more generic issues are conceivable: The representation in NEPSplorer makes no difference in survey mode.

In programming masters, conditions are used at many points: in the filtering process of the question (see *input and exit filter* in Figure 3), in the wording of the questions themselves (see *multiple pairs of a condition and an associated question* in Figure 3), and also in the formulation of possible answers. As for the conditional questions,

Figure 3 Overview of properties that can be stored within a survey instrument



the metadata schema are extended to improve the re-usability of questions. The conditions themselves, however, represent a fundamental problem: Their syntactic structure is often not consistent enough to be directly interpretable by software, although this would also be advantageous in the development and testing of the instruments.

Currently, 43 survey instruments are included in the meta-database. They consist of a total of 5,081 questions, with 11,228 variables. 3,891 questions are distinct, and 840 different response schemes are used.⁵

⁵ The counts from the SQL database in this chapter were determined by its administrator Manfred Dussold.

2.2 Representing the Scientific Use Files

The documentation of the datasets should be based on the metadata not only because the files themselves should be offered bilingually, but also in order to use the translation process in this area.⁶

It quickly became clear that the modeled structure of a survey instrument and an SUF consisting of multiple datasets is almost identical and that only the level of the item is redundant, which means that in the SUF case, every item is filled with only one variable.

The overview in Figure 4 shows highly simplified objects contained in the data structure and how they are used for representing instruments and SUFs. A detailed entity-relationship diagram is shown in Figure 11 in the Appendix.

The right side of Figure 5 shows the information structure for an SUF and how it strongly resembles that of a survey instrument (left). The main difference is the fact that questions in a dataset do not exist, and thus, the object in the database that stores the questions is used but always contains only one variable. The previously mentioned association between the SUF and one or more survey instruments can and should be represented, as well. This relationship is depicted in the NEPS metadata by explicitly modeling relationships between the variables in the SUF and the survey instrument.

In Figure 5, this relationship is represented by the dotted lines between the left and the right side. If such a relationship exists, the question of the survey instrument is mapped onto a variable in the dataset. This link is used for example in the codebooks.

With this reference, it is not necessary in many cases to define a new variable label: If no variable label is defined in the SUF, the variable label is inherited from the corresponding variable in the survey instrument. This also holds true for the response schema. In this manner, the redundancy in the metadata is limited, and re-use of translations is increased.

2.3 Organization of the Translation Process

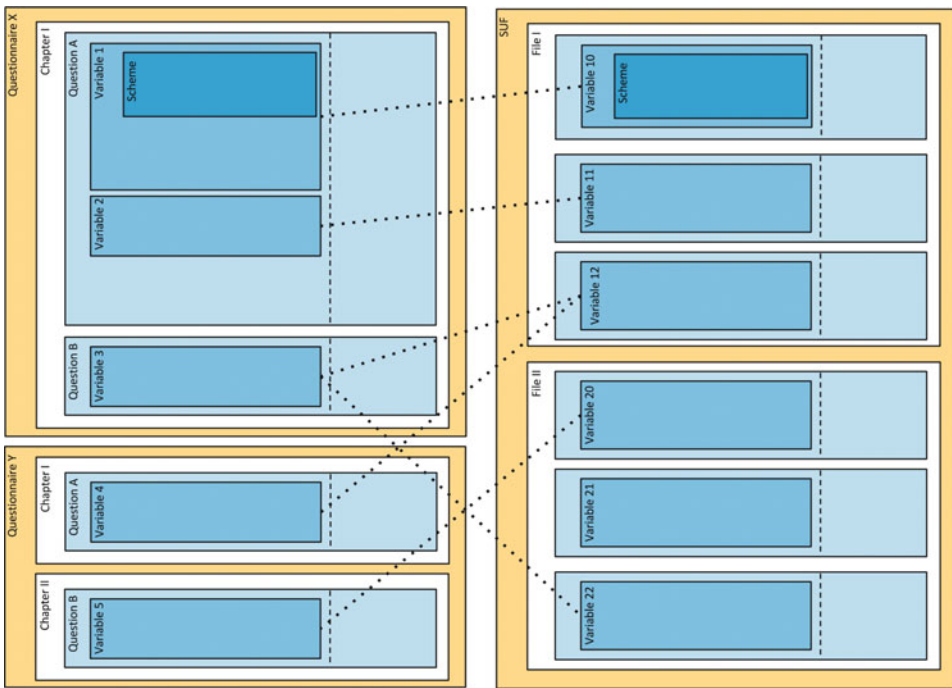
In respect to the international scientific community, all textual information in the metadata is translated into English. To minimize the translation cost and to ensure a certain degree of consistency of the translation, single text components are translated only once. This translation is re-used later on. The re-use of translations follows the logic of the re-use of metadata from survey instruments: Each response schema is translated only once. This also applies to all the text fragments within a question. All

6 The first conceptual attempt to expand the information of the survey instrument with additional fields proved to be unproductive: Although the SUFs were generated using the survey instruments, the differences are larger than one might initially suspect. For a documentation of SUFs, it is not enough to gather the survey instruments; rather, there has to be an entity for the SUFs themselves.

Figure 4 Use of objects in the database structure for documentation of questionnaires and Scientific Use Files

Object in database	Object, to be documented	
Instrument	Programming master or questionnaire with coding sheet	Whole SUF (collection of datasets)
Chapter	Chapter	File
Item (re-usable)	Question	Not used
Variable	One or (where required) more variables	One variable
Scheme (re-usable)	Response options	Value labels

Figure 5 A survey instrument (left) is connected to an SUF (right)



texts in the survey instruments themselves (e. g., headings and introductions of sections) must each be translated again, even if they have already been used in another instrument.

As an interface for translation agencies, the XML format XLIFF⁷ (XML Localization Interchange File Format) is used. In addition to the text pairs in the source and target language, this interface even comprises information about the translation workflow to some extent. The sequence of text fragments within the XLIFF files is similar to the survey instruments. Thus, despite the storage of text fragments in different tables of the database, the context that might support the translation itself is preserved.

Figure 6 shows a short instrument with only one question as a complete XLIFF file, including the already-translated fragments. The scheme (*yes/no*) and all other fragments corresponding to the question have already been translated, as is the case for the variable label *Zeitarbeit*. Only the chapter titles (*Kapitel A*) and the filter (*Falls ja, bitte weiter mit Frage 15*) have not yet been translated. These are the only two text particles that would be exported upon a request that the fragments be translated.

All previously translated fragments can be exported as a single XLIFF file. If suitable software is provided, this file can be used to support the translation process and increase the consistency of translation. For example, the free software Virtaal⁸ provides a feature to integrate the already-translated fragments in its “translation memory” and generates proposals for new texts to be translated. Figure 7 shows Virtaal presenting a translation proposal.

The translation database already contains 24,550 translated fragments with 174,262 words.

7 <http://docs.oasis-open.org/xliff/xliff-core/xliff-core.html>

8 <http://virtaal.translatehouse.org>

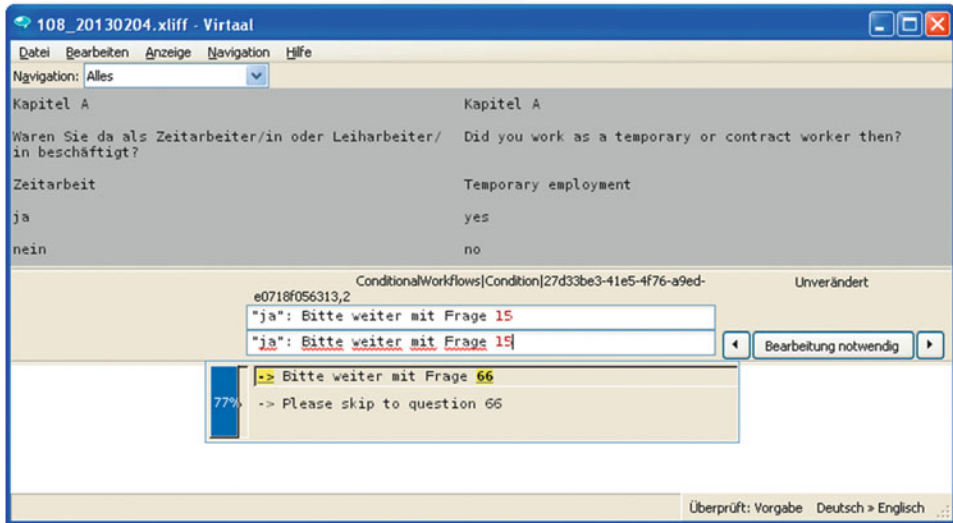
Figure 6 A short instrument with only one question as a complete XLIFF file, including the already-translated fragments

```

<?xml version="1.0" encoding="utf-8"?>
<xliff version="1.0">
  <file datatype="plaintext" original="MultiLingualeContent.fla" source-
language="de" target-language="en">
    <body>
      <trans-unit id="Chapters|Name|ac831872-1a72-412a-b8c2-ac747f39c891">
        <source>Kapitel A</source>
        <target>Kapitel A</target>
        <note>2013-02-04 17:44:22</note>
      </trans-unit>
      <trans-unit id="Questions|Text|6f379ef5-3e92-46e1-bfe9-
c20bc7c781af">
        <source>Waren Sie da als Zeitarbeiter/in oder Leiharbeiter/in
beschäftigt?</source>
        <target>Did you work as a temporary or contract worker
then?</target>
        <note>2011-04-15 08:12:24</note>
      </trans-unit>
      <trans-unit id="Variables|Label|a9e6b874-3184-4e1d-9a71-
1e94b42af990">
        <source>Zeitarbeit</source>
        <target>Temporary employment</target>
        <note>2011-04-15 08:12:24</note>
      </trans-unit>
      <trans-unit id="SchemeOptions|Label|6d097ff6-9d91-433f-b7b9-
85c589825224">
        <source>ja</source>
        <target>yes</target>
        <note>2010-12-02 12:12:02</note>
      </trans-unit>
      <trans-unit id="SchemeOptions|Label|7dc3da4a-19ed-4038-a68e-
ffe21bc8507a">
        <source>nein</source>
        <target>no</target>
        <note>2010-12-02 12:12:02</note>
      </trans-unit>
      <trans-unit id="ConditionalWorkflows|Condition|27d33be3-41e5-4f76-
a9ed-e0718f056313,2">
        <source>"ja": Bitte weiter mit Frage 15</source>
        <target>"ja": Bitte weiter mit Frage 15</target>
        <note>2013-02-04 17:47:11</note>
      </trans-unit>
    </body>
    <note>2013-02-04 17:47:11</note>
  </file>
</xliff>

```

Figure 7 A translation proposal presented by Virtaal



3 Use, Enrichment, and Maintenance of the Metadata

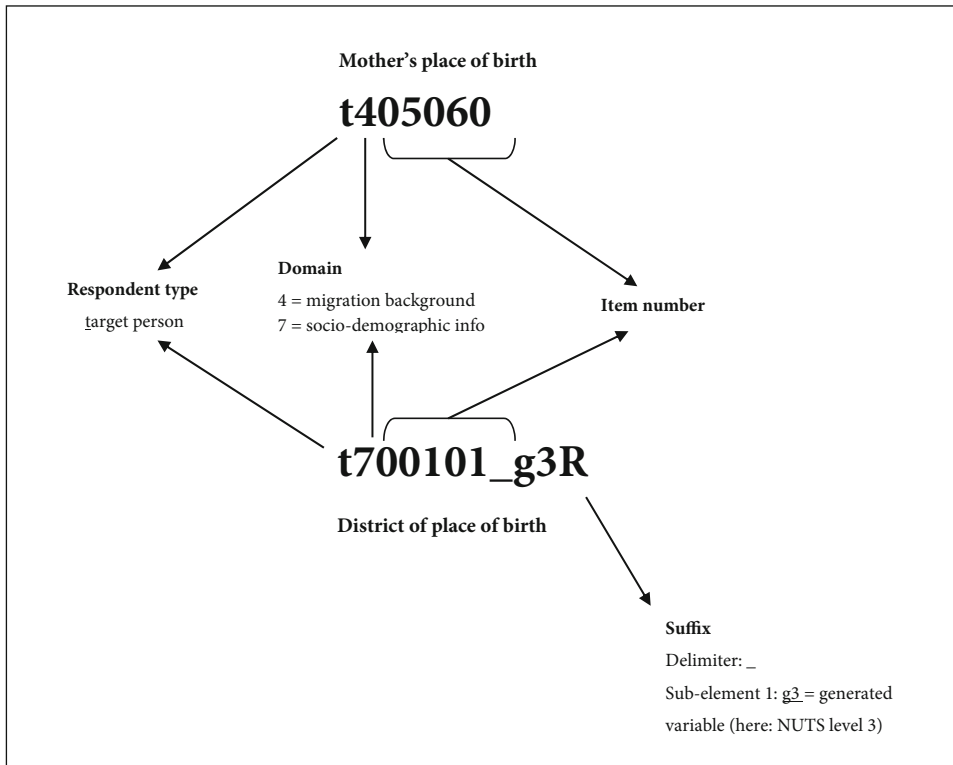
3.1 Metadata-Driven SUF Edition

There are many reasons not to use the variable names defined in the survey instruments in the datasets. Variable names derived from German words have been frequently used, which is not appropriate for an international audience and makes renaming necessary. Now, nearly system-free (non-speaking) variable names are used (see Figure 8), the name-change information is stored in the metadata, and the name changing during the edition is controlled by the metadata itself.

First, the item developers rename the variables in the survey instrument. The information is a so-called SUF variable name as a property of the question (which contains one or more variables) used in the survey instrument. This information is used to display the variable names in the generated survey instruments that accompany the SUF (so-called SUF versions). During the edition of the SUF, the old variable names are used until the datasets have the final structure. Then, a new variable name has to be determined for the (old) variable names in each file. The SUF object is initialized with a list that contains information of file names, new variable names, and old variable names.

In addition to the variable names, the order of the variables in the datasets is controlled by the metadata. Furthermore, variables that are not included in the metadata are not delivered in the final release.

Figure 8 NEPS variable naming convention (Leopold, Raab, Skopek 2012, p. 8)



Once the structure of the records is completely preserved in the metadata, the (variable and value) labels in the datasets are completely replaced with the appropriate information from the metadata during the edition. This seems risky at first glance, but as English datasets have to be produced, there is no alternative. If the labels of the German datasets are overwritten with the metadata, multistage computer-aided test loops help to ensure that labels are not overwritten in a distorting manner. These test loops are easier to handle by comparing two German text particles than by comparing a German/English text pair.

Currently—with the release of SC5 3.0.0—the NEPS metadata collection contains information on 9 SUFs with 142 files and 15,566 variables. Since the release of SUF NEPS SC4 (1.1.0), metadata of the SUF object have been frozen, and the metadata for the next release are based on a copy of the frozen object.

3.2 Mapping the Survey Design

The instruments can be documented not only as such but can be located within a multi-cohort sequence design (Blossfeld, Maurice, and Schneider, 2011, p. 14), the overall survey design of the NEPS, which consists of 6 starting cohorts and 2 additional studies. The corresponding data model is based on the following considerations:

- There are specific samples of the target population, namely newborns, children in Kindergarten and schools, students, and adults (see table “Samples” in Appendix, Figure 11).
- These samples are used at certain times (see table “Sample Waves” in Figure 11).
- At these time points, the decision as to whether the targets themselves and/or persons in the context (parents, teachers, principals) should be interviewed can be represented. These are the studies for which the survey instruments are actually developed (see table “survey” in Figure 11).
- The survey instruments can be assigned to surveys defined in such way (see table “Instrument Surveys” in Figure 11).

Furthermore, the sample itself can be assigned to a system of hierarchically structurable groups. The 6 cohorts and 2 additional studies are at the top level, and the distinction between special education and regular schools is at levels in between (see table “Sample Groups” in Figure 11). Using the information from the survey design, the survey instruments can be located in the survey matrix of the NEPSplorer (see Figure 9).

3.3 Metadata Products

As already described, the SUFs as such are products no longer conceivable without the metadata database. Whilst simple variable- and value labeling procedures could be implemented manually, translating all of this information in the datasets is not feasible with restricted manpower. However, the statistical package Stata—which is used for data edition of NEPS SUFs—is capable of quite comfortable automated access to SQL databases via Open Database Connectivity (ODBC). This enables the data editors at the NEPS Data Center to directly write information stored in the metadata database to the produced SUFs. Fortunately, Stata’s dataset format is not only well documented⁹ but also very flexible in storing meta-information. It allows for saving multilingual variable- and value labels as well as freely customizable additional texts attached to a variable via its “char” function. This enables NEPS SUFs to deliver a major surplus in usability to the data users using Stata. Not only is variable- and value

9 See <http://www.stata.com/help.cgi?dta> for details.

Figure 9 The NEPSplorer displays search results within the NEPS study design

Starting cohorts / Year		StudyDesign	SUF	2007/08	2008/09	2009/10	2010/11	2011/12	2012/13	2013/14
Early Childhood	panel						4 - 580			
Kindergarten	panel	1198					7 - 666			
Grade 5	panel	2306					5 - 721			
Grade 8	panel	2782					3 - 915	1 - 634		
First-Year Students	panel	1951								
Adults	panel	1724	1 - 43			1 - 668	1 - 823			
Organizational Reform Study in Thüringia	cross section	1910				3 - 171	3 - 170			
OS Reform in Baden-Wuerttemberg	cross section	1430					5 - 160	4 - 5		

label information seamlessly switchable between English and German using Stata's "label language" command, but old and new variable names, question texts, and subsidiary questions are also saved to a variable where applicable. This also happens multilingually, where appropriate. To make this information even more accessible to the user, the Stata program "infoquery" was written and is distributed via the NEPS website.¹⁰ It immediately displays the additional information (e. g., question text) attached to a queried variable to the user. A second program, "charren," enables the user to switch between old and new variable names. Generally, it is possible to attach even more information to variables. This, however, is restricted to the statistical package Stata. Other packages, such as SPSS, do not explicitly support (or do not document) such features. The distribution of datasets for other software is therefore limited to producing one SUF per language and statistical package.

The public web front-end for the documented metadata is an application we call NEPSplorer (see Figure 9). It locates the results of a powerful full-text- and keyword search directly in the NEPS study design and allows for extremely fast access to questions in survey instruments and variables in datasets of all already-published Scientific Use Files. The NEPSplorer is a module built for the free and open-source content management system *DotNetNuke*, which hosts the project website www.neps-data.de. The module utilizes the metadata that have been documented for every mayor study and Scientific Use File. Because the metadata are stored in a Microsoft SQL Server 2008R2, the metadata could be indexed with the powerful built-in full-text search facility, which makes fast access to huge text bodies possible in the first place. As a free alternative, *Lucene* from Apache¹¹ could also be used in other projects in which Microsoft technology is not available. For each item, every piece of textual information is indexed, such as question texts, interview instructions, variable names, possi-

10 Installation is easy with the Stata command "net install nepstools, from (<http://nocrypt.neps-data.de/stata>)." More information on this can be found in the data manuals.

11 <http://lucene.apache.org/core/>

ble answers, keywords, concepts, and so on. Based in the items that are relevant for a given query term, the interface provides the user with many pieces context information, such as the specific studies in which these items have been used and the instruments in which these items have been implemented. The same search options are also available for the variables of the Scientific Use Files. As all documentation and metadata of the Scientific Use Files are built from the information in the database, which is also used by NEPSplorer, users find the most up-to-date information directly on the website.

In addition to the direct search facility by entering a query string, scientists can also use the interface to browse through the different studies or use their choices as an additional filter for the search. Every study is located in a matrix of starting cohort versus year. The user can select any possible dimension within this matrix: A specific starting cohort, a year, or the intersection of starting cohort and year. As a special dimension, the user can also select the current SUF. As another possibility to either restrict the search or browsing results, scientists can also search by concepts. The NEPS uses about 1,500 concepts, which form a hierarchical tree. Users can select any concept they are interested in, and these concepts can be used as an additional filter.

Linking information that is also documented within the metadata database is visualized. The user can download additional documents as an example for any instrument or Scientific Use File that is part of the search results. Variables in questionnaires are linked to the corresponding variable in a Scientific Use File. A variable in a Scientific Use File is internally linked to the survey data, and first simple statistics can be seen in the NEPSplorer. Altogether, the NEPSplorer offers the user a fast and flexible user experience that can lead him or her from the very abstract study description to a first impression of the survey data.

The system is based on a user-rights management that allows the system to distinguish between different user groups. This allows us to provide different information to specific users. As a special service, we provide our data users with a preview of the instruments and Scientific Use Files that have not yet been released. Internal users of the data center are also able to have a look at any documented information.

The bilingual codebooks that are derived from the datasets and their associated metadata are a PDF-version of NEPSplorer. They include frequency counts with variable- and value labels sorted by datasets for each variable, as well as the related question, for each SUF.

The SUF versions of the survey instruments are also produced with the metadata. They are best suited to work with the data: The generated programming masters are very close to the original. Concerning the PAPI questionnaires, there are more differences between field and SUF versions because no layout information is available. The SUF version—which is well suited for the work with the data—is at least similar to the field version and contains variable names, variable labels, and numerical values assigned to the responses. The original or the SUF variables names can be displayed within the generated instruments for PAPI and CATI.

3.4 Enrichment of Metadata

At the level of questions in survey instruments, further information will be gathered in the future to facilitate the finding of questions and variables, increase the scientific quality of the documentation, and support internal processes.

Questions will be tagged with keywords from a classification system, namely the NEPS tree of constructs. The tags already used are visible in the NEPSplorer (tab “constructs”) and can be used for searching. We have begun to link our concepts to the concept of ontology of the *Gesis* (TheSoz/“Thesaurus Soziologie”) with the hope of both contributing to the idea of linked data that provide access to our metadata from many different sources and of improving the search.

Furthermore, the reference will be gathered for single questions. Thus, not will cross-references be made to other studies, but the claim “Give credit where credit is due” will be able to be redeemed.

Furthermore, the responsible working package and a contact person will be stored for single questions. This will allow for the creation of appropriate lists, which could relieve the item developers because they will not need to hold any more own documentation.

3.5 Maintenance of the Metadata and Software Environment

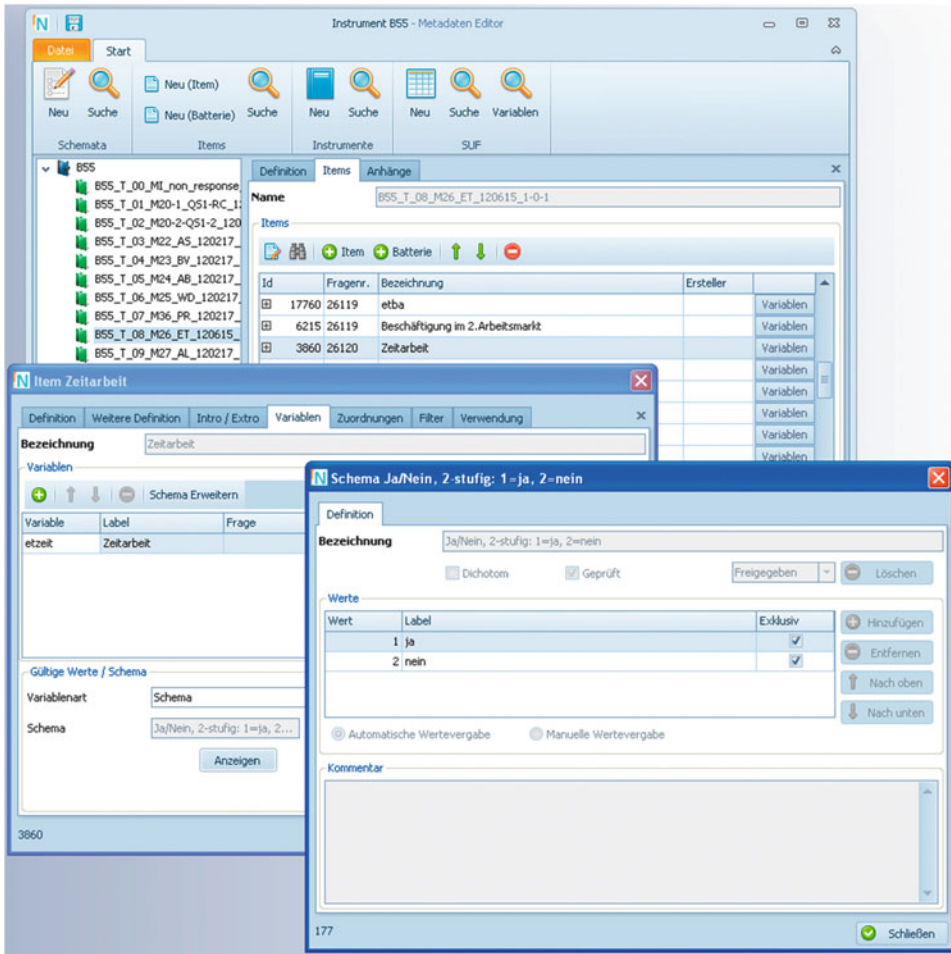
Special software has been developed to gather the metadata: the metadata editor (see Figure 10). It works as a front-end for the quite complex SQL database (see Figure 11) and allows the maintenance of almost all information from the survey instruments and SUFs. The structure of the user interface follows the re-use logic of the metadata. Schemes can be gathered for use in variables of questions. The questions are then placed in (survey) instruments. The “SUF” area allows for accessing the metadata of SUF objects. Here, it is possible to link a variable in the SUF to a variable in a survey instrument.

Large amounts of metadata can be imported using spreadsheets (e. g., Excel or LibreOffice) as alternative input interfaces. Whenever available information is already in a useful table shape, it is imported this way.

Development process

As stated above, the toolset for the NEPS Data Center is based heavily on Microsoft products. The whole software development was performed in C# within the *.NET Framework 4.0* using *Visual Studio 2010* and *Team Foundation Server 2010*. The decision to use a commercial database like *SQL Server 2008 R2* derived from the complexity of the database model. As advanced features, such as reporting for the creation of questionnaire overviews, *Extract-Transform-Load* (ETL) processes for importing and exporting, and *XML acceleration*, were requested, only the use of an enterprise-level

Figure 10 Three open windows of the NEPS metadata editor show an instrument in which the item (question) from Figure 2 is used



database like Microsoft, Oracle, or IBM DB2 was possible. In the academic world, Microsoft offers attractive conditions for universities for all these functionalities, so the decision was made to use its products. Nevertheless, all of the software that was produced in the process will be published under a license like *GNU Lesser General Public License* or *MIT*. Furthermore, if smaller studies than the NEPS are considering using the same solutions, they can use the whole infrastructure with free versions of the database product (e.g., *SQL Server 2012 Express Edition*), which is only limited in database size and scalability.

During the development process, which was split at times between the social scientists and requirement experts in Bamberg and the software developers in Frankfurt, two different environments were used to avoid leakage of data and instability of the system. The development process was handled by a combination of the Team Foundation Server 2010 (as project management and code control system), a Build server to automatically create nightly builds of the products (the Metadata Editor, the NEPS Portal including NEPSplorer, and the NEPS database itself including reports) plus a Test Server including mock data to test the functionalities of the individual products. When the software packages were, they were handed over to the IT administration of the NEPS Datacenter to roll them out into the production platform. The productive platform in Bamberg contains the productively used metadata, potentially also data within its database, and the latest stable release of the software. The nightly builds on the test server in Frankfurt were unstable and created only for the testing purposes of all included parties.

4 Outlook

The next steps in metadata development include the enrichment of the metadata with information that will help researchers to navigate through the NEPS data with a theoretical perspective: The sources of questions and key words have to be entered in many areas, which implies a lot of editorial work. Then, an infrastructure for documenting scales, which will be modelled as groups of variables, has to be established to provide materials known as scale manuals.

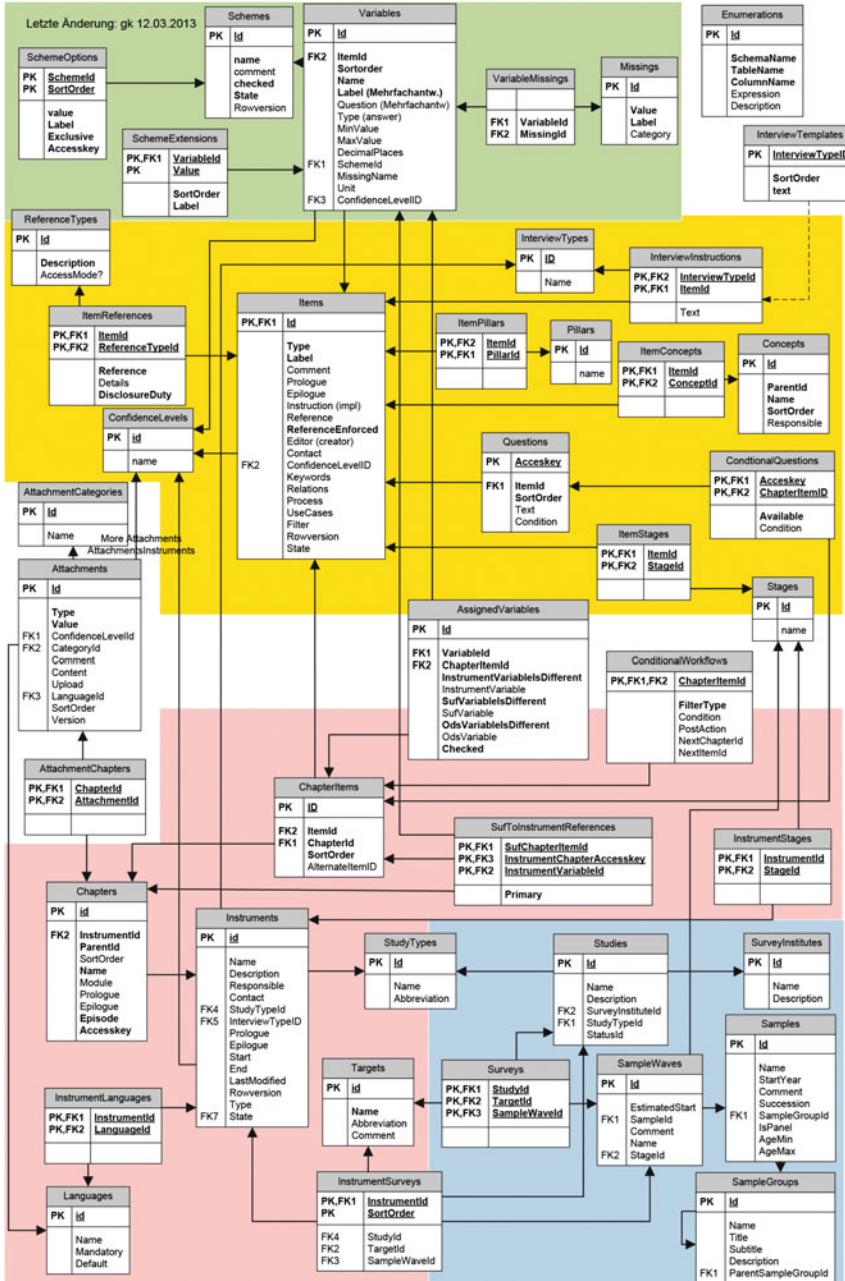
Progress in a more technical perspective would involve the development of interfaces for the DDI world. This development could benefit from community contributions in this area.

References

- Bela, D. (2016). Applied large-scale data editing (in this volume).
- Weinert, S., Artelt, C., Prenzel, M., Senkbeil, M., Ehmke, T., & Carstensen, C. H. (2011). Development of competencies across the life span. In H.-P. Blossfeld, H.-G. Roßbach, & J. von Maurice (Eds.), *Zeitschrift für Erziehungswissenschaft, 14. Education as a life-long process: The German National Educational Panel Study (NEPS)* (pp. 67–86). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Leopold, T., Raab, M., & Skopek, J. (2012). *Data manual—Starting Cohort 6: Adult education and life long learning* (NEPS Research Data Papers). Bamberg: University of Bamberg, National Educational Panel Study.

Appendix

Figure 11 The entity relationship diagram of the NEPS metadata database (Author: Gerhard Kraft)



About the authors

I. Barkow

University of Applied Sciences HTW Chur, Switzerland.

D. Bela

Leibniz Institute for Educational Trajectories (LifBi),
Bamberg, Germany.

C. Matyas

Leibniz Institute for Educational Trajectories (LifBi),
Bamberg, Germany.

M. Rittberger

German Institute for International Educational Research (DIPF),
Frankfurt am Main, Germany.

K. Wenzig

Socio-Economic Panel (SOEP),
German Institute for Economic Research (DIW Berlin),
Germany.
e-mail: kwenzig@diw.de

Applied Large-Scale Data Editing

Daniel Bela

Abstract

The dissemination of a huge collection of empirical data within the complex study framework of the National Educational Panel Study (NEPS) makes the collaborative and systematic preparation of the data indispensable. Both building up a collaborative infrastructure and committing all coworkers to principles that guide the data-preparation process are therefore crucial. In addition to leaving reported data unchanged and organizing the editing process in intermediate steps, the core principle is replicability, which is achieved via a completely syntax-based procedure using Stata®. The syntax elements of all collaborators are systematically linked to each other so that, in the last run, one press of a button generates all the scientific use data. This approach has two major advantages: It forces the staff to extensively document the process in order to make it comprehensible both at later points in time and for colleagues and reviewers. In addition, it facilitates the writing of generalized syntax, which can be reused across multiple editing projects. These guiding principles are supported by a technical framework to carry out data editing collaboratively. We came to organize the collaborative infrastructure by methods originating from software-development environments. The most important part of the infrastructure is a distributed version-control program, which enables us to keep track of any changes in syntax files. The writing of generalized syntax has resulted in an exhaustive library of additional Stata® subroutines for data editing. Due to their generality, these subroutines are shared with the scientific community to a large extent, providing data managers worldwide with convenient tools for their work in several fields of application. Furthermore, we pursue a strategy to involve all NEPS researchers in quality control. This is achieved by releasing early versions (comparable with “milestones”), enabling all other NEPS members to quickly evaluate the results of data editing during the process. An important advantage of this approach is that the data are carefully examined by many

researchers before their final release to the scientific community. This process enhances the data quality in an invaluable manner.

1 Introduction

In many research-oriented projects in the social sciences, data editing is not seen, planned, or funded as an integrated part of the data-generation process. Transferring this task to a potentially heterogeneous community of data users—which is common practice in most projects—may lead to inconsistent research findings due to diverse data editing.

Not only may every data user utilize different statistical software and data-management procedures to create ready-to-use datasets, but as a researcher, he or she is naturally interested in results and may not be an expert in the editing and preparation of (often complex) survey data. This may lead to irreproducible findings and statistical artifacts.

For the scientific community, there are several ways to avoid such issues. For instance, it is possible to force article submissions to contain syntax files that replicate all data-editing procedures. Some journal editorial boards are thinking about or have already implemented such requirements. Additionally, research projects that generate scientific use data could effectively and homogeneously integrate, prepare, and edit data prior to dissemination. The latter option, of course, is only feasible if the project is equipped with appropriate resources and staff.

In the case of the National Educational Panel Study (NEPS), data editing has been part of the project since the beginning of project plans. This has resulted in the creation of the NEPS Data Center as a unit in the methods department that holds expertise in data editing, preparation, dissemination, and documentation. Since the NEPS project is continued at the Leibniz Institute for Educational Trajectories (LIfBi), the LIfBi Research Data Center (LIfBi-RDC) succeeds the NEPS Data Center in all functions in the NEPS data-dissemination process.

Interestingly, literature discussing the best practices in data editing is not easy to find. While several international organizations have published handbooks on the topic of survey or census editing (see United Nations Department of Economic and Social Affairs (2010), United Nations Department of International Economic and Social Affairs (1984), and Inter-university Consortium for Political and Social Research (2012)), these works barely sketch out how to establish a productive workflow that effectively produces the high-quality results that these organizations propagate. In other words, these books often formulate aims of data quality and replicability but lack information on how to achieve these goals.

The current chapter briefly introduces works conducted by the LIfBi-RDC on data editing and preparation. After a short description of the initial (data) situation, methods and ideas of generalizing are introduced, discussed, and shown. This leads to a

“code of conduct” for NEPS data preparation that is to be understood as a best-practice approach of survey-data editing.

2 Motivation

The dissemination of a huge collection of empirical data within the complex study framework of the NEPS makes the collaborative and systematic preparation of the data indispensable. The NEPS provides data on six starting cohorts (Starting Cohort 1—Early Childhood through Starting Cohort 6—Adults) and two additional studies. Each of the panel cohorts is scheduled to be surveyed at least once a year. This comes to a (roughly estimated) data-dissemination time frame of two months per cohort, including data-consistency checking, correction, and editing, as well as documentation.

The workload is aggravated by the fact that the NEPS studies are not only plain cross-sectional surveys, but also implementations of complex designs. This includes—but is not limited to—the following specifics:

- a) In cohorts sampled in schools (Starting Cohort 2—Kindergarten through Starting Cohort 4—Grade 9), not only are target persons surveyed, but so, too, are several context persons (multi-informant perspective). This includes parents, educators/teachers, as well as the institutions’ headmasters.
- b) As the school cohorts’ target population “grows,” it diffuses into sub-populations. Some students remain in the regular schooling system, while others leave it to follow the vocational track. In addition, target persons may move or change schools and also have to be individually re-tracked. This often results in different instruments per sub-population and asynchronous surveying between these groups.
- c) In cohorts with an older target population (Starting Cohort 5—First-Year Students and Starting Cohort 6—Adults, in the future also Starting Cohort 4—Grade 9), biographical episodes are either retrospectively or—if not finalized in the preceding interview—recurrently surveyed. This makes the integration of panel waves even more complex.
- d) To a large extent, NEPS studies are not conventional panel surveys (i.e., surveys with congruent instruments that are repeatedly rolled out to a target population). On the contrary, the NEPS designed its surveys to be adequately fitted to the stage of each target person’s development. This results in diverse survey instruments and raw data that have to be harmonized in order to disseminate a coherent Scientific Use File for the corresponding cohort.

Despite all diversity between the different NEPS surveys, the LIfBi-RDC aims for Scientific Use Files to be as homogeneous as possible over cohorts. This aim demands great efforts not only in the LIfBi-RDC itself, but also in close cooperation with other

NEPS departments. As a reward for these efforts, data users are provided with well-documented, well-edited, and easy-to-use datasets in which as much complexity as possible has been removed from the data structure. Data-user feedback already shows that this provision is perceived and highly appreciated.

To reach and uphold this high level of data- and documentation quality, the LIfBi-RDC has implemented a structured, collaborative data-editing system. This article provides an introduction to this large-scale data-editing process and describes and explains the methods used by the LIfBi-RDC to achieve data dissemination. This includes a highly abstract, strictly syntax-driven data-editing process as well as the implementation of interfaces for collaboration with other NEPS departments.

3 Rethinking Data Editing

As described above, the LIfBi-RDC is facing several challenges in data editing. Not only is incoming data highly complex and diverse, but it is also very heterogeneous in its content. Thus, LIfBi-RDC staff cannot be familiar with the theoretical and/or empirical basis of the corresponding survey in all cases. Therefore, NEPS staff from different departments collaborates in data preparation.

This section provides an overview of how the NEPS handles this situation without much overhead in workload. The solution has been implemented using three major deviations from common data-editing procedures: implementing data preparation strictly in syntax files (by following certain style guidelines), abstraction and modularization of editing tasks (see Section 3.2 for both), and using a technically version-controlled environment to develop the corresponding syntax files (see Section 3.3). Although data editing in the LIfBi-RDC is primarily implemented in Stata^{*1} (with few excursions to R² and IBM[®] SPSS[®] Statistics³), all of these steps could be reassembled using any other general-purpose statistical package that is controllable via plain-text syntax files.

3.1 The LIfBi-RDC's Coordinating Role

The LIfBi-RDC is designed to be the interface between the NEPS studies and the data user. Thus, data documentation, user support, and data editing represent its major roles in the data-editing process. Furthermore, it has to coordinate all collaborative efforts of data preparation wherever other NEPS staff is involved.

1 <http://www.stata.com/>

2 <http://www.r-project.org/>

3 <http://www.ibm.com/software/analytics/spss/products/statistics/>

In the data-editing process, the LIfBi-RDC conducts the main work in checking, correcting, and integrating the data material that comes in from the field institutes. This is an iterative process: Once raw data are delivered by the field institute, integration begins. During this process, data errors and inconsistencies may come to light. Such problems are reported back to the field institute, and corrections that fix open issues are negotiated and carried out. Afterwards, updated and consolidated raw data are delivered to the LIfBi-RDC by the field institute, and data preparation can continue. In this process, editing is not a single person's task. On the contrary, it is an effort of a team of data managers at the LIfBi-RDC who have to collaborate efficiently and effectively within the team.

In addition to this main share of editing work, LIfBi-RDC staff invokes, coordinates, and integrates all other tasks that are to be carried out. All of these operations have to result in an "end product" that can be delivered to the scientific community in due time. To achieve this, a highly structured workflow has been imposed whose general aim is to implement parallelized steps that can be processed independently. The workflow includes exchange interfaces as well as technical solutions originating from software development.

3.2 Parallelization of Data Preparation

In many data-editing scenarios, a sequential workflow is standard: The first step in the editing process is supposed to be finalized before any subsequent steps can be worked on. With limited time in the data editing of a specific study, such a work process inevitably leads to delays. Thus, the LIfBi-RDC tries to implement parallelized working steps wherever possible.

Data editing via syntax

The first of the three main concepts mentioned at the beginning of Section 3 is not only a prerequisite for the latter two, but it also makes the whole editing process replicable: Syntax files can be archived and used to reproduce the same results in the future given the same source dataset files.

In order to collaboratively develop data-editing syntax, certain agreements about techniques and style have to be implemented. By doing so, many analogies to software development have been revealed. Correspondingly, many of the LIfBi-RDC's syntax guidelines originate from software development. It does not matter if a software engineer talks about "source code" and the social researcher instead speaks of "syntax files." Some additional analogies include "compiled program" versus "resulting dataset," "function" versus "program," as well as "variable" versus "macro."

The main paradigm involved was formulated by Hunt and Thomas (2000) in their book *The Pragmatic Programmer* and has trickled down into the programming community as "don't repeat yourself" (DRY). In writing programs, it is desirable to not

have to write a line of code multiple times. Instead, encapsulating techniques and procedures (which have to be applied multiple times) into functions makes the resulting code more easily readable and maintainable. Consequently, when a certain procedure has to be changed or corrected, only the function has to be modified. Thus, one change in the code manifests itself in several program steps. It is no longer necessary to search in every source file for changes to be applied to the offending procedure. As a downside, this maintenance-friendly removal of code redundancy is paid for dearly by a decrease in legibility.

A second directive might seem quite arbitrary to a social researcher: Data editing should take place based on generic rules rather than on individual cases. This means that coding⁴

```
replace var=<value> if (covar1==<value 1> & covar2==<value 2>)
```

is superior to

```
var=<value> if (idvar==<id>).
```

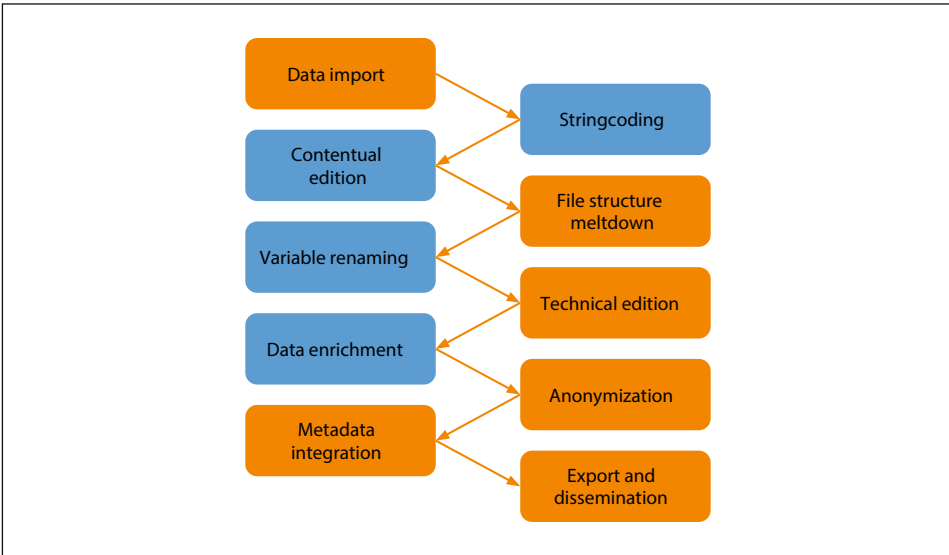
Writing rule-based editing syntax (as in the first example line) may seem cumbersome at first, but it makes the resulting code more robust and maintainable. Again, it helps with the de-duplication of written code lines—as soon as n cases meet a certain condition, $n - 1$ lines of code can be conserved at the cost of explicitly formulating the adequate condition. Beyond this, addressing a single case can be problematic. Identification via line numbers may change with the sorting order of the dataset, whereas access via a (perceived) id variable is only as strong as the persistence of the id system. As soon as id variables change—for example, due to anonymization procedures or data corrections—the code may not match any observation anymore, or, even worse, it may match a totally different observation than intended.

Modularization of Data-Editon Tasks

In the workflow of data editing for NEPS SUFs, a distinct structure of working steps took form very early on. After the first works on it had been performed, the LIfBi-RDC realized that these steps could—when arranged properly—be worked on quite independently. This results in a more parallel workflow for the data-preparation team. Figure 1 shows a schematic overview of the essential parts.

The steps are implemented in Stata syntax files and are performed on a shared directory that holds the dataset files. Each step in this directory has its own temporary sub-directory, enabling data editors to independently work on the code of a certain

4 Although this short example uses Stata syntax, it is hopefully generic enough to be easily understandable for readers not familiar to the Stata syntax language. It is important to note that, in this example, italic text in angle brackets denotes empirical values of a variable.

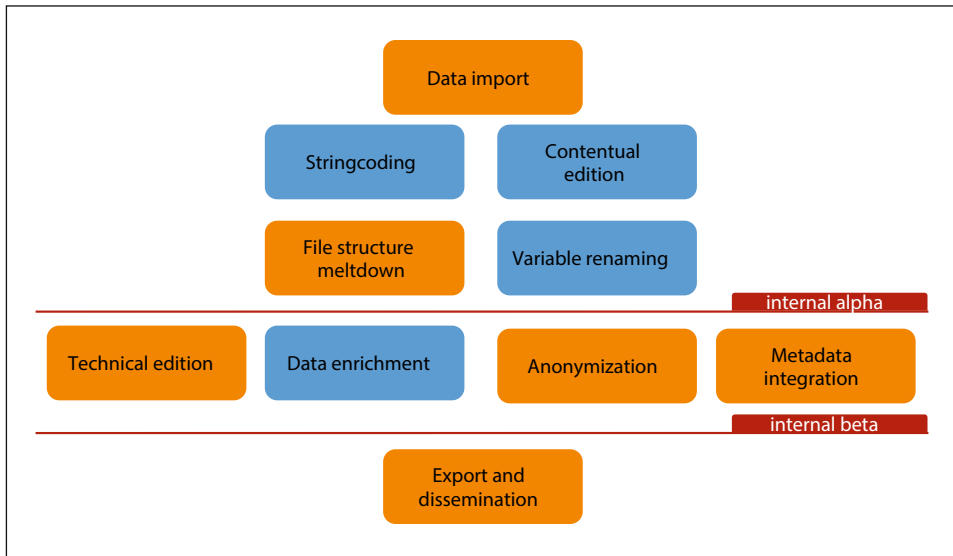
Figure 1 Sequential overview of NEPS SUF data editing

Note: Orange steps are performed by the LIfBi-RDC; blue steps are performed in collaboration with other NEPS departments.

procedure while other editors work on other sub-procedures. The script files are finally sequentially processed by a Stata wrapper syntax.

However, the workflow of establishing and creating the syntax files can also be viewed parallelized. In this view of the different steps, it is not necessary for all preceding steps to have been completed (in sequential logic) in order to work in a certain procedure. On the contrary, some steps can be developed perfectly in parallel, whereas others may need a few preceding steps to be implemented. This is illustrated in Figure 2. Additionally, as soon as a certain part of the work has been conducted and the corresponding procedures are in a stable state, the resulting data files are declared “alpha” or “beta” releases. These are disseminated inside the NEPS consortium to let all involved staff check the quality of the (preliminary) SUF data. More on these intermediate “milestones” is described in Section 3.4.

Data import: The starting point for any SUF data work is importing source-data files into a project directory shared by all data editors. Although this might sound like a trivial task, it is not: Not only is source data scattered in the file system of the NEPS data server, but identifying relevant data files is cumbersome since data incoming from the survey institutes may have been updated, withdrawn, or revised (as described briefly in Section 3.1). Consolidating these files and documenting them in the consolidation scripts is essential in order to replicate the whole process later on.

Figure 2 Overview of parallelized NEPS SUF data editing

Note: Orange steps are performed by the LifBi-RDC; blue steps are performed in collaboration with other NEPS departments.

Additionally, sub-steps of this working process include an id-change procedure as well as basic file cleansing (i. e., erasing empty files, flagging technically invalid observations) and generating an overview dataset containing all observations, which is later called CohortProfile and documents each participant’s “trajectory” in all survey waves.

Stringcoding: The first step in the editing of content is the coding of open questions into categorical variables. While the most complicated items to process in this part are occupational variables (as the coding scheme is the most complex to date) coded by the LifBi-RDC (see Munz, Wenzig and Bela in this volume), many other items are encoded by NEPS staff from other departments. In order to organize this collaboration effectively, a spreadsheet interface has been implemented for import and export (see Section 3.3 for a more detailed description).

The reason for this step being conducted at a very early stage in the editing process is that the file structure has not yet been modified. This means that all data files and variable names correspond to the versions that the NEPS staff is familiar with from internal data dissemination, making the exchange of coding information quite accessible for all involved personnel. Additionally, as the coding process may produce a large workload, it is ideal to start it at a very early point in time. Consequently, the coworkers have a larger time frame for finalizing this part.

Contentual edition: As the sole part of the editing process, content editing is autonomously performed by NEPS staff outside of the LIfBi-RDC, with colleagues from the latter only being involved in the integration of the resulting syntax files. The main reason for this is that researchers near the field phase and instruments are the only ones familiar with the theoretical constructs involved in the survey. This part of data preparation is optional and may be omitted if the NEPS colleagues do not see a need for it.

As a special issue in the context of content editing, data from competence assessments are scored autonomously by the responsible item developers. The resulting data are delivered as dataset files and concatenated appropriately by the LIfBi-RDC.

Again, the early positioning of content editing in the process is due to the fact that data editors outside the LIfBi-RDC are familiar with the unmodified file structure.

File-structure meltdown: A major part of the knowledge and considerations of the LIfBi-RDC's data-editing team manifests itself in the final data and file structure of the SUFs. The arrangement of data in panel files, cross-sectional files, and spell files is a challenge that has to be met before every data release.

Once the final data structure has been planned, the "meltdown" process is performed in two separate sub-steps. In a first preparative step, the original data files are modified and edited where appropriate to enable file joining. This includes the correction of variable labels and value labels as well as—most importantly—adjustments in the scaling of variables wherever they would not match each other between datasets.

Finally, as soon as all data sources that have to be integrated into a resulting SUF dataset are consistent with each other, files are joined together. This not only makes use of "simple" procedures, such as vertical (cross-sectional data) and horizontal (panel or spell data) concatenation, but moreover, data cells have to be filled with appropriate missing codes when a variable is present in less than all of the source files, when variables flagging the source of an observation have to be created, and in additional cases. In the case of the integration of panel data with pre-loaded information for dependent interviewing, whether or not observations from all sources match each other has to be checked. This is especially true for spell data, for which censored episodes in an interview in wave $n - 1$ have to be attached to the follow-up episode in wave n . The LIfBi-RDC requires that these data-integration steps be implemented in as standardized a manner as possible. This means that the corresponding procedures are encapsulated in programs (Stata-speak: ado-files) and perform the same tasks in all SUFs in the same way. In the future, as soon as the procedures are sophisticated and well-documented enough, the plan is to publish these programs online for broader use in other data-editing contexts. The result of this "meltdown" is the final file structure of an SUF, as is explained later in this chapter. This means that far more than 100 data files are condensed to no more than two dozen, especially in cohorts including CATI or CAPI surveys.

Variable renaming: When using a dataset, variable names based on natural language can be irritating, misleading, or (at least) difficult to understand if they are not in the user's native language. Thus, the NEPS has settled for a (nearly) system-free approach for naming variables. However, these names are mostly not used in the survey instruments and raw data. Accordingly, they have to be introduced to the datasets by renaming the original variables with a more generic name. To achieve this, the LIfBi-RDC requests new variable names from the responsible item developers via an online document. The resulting information is imported to the NEPS metadata database and automatically applied to the datasets (see Section 3.3 for a detailed explanation of metadata access; see Wenzig, Matyas, Bela, Barkow, and Rittberger in this volume for a detailed explanation of the NEPS metadata system). After this step has been completed, the SUF data files are in their final form and contain their final variable names.

Technical edition: In this procedure, data content is technically edited and cleansed. This is a strictly rule-based correction of technical data errors, inconsistencies, and (more-or-less cosmetic) blotches. It includes the recoding of missing values to comply with the NEPS standard missing codes, the resolving of non-response in multiple-response batteries, and the generation of regional variables. As a second contrast to the "contentual" editing (see above), all of these tasks are performed on already-integrated SUF data with new variable names (instead of nearly unmodified raw data). This has two essential benefits: First, redundancy is minimized because only integrated data files are processed. It is not necessary to work on all source files. Second, all data changes are documented by the produced syntax. This syntax is thereby as easy as possible for data users to understand because they are already familiar with the data structure and variable names. This point is the main reason to implement technical-editing tasks at this late time in the sequential data-editing workflow. However, these scripts have not yet been publicized. The plan is to do so as an enhancement to the NEPS metadata products in the future.

Data enrichment: As a final step of data editing, additional variables and/or dataset files are added to the cleansed and integrated SUF structure. Again, this step is a process of collaboration with staff from other NEPS departments. Standardized syntax files are delivered to the LIfBi-RDC and then integrated into the data. Section 3.3 goes into more detail on this collaboration.

Anonymization: Once data editing, integration, and enrichment have been finalized, the final SUF data have to be modified to comply with the NEPS anonymization guidelines. Dataset files are therefore split up into four versions: a master version (including all information that has been included up until now) and one version per disseminated anonymity level (*Download*, *RemoteNEPS*, and *On-site*). In each of these three stages, data are modified to reflect the appropriate level of confidentiality. A detailed description of the anonymization procedures and the underlying consid-

erations can be found in Koberg (in this volume). The modifications to the data are implemented semi-automatically: In a human-readable spreadsheet, all changes to be made are defined; this sheet is automatically read and interpreted by the anonymization Stata scripts, and modifications are performed to the data. This also includes omitting all variables that are not to be disseminated (e. g., temporary variables from data editing) from the data files.

Metadata integration: Directly before dissemination, all metadata (e. g., variable- and value labels) of the Stata datasets are erased and re-written. Although this might seem prone to failure, it is a useful step: As the NEPS aims to deliver multilingual datasets (currently in English and German), all of the metadata have to be translated and edited for (at least) all languages that differ from the dataset language originally surveyed. Moreover, SUF metadata manifest in several products, such as codebooks, the web application NEPSplorer, and generated survey instruments. In order to reliably keep these documentation products in sync with the main product (the SUF), the corresponding information has to be applied to all appliances from a “single source of information”. To achieve this, the NEPS metadata database was developed (see Wenzig et al. in this volume for details on this elaborate database system and its applications). In data editing, this database is directly used from within Stata to retrieve the metadata to be written to the dataset. A more detailed description of this process can be found in Section 3.3.

Export and dissemination: Finally, all dataset files are re-checked and exported. Not only does the checking algorithm assure the correct naming of data files and the cleansing of all unwanted notes from data, but it also includes the polishing of the data files with additional features. As NEPS SUFs are registered as *persistent identifiers* with Digital Object Identifiers (DOIs), the corresponding DOI is written as a label of the dataset together with a short, human-readable comment. This makes NEPS SUFs directly citable in publications (see Wenzig (2012) for a detailed explanation of the NEPS DOIs and a citation of NEPS SUFs). Additionally, Stata has the capability of calculating a so-called “datasignature” and saving it to the data. This signature is comparable with a check sum. Whenever a data user opens an NEPS SUF dataset, it is possible to double-check that the data have not been modified since their dissemination by the LIfBi-RDC by entering “datasignature confirm” into Stata’s command prompt. The software performs a signature check and reports if the data have been modified since dissemination. Lastly, all three dataset versions are exported into separate directories. A derived version not containing any observations is built from the *On-site* version. This spin-out is referred to as “Semantic Data Structure File.”

Once the Stata data files have been created, an automated IBM SPSS Statistics batch job is created and executed. It reads each Stata data file, marks NEPS missing-value codes as *missing*, and saves the result as an SPSS dataset file separately for each metadata language. All resulting files are ready for shipping and are disseminated via

web access or copied to the appropriate directories for remote or on-site access. Additionally, the Semantic Data Structure File is made available online without any access restrictions as a part of the documentation material.

3.3 Defining Interfaces for Collaboration

As mentioned above, in such a large-scale approach, collaborating with coworkers throughout the NEPS in order to gain optimal results is inevitable. This naturally includes diverse expertise in statistical software packages as well as a heterogeneous landscape of operating systems and other technical circumstances to be avoided. This challenge is not always easy to solve. The LIfBi-RDC, as the coordinating instance of the whole editing process, decided to go for a quite abstract approach.

This approach is based on defining common interfaces for information exchange. An “interface” in this context can be defined as any agreement between the participating coworkers on how to exchange data, metadata, syntax files, etc. To be practicable, such an interface has to meet several criteria.⁵

Comprehensibility

Every person who is supposed to be part of the collaboration process by using a specific interface should be familiar with the interface’s implications and format. This does not necessarily mean that the coworker has to fully understand how the exchanged data are derived from more complex information in the first place. However, the interface should be designed appropriately to reflect the users’ technical skills. A colleague originating from a research department who may not be very tech-savvy should not, for instance, be burdened with hundreds of lines of programming code written by a software engineer. Thus, exchange formats must be designed to meet the lowest common denominator of all coworkers.

Universality

It is possible to develop many different interface procedures for very distinct types of collaboration. When doing so, however, the interface designers run into trouble very quickly. As soon as the data-exchange procedure has to be modified, one or several defined interfaces have to be applied accordingly. Sooner or later, a vast quantity of different processes has to be maintained and updated regularly. To avoid this workload, it is much more practicable to implement deliberate interfaces in the first place. By excogitating the exchange process and the needed formats prior to implementation, it may be possible to abstract an interface design in a manner that fulfills sev-

⁵ This list may not be exhaustive; however, the named concepts have proven sufficient for implementing the NEPS editing interfaces.

eral purposes. Moreover, this abstracted interface further reduces the maintenance workload if it has already been prepared for deviations from the standardized workflow.

Platform Independence

When exchanging information or data, not all collaborators may use the same technical platform. This is not limited to operating systems (e.g., Microsoft Windows vs. Apple Mac OS vs. Linux), but also applies to additional client software: Licenses for special statistical programs may not be available, and/or coworkers may not have the necessary administration privileges to install applications. Moreover, Microsoft Office may not be available. This leads to the necessity for exchange formats to be as common as possible. Fortunately, this is not a challenge that is very difficult to meet. Nearly any situation can be resolved with plain-text files, and nearly every office suite can handle the Microsoft Office Open XML (OOXML) format. Such data formats, which are exchangeable between working units, represent the interface language of choice.

Coding of Open Answers

A major case of exchanging data in the NEPS editing framework is the coding of open answers. In many survey situations, querying information by closed-answer categories is not practicable. This is especially true when the categorical system would include hundreds of possible answers or if the respondent presumably does not know the exact category to which her or she would belong (in other words, if the respondent is not familiar with the classification system).

This problem is solved by querying open answers from the respondent and later coding these answers into an appropriate classification system. The LifBi-RDC is thereby responsible for some variables (mainly occupational variables, economic sectors, and education), but not all of them: As coding into a classification system requires expertise in the theoretical framework of the classification, many topics have to be worked on by the NEPS staff, which is responsible for survey-instrument development in the first place. These colleagues are acquainted with the theory behind their questions and therefore know best how to deal with the open questions correctly.

To deal with the diverse environment throughout the NEPS consortium, the interface for data exchange in the context of string coding has been defined as a spreadsheet format. This spreadsheet contains all (de-duplicated) open answers of a variable and, if appropriate, the content of auxiliary variables that a coder might need. In addition, empty columns are integrated for every variable a coder wants to encode. The data-editing scripts export these spreadsheets in a fully automated manner. They also watch import directories in which coded spreadsheets may already have been placed. If detected, the encoded variables are integrated into the dataset. This coding procedure may include fully-automatic or suggestion-based pre-coding of strings based on

already-known “dictionaries” of coding. A more detailed description of string coding for NEPS SUFs and the underlying theoretical and practical considerations can be found in Munz et al. (2013) in this volume.

Enrichment With Generated Variables

Beyond the coding of open answers, other measures to enrich the SUFs are to be carried out. Again, not all of these can be done at the LIfBi-RDC for the same reasons as in collaborative string coding. The simple solution implemented in the NEPS data-editing workflow is a syntax-based approach. When an NEPS researcher wants to integrate a new variable (i. e., a sum score of an item battery) into an SUF dataset, a syntax file that does so is delivered to the LIfBi-RDC. After a short review by the staff, it is translated to Stata (if written for any other statistical package) and standardized to integrate into the data-editing syntax tree. The main editing scripts search in a specific directory for these syntax files and automatically execute them. Upon execution, the standardized syntaxes are parsed and documented in log files. The generated content could be one or more variables as well as complete datasets, increasing accessibility for the data users.

Importing Metadata From the NEPS Metadata Database

As has been mentioned before, the NEPS SUFs’ metadata are completely re-written during the editing process. This happens based on the comprehensive NEPS metadata structure (see Wenzig et al. (2013) in this volume for a more detailed description of the structure’s functionality and application) and represents only one of several “outcomes” from this database. To achieve the exchange between the metadata database and the data editors, Stata’s ability to read from and write to Open Database Connectivity (ODBC) sources has proven extremely helpful. As the Microsoft SQL Server database underlying the metadata system natively supports ODBC and comes with the appropriate provider to translate queries accordingly, the data-editing scripts gain direct (read-only) access to the database.

Accessing this information is—as described earlier—used in two parts of the NEPS data editing, and it completely manages the process of renaming variables. All and only variables that contain an alias name in SUF metadata are renamed to this alias. Second (and even more important), all metadata from the datasets are completely erased and overwritten with new information. This information includes multilingual labels and additional metadata like question texts. During this process, all changes are logged, and a detailed error log is generated. Experience shows that writing (and reading) a very detailed log at the very last part of the editing often reveals errors in the metadata database as well as in the editing procedures themselves.

Collaborative-Syntax Development in Version-Controlled Environments

Wherever several persons work on shared files at the same time, problems occur regarding restrictions in accessing the same data. Nearly every researcher who has

worked with shared-network drives has had the experience of not being able to open a file because a coworker was already holding a file handle on it.

When the NEPS SUF data editing began, these problems were anticipated: A team of five to ten members would have to write the editing files, and sharing violations would arise. However, software that solves these issues comfortably (at least when authoring plain text files) is already widely used in software development. The so-called “version control system” (VCS) is able to track changes, merge results from different coworkers, and perform many other tasks. The basic concept behind this type of collaboration software is that every project member works on his or her own copy of the files to be edited, and when finished, communicates his or her changes to all other team members. This communication process is handled by VCS. In conventional VCS systems, such as the “Concurrent Versions System” (CVS)⁶ and “Apache™ Subversion™” (SVN),⁷ this happens in a server-client relationship: A version system server is the instance that assures the correct handling of versions. Clients are used to submitting changes between the members. More modern approaches leave the server instance alone and allow clients to communicate directly. These “distributed version control systems” (DVCS) mostly implement version management directly in the file system, where each client holds all versioning information. Three major packages are the most common: Mercurial,⁸ Git,⁹ and Bazaar.¹⁰

When choosing a VCS, the LIfBi-RDC had clear preferences: The system to be used should (a) work without a dedicated server instance, (b) be licensed free of charge and ideally be open-source software,¹¹ and (c) feature a graphical user interface that works without deep intrusions into the operating system. As a result, Bazaar was selected.

All SUF data-editing syntax was developed in a Bazaar repository, a shared directory that holds the version-control information. The workflow imposed can only be skimmed here. For additional information, please refer to the *Bazaar User Guide*, published by the Bazaar Developers (2013).

NEPS data editors fetch a local working copy of the scripts and work on these copies. As soon as their working step is finished, they “commit” their changes back to the repository, and other coworkers can learn what has been changed. The whole project receives its version number (called “revision”) in an incremented manner, and all changes of the commit are logged. This enables version-control software to reflect all changes in a log view (see Figure 3) and even to display the originator of every single line in a specific syntax file. All tracked changes can be withdrawn, and all preced-

6 <http://cvs.nongnu.org/>

7 <http://subversion.apache.org/>

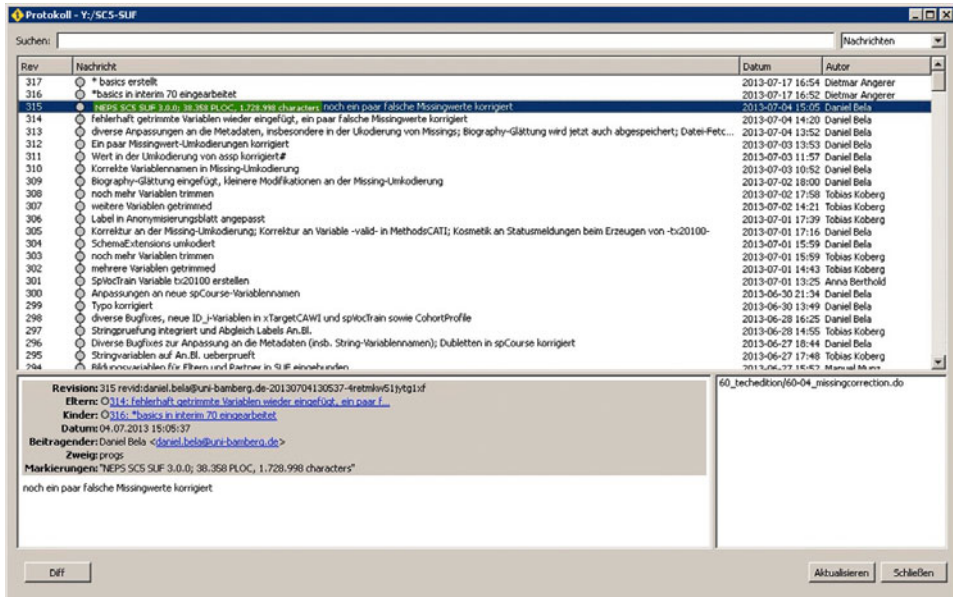
8 <http://mercurial.selenic.com/>

9 <http://git-scm.com/>

10 <http://bazaar.canonical.com/>

11 This constraint is the reason for not evaluating commercial solutions.

Figure 3 Bazaar protocol view of the data-editing scripts for SUF SC5



ing revisions can be reconstructed. All in all, version control greatly enhances collaboration in NEPS data editing and yields no more than slight adjustments in workflow.

3.4 Internal Test Releases of SUF Data

As the NEPS is organized as a consortium with many participating researchers distributed all over Germany, data editors wanted to benefit quickly from others' expertise in various fields. This benefit explicitly includes the editors' experience in data analysis and use. To achieve this, the LifBi-RDC installed two data pre-release versions of SUF data, as mentioned above. The first version (called alpha release) features the nearly completed file structure of the SUF data to be published. The NEPS coworkers are asked to provide feedback on the structure of this preliminary version. Additionally, the coworkers can begin developing syntaxes that generate additional variables and/or dataset files that should be distributed with the final release (see section "Enrichment With Generated Variables").

The beta version consists of a nearly complete SUF, including all feature-complete generated variables and datasets as well as most metadata (labels, question texts). Additionally, all feedback from the alpha release has been worked into a data editing at this point. Again, the coworkers can give feedback on the nearly finalized data, revealing metadata- or data-editing glitches in the LifBi-RDC.

As soon as the mandatory steps have been implemented in the data editing (as depicted in Figure 2), each test version is generated and distributed throughout the NEPS consortium. Although not all test versions have been generated for all published SUFs to date (mostly due to time restrictions in the data-dissemination schedule), experience shows that the feedback from the NEPS consortium has been extremely helpful in uncovering deficiencies that the data might have. By relying on this test procedure, data quality is enhanced considerably.

4 Concluding Remarks and Outlook

In this article, the extensive work the NEPS is investing in data editing has been described from a conceptual perspective. However, it is important to note that developing the appropriate procedures does not represent an end in and of itself. On the contrary, adjusting the workflow to follow the formulated guidelines greatly increases efficiency and productiveness as well as data quality when disseminating NEPS Scientific Use Files. Accordingly, the NEPS—at the time of writing—has already published a total of nine SUF packages containing data from 14 survey waves.¹² Additional data can be found in Table 1.

Future enhancements to the editing process are currently being planned, including the expansion of the NEPS metadata database to produce a semi-automated report on scales derived from survey variables. This information can additionally be integrated into the datasets. Furthermore, feeding back more information from the data to the metadata is taking place: The scripts that generate derived variables can send their generation directives to the database since these directives were already parsed in the generation process. This feedback could result in the scale handbook including the exact source code used to generate a variable (e. g., a sum or median score).

Through implementing a rather abstracted and generic editing framework, the NEPS has been able to render SUF data generation in all data products as homogeneously as possible. This upholds a standard of data quality seldom assured in smaller research projects. As a result, researchers using NEPS data can be sure that artifacts in their findings will be minimized and that the user support and training provided by the LifBi-RDC will continue.

In summary, a first draft for common data-editing procedures has been made and could be extended to a “code of conduct” for data editing in social research in the future. However, this article can only be seen as a first step leading the way. A discussion about such guidelines will hopefully arise in the scientific community and diffuse into university teaching in the long term.

12 These include the school-reform studies in Thuringia and Baden-Württemberg, each containing two waves of published (cross-sectional) data.

Table 1 Data-Editing Metrics for NEPS Scientific Use Files

Metric	SC2/SC3/SC4 1.1.0 ^a (School cohorts)	SC5 3.0.0 (Students)	SC6 3.0.0 (Adults)
<i>PLOC^b in step...</i>			
Data import	1,748	1,256	248
Stringcoding	10,260	312	7,473
Contentual edition	7,304	19,191	128
File-structure meltdown	1,089	4,195	16,644
Variable renaming	35	30	30
Technical edition	462	1,168	3,707
Data enrichment	12,752	7,198	1,343
Anonymization	1,597	1,493	1,289
Metadata integration	387	385	386
Export and dissemination	307	307	307
other	3,631	2,823	4,922
<i>Total</i>	<i>39,572</i>	<i>38,358</i>	<i>36,277</i>
No. of Bazaar revisions	584	315	1,229
No. of imported spreadsheets ^c	217	46	4 ^d
No. of original raw-data files	266	113	205
No. of resulting data files	42	23	27

Note.

^a The data editing for Starting Cohorts 2 through 4 was implemented in a joint project up to SC4 version 1.1.0, from which the presented metrics originate.

^b Abbreviation for "Physical Lines Of Code."

^c Spreadsheets imported in the string-coding procedure, resulting in coded variables.

^d Stringcoding using the spreadsheet interface was not implemented before finalizing SC6 SUF data.

References

- Bazaar Developers (2013). *Bazaar user guide*. Retrieved from <http://doc.bazaar.canonical.com/bzr.2.5/downloads/pdf-en/bzr-en-user-guide.pdf>
- Hunt, A., & Thomas, D. (2000). *The pragmatic programmer: From journeyman to master*. Boston: Addison-Wesley.
- Inter-university Consortium for Political and Social Research (ICPSR) (2012). *Guide to social science data preparation and archiving: Best practice throughout the data life cycle* (5th ed.). Ann Arbor: Institute for Social Research University of Michigan. Retrieved from <http://www.icpsr.umich.edu/files/ICPSR/access/dataprep.pdf>
- Koberg, T. *Disclosing the National Educational Panel Study* (In this volume).
- Munz, M., Wenzig, K., & Bela, D. *String coding in a generic framework* (In this volume).
- United Nations Department of Economic and Social Affairs (2010). *Handbook on population and housing census editing* (Studies in Methods No. 82). New York: United Nations. Retrieved from http://unstats.un.org/unsd/publication/SeriesF/seriesf_82rev1e.pdf
- United Nations Department of International Economic and Social Affairs (1984). *Handbook of household surveys* (Studies in Methods No. 31). New York: United Nations. Retrieved from http://unstats.un.org/unsd/publication/SeriesF/SeriesF_31E.pdf
- Wenzig, K. (2012). *NEPS-Daten mit DOIs referenzieren*. (RatSWD Working Paper Series No. 202). Berlin: Rat für Sozial- und Wirtschaftsdaten. Retrieved from http://www.ratswd.de/download/RatSWD_WP_2012/RatSWD_WP_202.pdf
- Wenzig, K., Matyas, C., Bela, D., Barkow, I., & Rittberger, M. *Management of metadata: An integrated approach to structured documentation* (In this volume).

About the author

D. Bela
Leibniz Institute for Educational Trajectories (LifBi),
Bamberg, Germany.
e-mail: daniel.bela@lifbi.de

Life-Course Data and the Longitudinal Classification of Education

Jan Skopek and Manuel Munz

Abstract

In this chapter, we present a longitudinal approach to the classification of education as applied to data from Starting Cohort 6 of the NEPS. Arguing that educational achievement is a time-dependent process involving the timing and sequence of transitions in an educational state, we examine the following two questions: 1) How can inter- and intra-individual variations of educational achievement be analytically described and compared? and 2) How can longitudinal data on educational careers be adequately measured and coded in analytically meaningful ways? We present CASMIN and ISCED-97 as helpful coding frames to capture educational achievement. Referring to life-course data from NEPS Starting Cohort 6, we present a longitudinal assignment scheme of educational attainment that we implemented in a generated transition data file called *Education*, which accompanies the Scientific Use File. *Education* provides upward transitions in ISCED and CASMIN for respondents in an easy-to-manage event-time format. Using the file, researchers can easily reconstruct the educational level measured in standard classifications for each respondent at each point in the recorded lifetime. Finally, we demonstrate the power of *Education* through two simple exemplary analyses.

1 Introduction

This chapter presents a longitudinal approach for classifying educational achievement that we have applied to recently published data from the NEPS Starting Cohort 6. Education is without a doubt one of the major resource structuring social chances and forms of participation of individuals in modern societies (Blossfeld, 1985). Since educational attainment is a predominant mechanism of status attainment and social mobility (Blau & Duncan, 1967; Müller & Mayer, 1976), it is the subject of a wide ar-

ray of research on social inequality and stratification (e.g., Shavit & Blossfeld, 1993; Breen & Jonsson, 2005; Breen et al., 2009). As a result, educational level is one of the most considered variables in empirical studies of social-science research. Hence, providing data on individuals' educational attainment is of crucial importance for survey-data providers.

Moreover, educational level is also a subject of change over time and is thus substantially intertwined with a broad range of individuals' life-course events. However, commonly used datasets provide information on education only at a certain point in time (for instance, the highest educational level of an individual at the time of interview), thereby limiting a methodologically adequate consideration of education as a time-dependent variable. In terms of interview time, this cross-sectional approach might be quite efficient for many survey contexts. However, this practice has serious shortcomings if one considers education from a more substantial perspective. First, educational attainment should be conceived as a *time-dependent process*, and individuals' highest educational level at a certain point in time only yields a temporary snapshot of the preliminary state of this process. Second, the cross-sectional approach is not capable of capturing the *timing* of educational transitions. Since educational attainment is likely to be entangled with other life domains, such as family formation, it is vital to know *when* in their lives individuals attained certain degrees and qualifications. Third, if we only know an individual's highest educational level, we actually know nothing of the individuals' history of *sequence* of educational transitions. Therefore, the standard approach is not apt to account for the heterogeneity of different educational pathways individuals might have gone through to achieve a certain educational degree.

In contrast, the National Educational Panel Study aims at collecting seamless data on individuals' educational careers. The adult panel of the NEPS, namely Starting Cohort 6, collects comprehensive longitudinal data on schooling, vocational training, adult education, and lifelong learning, which enables a measurement of individual educational levels at every time point within the observation window. To achieve this, the NEPS combines retrospective and prospective panel methods for collecting event-history data capable of tracing educational trajectories in unprecedented detail (Allmendinger et al., 2012).

However, the reconstruction of analytically meaningful categories of achieved education from fine-grained event-history data can be demanding in theoretical terms on the one hand as well as demanding, error-prone, and time-consuming in terms of data management on the other hand. Thus, this chapter focusses on two questions. First, we investigate how one can analytically describe and compare the inter- and intra-individual variation of educational attainment. We argue that standard educational classifications like CASMIN ('Comparative Analysis of Social Mobility in Industrial Nations') and ISCED ('International Standard Classification of Education') provide helpful coding frames for education and qualification. Relying on standard schemes is helpful for a vast range of research agendas and facilitates a national and

international comparability of results. However, one should be aware of certain limitations and weaknesses when working with standard classifications. Second, after having chosen a certain classification scheme, the question of how to code diverse longitudinal data on educational participation adequately into the scheme comes into play. Referring to the life-course data from Starting Cohort 6, we present a longitudinal assignment scheme of educational attainment according to CASMIN, ISCED-97, and the average years of education. Resulting from this endeavor, we generate a data file, “*Education*,” which provides user-friendly coded data on individuals’ transitions in the state space of educational classifications. In a second part of the chapter, we present the file structure as well as some first illustrative empirical examples for using the file. Finally, we provide recommendations as to how the file can be best used for empirical analyses.

2 Educational Classifications

Why rely on educational classifications at all? Several reasons for classifying educational information can be considered. Undoubtedly, a basic motivation behind the development of classifications is the ability to obtain a standardized measure of the educational status of an individual at a certain point in time.

First, classifications standardize educational information over different data-collection designs, thereby making them more comparable. Many survey studies, such as the National Educational Panel Study (NEPS), the German General Social Survey (ALLBUS), the German Ageing Survey (DEAS), and the micro census, as well as different administrative data, collect and provide data on the educational level of individuals. However, this might be done in very different ways and with very different levels of precision. For example, one could consult educational degrees in closed-category schemes of different granularity, or alternatively, one could investigate a degree in an open-question format. Moreover, the highest schooling degree and the highest vocational degree might be collected jointly in one question or separately in different questions in survey projects. With regard to time, a questionnaire design can ask only for the highest educational degree at the time of interview, or alternatively, it can collect time-related data on past educational achievements in a retrospective fashion. While the former approach provides only a very static view of educational outcomes, the latter yields a more dynamic picture of education pertaining to a life-course perspective. Additionally, beyond sheer attainment, one could consider surveying the overall participation in education involving episodes of successful (degrees) and unsuccessful (no degrees) schooling and trainings. For example, in Starting Cohort 6 of the NEPS, there are complete histories of schooling, vocational preparation, and training, which are collected retrospectively and prospectively, regardless whether or not these education episodes were finished successfully. Furthermore, degrees for successful trainings are asked for in fine-grained category schemes, enabling the re-

spondent to specify his or her degree very precisely. In this regard, educational classifications could be helpful to standardize educational information over heterogeneous measurement approaches in different survey studies and datasets. Consequently, results achieved with data from different surveys gain in comparability and reliability.

Second, classifications are helpful for comparing educational structures over disjunctive populations with different institutional and cultural settings with respect to education. In particular, classifications are extremely valuable in internationally comparative educational research in which raw country-specific degree information is almost useless for quantitative analyses. In this regard, scholars like Schneider (2008) have strongly pushed comparative research by developing international coding schemes that map country-specific educational degrees into educational classifications like the International Standard Classification of Education (ISCED, cf. UNESCO, 2006).

Third, educational classifications facilitate the comparability of educational status over time, which is particularly germane to longitudinal studies. Social, economic, and technological change in contemporary societies shapes the demand for qualifications that the educational system provides. Hence, educational certificates and institutional arrangements of education change as a natural result of societal and economic development. A prominent example in Germany is the displacement of traditional diploma degrees with bachelor and master degrees, which haven recently been introduced in Germany to comply with the agenda of the Bologna Reform. In such cases, classifications like ISCED and CASMIN provide an analytical lens for education at a higher level of abstraction. Thus, classifications as coding frames make educational attainment comparable over a changing landscape of concrete educational certificates. Of course, at the same time, classifications are limited to mirroring changes in the relative value of educational levels (like the economic and social prospects of lower secondary degrees vs. higher secondary degrees), for example, by a skill-biased technical change or an educational expansion. Nonetheless, one can detect changes and effects related to absolute levels of educational attainment on a comparable basis by using classifications.

2.1 Available classifications and measures

For our longitudinal approach of classifying educational achievement, we rely on two commonly available schemes: CASMIN (“Comparative Analysis of Social Mobility in Industrial Nations,” cf. Lüttinger & König, 1988) and ISCED-97 (“International Standard Classification of Education 1997,” cf. UNESCO, 2006). Additionally, we implement a metric scale that reflects the standardized years of education (cf. Brüderl & Diekmann, 1994: 62f). All of these schemes are widely accepted and applied in national and international empirical research. Hence, relying on such classifications is appealing to researchers if they want to obtain comparable results between studies

and populations and over time. We opted for providing both schemes, ISCED and CASMIN, because they follow different theoretical concepts and measurement ideas and consequently display specific limitations. The provision of different educational scales in a dataset not only allows for complementary perspectives on education but also facilitates robustness checks of the results. In the following section, we briefly discuss the basic ideas behind the schemes of ISCED and CASMIN.

The International Standard Classification of Education (ISCED-97)

The ISCED-97 scale utilizes information on general (schooling) as well as vocational-training information and passed vocational-preparation measures. In general, ISCED is intended to be a certification-based classification (Schneider, 2008; Schroedter, Lechert, & Lüttinger, 2006) that indicates information on general and vocational educational attainment. Both types of training appear in separate values of the classification. Table 1 shows how we composed the ISCED-97 scale from the NEPS data.

To best fit the NEPS data, we slightly modified the official ISCED-97 provided by UNESCO (2006). Since values *0A* (attendance at a kindergarten, nursery, or play schools) and *1A* (primary-school degree) of ISCED-97 were reported jointly, we collapsed them into one category. However, neither pre-school- nor primary-school attendance leads to an effective educational degree in Germany as there are 9 years of compulsory schooling across most federal states. Furthermore, the available categories for the information on the schooling level do not distinguish between pre-school and primary school. The first possible schooling level is a “degree” from primary school. Hence, pre-school and primary school are coded in a combined class *0A/1A*. We split Level 2, which, by default, contains several lower secondary schools with access to general vocational training, into two sub-levels: *2B* marking a lower secondary-school degree (“Hauptschule”) and *2A* marking an intermediate general-educational degree (“Realschulabschluss”) in Germany.

When adapting ISCED-97 to the NEPS data, we also considered two “second cycle” levels in line with the official ISCED standard. Class *4A* is coded if an individual accomplishes level *3A* after the completion of a degree from level *3B*. However, we assigned class *4B* if an individual receives an educational level of *3B* conditional on a level of *3A*. Obviously, the coding of these levels is only feasible in the presence of longitudinal data on the respondent’s educational attainment.

Finally, tertiary education levels *5A*, *5B*, and *6* were not adjusted and thus correspond to the official ISCED-97 scale.

Comparative Analysis of Social Mobility in Industrial Nations (CASMIN)

The CASMIN scale highlights different educational levels and their strong relation to socio-cultural class-construction theories and to social mobility (cf. Lüttinger & König, 1988: 6 ff). Conceptually, CASMIN combines the different general and vocational training levels with each other (see Table 2). Only the lower and higher tertiary-educational certificates of CASMIN levels *3a* and *3b* feature solely vocational-training

Table 1 NEPS Adaption of ISCED-97

Level	Degrees (English)	Degrees (German)
0/1	A Inadequately completed general education	kein Abschluss
2	B Lower general education	Haupt- oder Volksschulabschluss, Berufsvorbereitende Maßnahme
	A Intermediate general education	Mittlere Reife, Realschulabschluss
3	A Full maturity certificates (e. g., the Abitur, A-levels)	Fachhochschulreife (Fachabitur), allgemeine Hochschulreife (Abitur)
	B Basic vocational training, vocational full-time school, health-sector school (less than two years), civil servant of the lower grade, vocational basic skills	Lehre, Berufsfachschule, Fachschule des Gesundheitswesens (weniger als zwei Jahre), Beamter einfacher Dienst, berufliche Grundkenntnisse
	C Civil servants of the medium grade	Beamter mittlerer Dienst
4	A Full-maturity certificates (e. g., the Abitur, A-levels) (second cycle)	Fachhochschulreife, Hochschulreife (zweiter Bildungsweg)
	B Basic vocational training, vocational full-time school, health-sector school (less than two years), civil servant of the lower grade, vocational basic skills (second cycle)	Lehre, Berufsfachschule, Fachschule des Gesundheitswesens (weniger als zwei Jahre), Beamter einfacher Dienst, berufliche Grundkenntnisse (zweiter Bildungsweg)
5	B Diploma (vocational and other specialized academies, college of public administration), qualification of a two- or three-year health-sector school, master's/technician's qualification	Fach- und Berufsakademische Abschluss, Verwaltungsfachhochschule, Fachschule des Gesundheitswesens (mindestens zwei Jahre), Meister/Techniker, anderer Fachschulabschluss, Beamter gehobener Dienst
	A Bachelor, master, diploma, state examination, civil servants of the highest grade	Bachelor, Master, Diplom, Magister, Staatsexamen, Beamter höherer Dienst
6	Doctoral degree and postdoctoral lecture qualification	Promotion

information without a further differentiation between past school degrees. The classes reflect the typical school-leaving certificates (“Hauptschulabschluss,” “mittlere Reife,” “Hochschulreife”) in combination with information on whether a vocational training was completed or not. Furthermore, CASMIN distinguishes between lower and higher tertiary education. Contrary to the modified version of ISCED-97, the CASMIN classification of the NEPS corresponds one-to-one with the original CASMIN scheme.

Standardized Years of Education

We additionally provide a variable for standardized years of education that meets the demand for different scale levels allowing for different analytical perspectives. The metric values of this scale represent a direct and simple derivation from the CASMIN values (see Table 2). Each CASMIN level is assigned to a standardized number

Table 2 NEPS Adaption of CASMIN

Level		Standardized Years	Degrees (English)	Degrees (German)
1	A	8	Inadequately completed general education	Kein Abschluss
	B	9	General elementary education	Hauptschulabschluss ohne berufliche Ausbildung
	C	12	Basic vocational training above and beyond compulsory schooling	Hauptschulabschluss mit beruflicher Ausbildung
2	B	10	Intermediate general education	Mittlere Reife ohne berufliche Ausbildung
	A	13	Intermediate vocational qualification, or secondary programs in which general intermediate schooling is combined with vocational training	Mittlere Reife mit beruflicher Ausbildung
	c_gen	13	General maturity: full maturity certificates (e. g., the Abitur, A-levels)	Hochschulreife ohne berufliche Ausbildung
	c_voc	15	Vocational maturity: full maturity certificates including vocationally specific schooling or training	Hochschulreife mit beruflicher Ausbildung
3	A	16	Lower tertiary education: lower-level tertiary degrees, generally of shorter duration and with a vocational orientation	Fachhochschulabschluss
	B	18	Higher tertiary education: the completion of traditional, academically orientated university education	Universitätsabschluss

of years of education that are typically needed for completing the respective level (cf. Brüderl & Diekmann, 1994: 62 f). Compared with the ordinal logic of CASMIN, standardized years of education account for distances between different educational levels. Of course, one can consider other ways of recoding CASMIN classes to duration data. For example, it is possible to calculate the median of the real reported durations of related certificates.

2.2 Commonalities and Differences Between the Scales

To enhance the usability of Scientific Use Data, the NEPS aims to provide various scales involving different theoretical conceptions and properties. Correspondingly, we discuss major differences between the aforementioned scales and resulting practical implications in the following section.

Scale level

The type of scale is quite different for the discussed classifications. ISCED-97 is only scaled on an ordinal level partially because there are classes that cannot be ordered in a straightforward manner. For example, it is not possible *a priori* to distinguish whether it is better to have a certificate of level 3A (e.g., “Abitur”) or of 3B (e.g., completed basic vocational training). As a result, ISCED provides nominal information and educational attainment to some extent. CASMIN, on the other hand, is clearly ordinal. It is possible to arrange the different levels into a nicely ordinal sequence by using the standardized years of education as a sorting criterion. Moreover, the standardized years of education allow not only ordinal comparisons but also metric comparisons of educational attainment. This feature can be useful in circumstances in which the aim is to model education parsimoniously as a metric variable (dependent or independent).

Theoretical background

The presented scales and classifications follow different theoretical ideas and analytical purposes. While the ISCED-97 scale reports different tracks of education—general and vocational—in separate classes, the CASMIN scale integrates both types of education. Furthermore, contrary to CASMIN, the ISCED-97 employs information on participation at vocational-preparation measures.

ISCED-97 distinguishes between different types of vocational training by constructing the levels 3B, 3C, and 5B. Importantly, this allows for determining whether a specific educational level is achieved by initial vocational training or by further vocational education. This differentiation is especially relevant in Germany to illustrate the difference between a certificate received from a basic vocational training (3B) and certificates received at technical colleges (“Meisterabschluss,” 5B). Relying solely on CASMIN, it is not possible to detect this difference since CASMIN unites all kind of vocational training (basic and further vocational training) in some of its classes (e.g., 1b, 2a, 2c_voc).

Moreover, ISCED identifies educational degrees that have been achieved via the second cycle (classes 4A and 4B). However, the CASMIN codes do not capture these degrees; instead, it allows for differentiating between lower and higher levels of tertiary education.

To sum up, each classification has its own strengths and weaknesses. Hence, choosing the “proper” classification clearly depends on the research question. The NEPS supports its users by offering a diversity of classifications by default.

3 Longitudinal Coding of Educational Achievement in Starting Cohort 6

Using the NEPS data, it is possible to trace individual trajectories of educational attainment over the whole life span. In particular, Starting Cohort 6 collects comprehensive longitudinal data on formal education, such as schooling, vocational preparation, and vocational training, as well as data on non-formal and informal further education for an adult population (Allmendinger et al., 2012).

We utilized data on formal education to generate the transition file “*Education*,” which is introduced in the Section 4. More precisely, we exploited comprehensive retrospective information on the respondents’ school, vocational preparation, and vocational-training history. Owing to conceptual differences of ISCED and CASMIN, variables for both schemes were coded separately (see Table 3 for an overview).

Moreover, we constructed auxiliary variables to distinguish between the standardized schooling and vocational tracks that are needed to classify individuals properly into either ISCED or CASMIN because both classifications consist of schooling and vocational tracks that were either combined or included separately. The schooling track contains all surveyed information on school-leaving qualifications, the type of schools, and, for the ISCED scale, also the vocational-preparation information since vocational preparation courses are a central component of the ISCED class 2A (UNESCO, 2012). The auxiliary classes for the schooling track were constructed identically for both CASMIN and ISCED-97 and contain the values “no general education,” “lower general education,” “intermediate general education,” and “higher general education.” Using this information, we derived the ISCED classes 2B, 2A, and 3A, as well as the CASMIN classes 1b/c,¹ 2b/a, and 2c_gen/_voc in a following step.

Subsequently, we transferred all information on vocational training and the type of vocational training into the more complex vocational-training track of the ISCED. We were able to construct the ISCED-97 classes 3B, 3C, 5B, 5A, and 6 directly from the information on vocational training provided in the original dataset. The ISCED-97 second-cycle classes 4A and 4B account for the sequence of an educational career. If an individual first received a certificate relating to ISCED level 3B and afterwards received one relating to the degree of level 3A, this ultimately led to ISCED level 4A (and vice versa for 4B).

As a result, recoding the vocational training track of the ISCED scale allowed for an easy derivation of the vocational training track of the CASMIN scale. We used the constructed ISCED-97 levels 3B, 3C, and 5B as indicators for an achieved vocational-training certificate. This distinguishes between the CASMIN classes 1b/c, 2b/a, and 2c_gen/_voc in combination with the schooling information that had been created previously.

1 These classes were further distinguished by including the information of whether or not a vocational training degree was achieved.

Table 3 Assignment Scheme ISCED—CASMIN

ISCED class	ISCED label	Assigned CASMIN value
0A/1A	Inadequately completed general education	1a
2B	Lower general education	1b/1c
2A	Intermediate general education	2a/2b
3A	Full maturity certificates (e. g., the Abitur, A-levels)	2c_gen/2c_voc
3B	Basic vocational training, vocational full-time school, health-sector school (less than two years), civil servant of the lower grade, vocational basic skills	Part of indicator (vocational training available yes/no)
3C	Civil servants of the medium grade	Part of indicator (vocational training available yes/no)
4A	Full maturity certificates (e. g., the Abitur, A-levels) (second cycle)	2c_voc
4B	Basic vocational training, vocational full-time school, health-sector school (less than two years), civil servant of the lower grade, vocational basic skills (second cycle)	2c_voc
5B	Diploma (vocational and other specialized academies, college of public administration), qualification of a two- or three-year health-sector school, master's/technician's qualification	Part of indicator (vocational training available yes/no)
5A	Bachelor, master, diploma, state examination, civil servants of the highest grade	3a/3b
6	Doctoral degree and postdoctoral lecture qualification	3a/3b

The CASMIN classes *3a* and *3b* (lower- and upper-tertiary educational level) were rearranged by using the original information on the tertiary-educational certificates from the original dataset on the vocational-training information since the ISCED class 5A combines the lower- and upper-tertiary educational degree.

We assigned the ISCED classes *4A* and *4B* to the CAMIN class *2c_voc*. Moreover, we coded the levels *0A/1A* for ISCED and *1a* for CASMIN if neither a schooling degree nor a vocational-training certificate had been achieved. These constructed auxiliary tracks were combined (CASMIN) and sorted (ISCED) to generate the CASMIN and the ISCED scale, respectively.

We included the CASMIN and the ISCED classifications in the Education dataset after the construction of both classifications to provide a longitudinal dataset with the educational career of the target persons.

4 The Transition File “Education”

The generated file *Education* provides longitudinal information on effective transitions in respondents’ educational careers. Hence, the file comprises only those respondents who had achieved an educational degree at least in lower secondary education at the time of the interview. On the other hand, respondents who (1) did not finish compulsory schooling (i. e., have no educational degree) or (2) did not provide sufficient information to reconstruct an educational degree are excluded from the file. We coded the transitions in a long event-time format, meaning that each row corresponds to a transition in at least one classification (CASMIN and/or ISCED-97). Hence, one respondent might have multiple entries in case of multiple transitions. Variables in month and year of the transition specify the event time. We only consider upward educational transitions in CASMIN levels and upward as well as lateral transitions in ISCED-97 levels because CASMIN is an ordinal scheme, whereas ISCED-97 has some nominal elements. Since ISCED-97 and CASMIN follow different concepts, some educational transitions (approximately 7% in these data) are effective for only one of these classifications.

Table 4 provides an overview to the file’s variables. There are variables coding the highest CASMIN level and the recent or highest ISCED level accomplished by the transition as well as an indicator for the classification scheme for which the transition

Table 4 Variables in File Education

Variable	Coding	Description
<i>ID_t</i>	(number)	Identifier of respondent. Needed for merging respondent data.
<i>number</i>	1–6	Number used for sorting transition.
<i>splink</i>	(number)	Identifier of the respondent’s school, vocational preparation, or vocational training spell that caused the transition.
<i>sptype</i>	22, 23, 24	Type of spell that caused the transition. Possible types are school (=22), vocational preparation (=23), and vocational training (=24).
<i>datem</i>	1–12	Month of the transition.
<i>datey</i>	(calendar year)	Year of the transition.
<i>tx28101</i>	0–8	Highest CASMIN level achieved by the transition: level 1a (=0) up to level 3b (=8).
<i>tx28102</i>	0–18	Standardized years of education as a function of CASMIN level.
<i>tx28103</i>	1–10	Recent/highest ISCED level achieved by the transition: level 2B (=1) up to level 6 (=10).
<i>tx28109</i>	1–3	Indicator variable for the classification scheme for which the transition is effective: only CASMIN (=1), only ISCED (=2), both (=3).

was effective. Additionally, there is a month- and a year variable for the date of the transition as well as identifier variables for the respondent and the spell that caused the transition.

Table 5 provides an exemplary data snapshot illustrating the basic structure of the file, which is very easy to grasp. The snapshot depicts the educational transitions over time for two respondents. The first respondent ($ID_t = 8000507$) obtained a lower secondary degree (“Hauptschulabschluss”) in March 1966. Consequently, both the CASMIN ($tx28101$) and the ISCED-97 ($tx28103$) variables assume a value of 1. The variable $tx28109$ indicates that a change took place in both classification schemes (denoted by the value 3). This always applies to the first event spell of a respondent in this dataset. In September 1969 (second event spell), the respondent completed a vocational training (“Lehre”). Hence, CASMIN increases to a value of two (lower secondary degree with completed vocational training) and ISCED-97 to a value of four (basic vocational training). Because this upward transition concerns both classifications, the indicator variable $tx28109$ is again 3. Three years later (September 1972), the respondent experienced a vocational upward transition (e.g., master’s qualification, “Meisterabschluss”). Only the ISCED-97 scheme captures this transition ($tx28103$ increases from value 4 to 8, i. e., diploma from vocational and other academies, college of public administration, etc.); CASMIN remains at the value 2 because CASMIN does not distinguish between basic and advanced vocational trainings. As a result, $tx28109$ is set to a value of 2, meaning that only ISCED-97 changed its value. The reverse is true for the fourth (and final) event spell of this respondent, in which an educational upward transition is recorded. This change is effective only for the CASMIN classification. The corresponding value of CASMIN ($tx28101$) is 6 (full maturity certificates including vocationally specific schooling or training), indicating that the respondent has attained an A-level qualification (or equivalent) in addition to the vocational training that had already been completed. Therefore, $tx28109$ has the value 1, denoting a change only in the CASMIN scheme. The variable $sptype$ specifies the kind of spell that initiated the transition.

To give some summarizing statistics, *Education* contains 26,094 transitions for 11,505 out of 11,649 total respondents in the first NEPS wave of Starting Cohort 6; thus, 98.8% of all respondents have any determinable educational degree. About 92.9% of all transitions are effective in both schemes. Since ISCED-97 is partly sharper than CASMIN, 5.8% of transitions are only effective in ISCED, while only 1.3% of them are effective solely in CASMIN. The vast majority of respondents have two transitions, while about 5% of them have four up to a maximum of six transitions.

Table 5 Exemplary Data Snapshot

<i>ID_t</i>	<i>Splink</i>	<i>datem</i>	<i>Datey</i>	<i>tx28101</i>	<i>tx28103</i>	<i>tx28109</i>
8000507	220001	3	1966	1	1	3
8000507	240001	9	1969	2	4	3
8000507	240002	9	1972	2	8	2
8000507	220002	9	1974	6	8	1
8000512	220001	8	1968	1	1	3
8000512	240002	9	1974	2	4	3
8000512	240004	8	1986	7	9	3

5 Examples

The coded data on educational transitions are probably most useful for reconstructing the temporal order of states in educational attainment and other life-domain processes in the context of longitudinal analyses. To illustrate working with the *Education*, we present two simple exploratory analyses that make substantial use of the transition data recorded in the file. For the sake of demonstration, we restrict the following to CASMIN transitions only. In a first step, we inspect the prevalence of educational states over the life span. Separately for both genders, we calculate the distribution of the highest levels of education over each month of the life span, paint probability plots, and state distribution plots for visualization. This gives us an idea regarding how educational achievement is structured in the life courses of men and women at an aggregate level. Thus, we actually take on a kind of macro perspective on the process of education. However, this perspective does not tell us anything about how pathways of education evolve on an individual basis. In other words, a macro perspective hides individual educational transitions. To approach a pathway perspective on education, we investigate probabilities of transitions between different educational classes in a second step. Once again, we do this separately for men and women. For simplicity, we run all analyses without weights.

5.1 Distribution of Educational States Along the Life Course

In our first analysis, we explore how educational levels are distributed over life-course age. If the file *Education* is used, it is only necessary to do some minor data preparation for this kind of analysis. We included the whole sample of the first NEPS wave of Starting Cohort 6 (2009/2010), which is provided in the Scientific Use File (NEPS

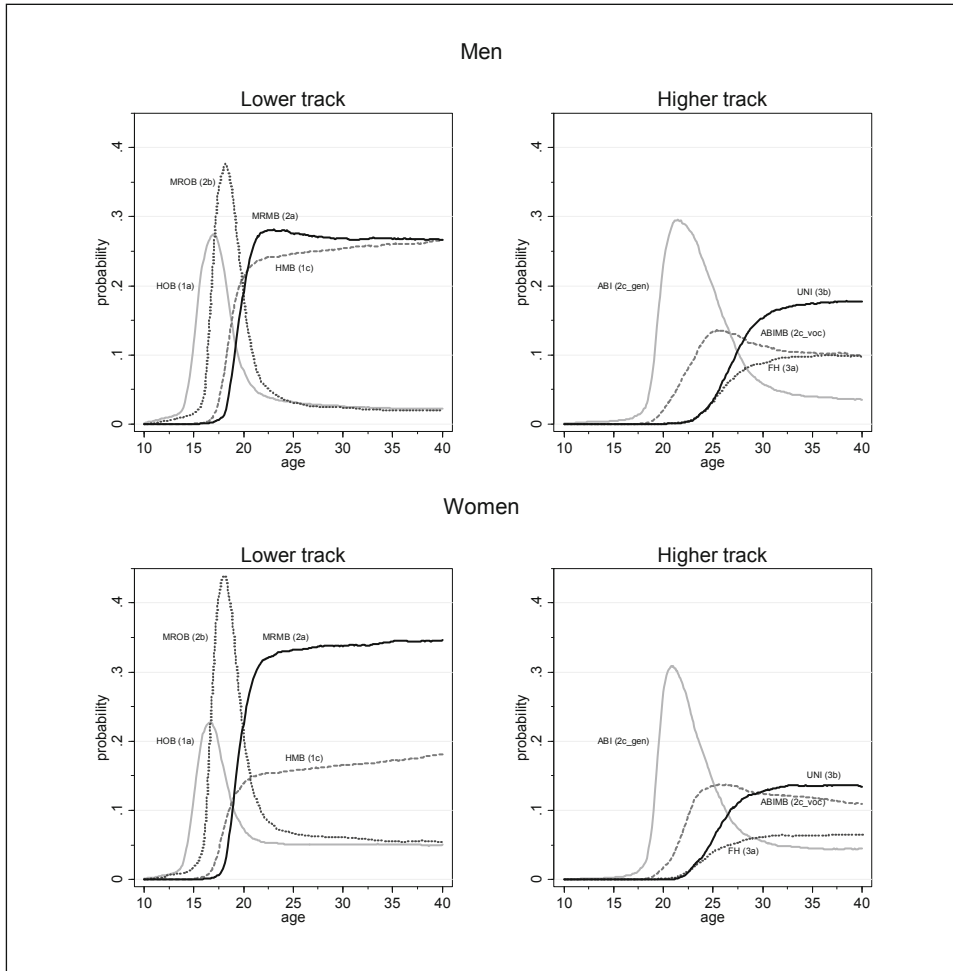
SC6 1.0.0).² The sample consists of 11,649 persons born from 1944 to 1986, 50.95 % of whom are women. Using the Education file, we determined the highest level of education (CASMIN) at each month of life for each man and each woman of the sample. For example, if a man is aged 30 at the time of the interview, we looked up the highest CASMIN level for each of his $30 \times 12 = 360$ months of lifetime. Afterwards, we aggregated over all respondents by gender and life month by computing the fraction of each CASMIN level observed in each gender in each life month. We used this as an estimator for the gender-specific probability of being in a CASMIN state at some age.

Since we are working with retrospective data from an age-heterogeneous sample ranging from individuals between age 23 and 64 at the time of the interview, there is an increasing amount of right censoring at age 23 and above. Consequently, this leads to distributions of states that are increasingly biased towards older birth cohorts with age. To alleviate this selectivity issue, we restricted our analysis to a maximum age of 40. We slightly relabeled the CASMIN levels to enhance the readability: HOB (“Hauptschule ohne Berufsausbildung,” lower secondary) refers to level *1b*, HMB (“Hauptschule mit Berufsausbildung,” lower secondary plus vocational training) to *1c*, MROB (“Mittlere Reife ohne Berufsausbildung,” intermediate secondary) to *2b*, MRMB (“Mittlere Reife mit Berufsausbildung,” intermediate secondary plus vocational training) to *2a*, ABI (“Abitur,” upper secondary) to *2c_gen*, ABIMB (“Abitur mit Berufsausbildung,” upper secondary plus vocational training) to *2c_voc*, FH (“Fachhochschule,” lower tertiary) to *3a*, and UNI (“Universität,” higher tertiary) to *3b*. Finally, we rescaled months to years to obtain a plain representation of time.

Figure 1 shows the empirical probabilities of the different CASMIN states over life age. Probability plots are drawn separately for gender and by a lower (CASMIN < *2c_gen*) and higher (CASMIN > *2a*) educational track. Additionally, we plotted these probabilities in a stacked fashion, resulting in a state-distribution plot shown in Figure 2. When inspecting the probabilities of CASMIN levels in Figure 1, we notice that men and women share very similar shapes in the age-dependent probability of having HOB and MROB education. Estimated probabilities for both sexes peak at an age of around 17 for HOB and around 18 for MROB. However, women possess a higher probability of having those levels over broad intervals in the life course. Levels with added vocational training, namely HMB and MRMB, are increasingly crowding out HOB and MROB along age. Interestingly, men show higher probabilities of having achieved HMB compared with women, and women show higher probabilities of having achieved an MRMB compared with men. In general, after jumping up to high levels at the beginning of the twenties, the probability of MRMB slightly decreases, and the probability of HMB increases again. An explanation for the lowering probability of MRMB could be that persons move to upper educational levels, for example, by beginning tertiary education. However, as noticed above, we are not able to discern educational-career processes from cohort selectivity in this analysis. The lifted prob-

2 Doi:10.5157/NEPS:SC6:1.0.0.

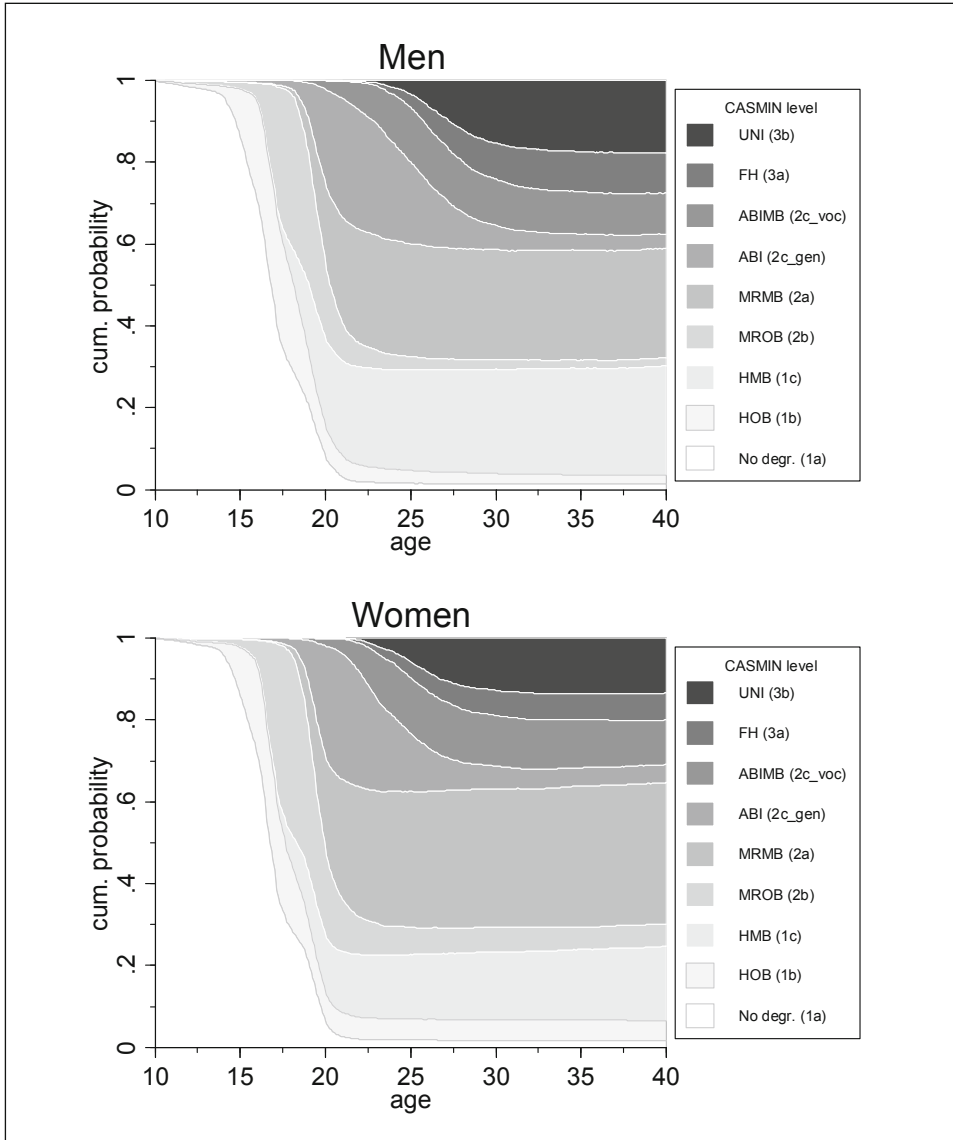
Figure 1 Empirical probabilities for being in a CASMIN state by life age and gender. Calculations are based on the CASMIN trajectories of 5,714 men and 5,935 women. Data from NEPS SC6 1.0.0, own calculations using *Education* file



ability of HMB in higher ages, in particular, is suspicious and appears to be an artifact of cohort selectivity that increases with the age axis.

Finally, we now take a look at the higher-track levels. The curve of empirical probabilities of ABI is again very comparable between men and women. As was the case for the schooling education levels in the lower track (HOB and MROB), probability increases, peaks at some age (~21 years), and then decreases again rapidly. Interestingly, the curve is a bit bolder for men, suggesting that they stay longer in the ABI state. This is an obvious result of military service that was obligatory in Germany for

Figure 2 CASMIN state distributions over life age and gender. State distribution plots that stack probabilities plotted in Figure 1. Calculations are based on the CASMIN trajectories of 5,714 men and 5,935 women. Data from NEPS SC6 1.0.0, own calculations using *Education* file



men and consequently delayed males' entry into post-secondary education. Consistent with this finding, the curves for tertiary education (FH and UNI) are shifted slightly to the right for men compared with women. However, compared with women, men reveal a higher probability of having tertiary levels of education at age 27 and above. On the other hand, there is some evidence that women are more likely to have an upper-secondary degree with additional vocational training (ABIMB). Finally, by additionally including probabilities for no degree (1a), Figure 2 visualizes the CASMIN state distribution over age in a denser manner for men and women.

5.2 Transitions Between Educational States

In our second example, we are interested in examining the transition probabilities of CASMIN levels. Contrary to the state distributions investigated above, analyzing transition probabilities provides insights into pathways to educational levels by assessing how likely a transition to the university level is to occur given that the 'Abitur' has been achieved before. Since each row of the *Education* file represent a piece of transition data, the preparation for his investigation was simple. Using *Education*, we simply generated an additional variable containing the educational state before the transition. It is important to note that persons only enter the file if they had at least one transition. Consequently, we coded the previous state for first transitions to "no degree." In the next step, we conducted a cross tabulation for obtaining the empirical transition probabilities between previous and actual state. Tables 6 and 7 present the results, separated by men and women. Additionally, Tables A1 and A2 in the Appendix inform about the absolute frequencies of transitions. Cells in transition matrices containing a dash represent transitions that are impossible by the design of the *Education* file.

When looking at the absolute frequencies of transitions (Tables A1 and A2), it is not surprising to find that transitions from no degree to any degree represent the majority of all transitions (~45 % of all men's and ~48 % of all women's transitions). Of course, this is the first transition, and almost all individuals in the sample have at least one transition. While ABI as the first transition has a probability of about 24–25 % for both men and women, we find gender differences with regard "Hauptschule" and "Mittlere Reife." Compared with women, men are more likely to achieve an HOB in the first transition. A few cases transition directly from no degree to vocational degrees like UNI or FH. These individuals most probably did not report valid school episodes, and we were thus not able to capture any transition to a school-level degree. Subsequent to an HOB level, vocational training leading to HMB is most likely. It is important to note that men go on to vocational training in the next step more often compared with women in this regard, whereas women go on to intermediate secondary training (MROB) more often compared with men. Women are also more likely to transition from HOB to ABI, albeit at a low level. Similarly, intermediate sec-

Table 6 Empirical Probabilities of Level Transitions in CASMIN (Men, Row Percentages)

CASMIN	Transition... from	to									
		1a	1b	1c	2b	2a	2c_gen	2c_voc	3a	3b	Total
No degree	1a	–	33.46	2.94	39.52	0.09	23.71	0.23	0.02	0.04	100.00
HOB	1b	–	–	80.06	14.43	1.72	3.10	0.40	0.00	0.29	100.00
HMB	1c	–	–	–	–	64.83	–	27.97	5.08	2.12	100.00
MROB	2b	–	–	–	–	75.98	20.42	2.41	0.89	0.30	100.00
MRMB	2a	–	–	–	–	–	–	73.12	20.95	5.93	100.00
ABI	2c_gen	–	–	–	–	–	–	36.50	15.00	48.51	100.00
ABIMB	2c_voc	–	–	–	–	–	–	–	62.24	37.76	100.00
FH	3a	–	–	–	–	–	–	–	–	100.00	100.00
UNI	3b	–	–	–	–	–	–	–	–	–	–
Total		–	15.21	12.56	19.98	15.97	15.09	8.57	4.88	7.74	100.00

Note. Calculations based on 12,410 CASMIN transitions of men coded in the *Education* file of the Scientific Use File of NEPS Starting Cohort 6 (Data: NEPS SC6 1.0.0, own calculations).

ondary education (MROB) is mostly followed by vocational training (MRMB), with transition probabilities being quite comparable between men and women (76–79%). However, there is a remarkable probability of 19–20% for proceeding with ABI after MROB.

The observed gender differences are quite interesting when it comes to the upper-secondary track. Proceeding to an upper tertiary degree (UNI) is most likely for men, conditional upon having an upper secondary degree (ABI). On the other hand, women with an ABI most likely turn to vocational training in the next step. In addition, the probabilities of proceeding with tertiary education in general (FH and UNI) are lower for women in absolute terms. Although it is more likely for women possessing levels of MRMB and ABIMB to transition to UNI than it is for men, these transitions occur relatively seldom. Hence, our pathway analysis corroborates our result obtained from the first example regarding the lower tertiary state probabilities of women. Finally, the transition probability from FH to UNI is 100%. This is trivial since the *Education* file comprises only upward transitions in CASMIN.

Table 7 Empirical Probabilities of Level Transitions in CASMIN (Women, Row Percentages)

CASMIN	Transition... from	to									Total
		1a	1b	1c	2b	2a	2c_gen	2c_voc	3a	3b	
No degree	1a	–	28.64	1.68	44.44	0.05	24.93	0.19	0.02	0.05	100.00
HOB	1b	–	–	63.92	27.56	1.14	6.61	0.50	0.21	0.07	100.00
HMB	1c	–	–	–	–	69.33	–	21.33	4.00	5.33	100.00
MROB	2b	–	–	–	–	78.58	18.61	1.24	0.97	0.60	100.00
MRMB	2a	–	–	–	–	–	–	59.01	31.98	9.01	100.00
ABI	2c_gen	–	–	–	–	–	–	47.14	12.73	40.13	100.00
ABIMB	2c_voc	–	–	–	–	–	–	–	46.35	53.65	100.00
FH	3a	–	–	–	–	–	–	–	–	100.00	100.00
UNI	3b	–	–	–	–	–	–	–	–	–	–
Total		–	13.76	8.20	24.53	17.83	16.82	8.20	3.52	7.14	100.00

Note. Calculations based on 12,168 CASMIN transitions of women coded in the *Education* file of the Scientific Use File of NEPS Starting Cohort 6 (Data: NEPS SC6 1.0.0, own calculations).

6 Conclusion

In this chapter, we proposed a longitudinal approach to the classification of education as applied to data from Starting Cohort 6 of the NEPS. Arguing that educational attainment is a time-dependent process involving timing and the sequence of transitions in an educational state space, we asked two questions: 1) How can one analytically describe and compare inter- and intra-individual variations of educational attainment? and 2) How can one adequately measure and code longitudinal data on educational careers in analytically meaningful ways? With CASMIN and ISCED-97, we presented two classifications as helpful coding frames for measuring educational attainment. We highlighted differences and commonalities between both schemes. Referring to life-course data of NEPS Starting Cohort 6, we presented a longitudinal assignment scheme of educational attainment that we implemented in a generated transition file called *Education*, which accompanied the Scientific Use File package of Starting Cohort 6. Using this file, researchers can easily reconstruct the educational level measured in standard classifications for each respondent at each point in his or her recorded lifetime. Finally, we demonstrated the power of *Education* via two simple exemplary analyses.

One can consider a variety of other research scenarios for which *Education* might be a highly useful dataset. This does not depend on whether the educational level is

considered as a dependent or independent variable. For example, if one is interested in the timing of the first marriage and has a specific hypothesis that educational attainment affects entry into marriage, it would be preferential to consider educational level as a timing-varying covariate. Our file *Education* easily provides this data. If one is interested in respondents' highest-achieved educational level, the last entry in file can be consulted. Alternatively, if one is interested in the level achieved first in the life course, the first entry of a respondent can be consulted.

Of course, some words of caution are warranted. It is important to be aware that *Education* is restricted to upward transitions only (as well as some lateral transitions in ISCED). Hence, if a respondent earns a university degree and then receives a non-tertiary vocational degree (e.g., "Lehre"), this transition—albeit potentially interesting for analyzing educational downgrading—does not enter the file. Moreover, our approach only considers 'successful' education, namely graduation. Hence, if one is interested in participation in education as compared with the effective attainment of a degree, the file might be of limited value.

Taken together, we would like to emphasize that the *Education* file provides just one specific perspective on education in the life course. We chose this perspective because we believe that it will serve most of the analytical requirements well. Nevertheless, it would be wonderful to complement this with alternative perspectives. Researchers are welcome to contribute by constructing and publishing additional files in the future.

References

- Allmendinger, J., Kleinert, C., Antoni, M., Christoph, B., Drasch, K., Janik, F., ... Ruland, M. (2011). Adult education and lifelong learning. In H.-P. Blossfeld, H.-G. Roßbach, & J. von Maurice (Eds.), *Zeitschrift für Erziehungswissenschaft, 14. Education as a lifelong process: The German National Educational Panel Study (NEPS)* (pp. 283–299). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Blau, P. B., & Duncan, O. D. (1967). *The American occupational structure*. New York: Wiley.
- Blossfeld, H.-P. (1985). *Bildungsexpansion und Berufschancen: Empirische Analysen zur Lage der Berufsanfänger in der Bundesrepublik*. Frankfurt: Campus.
- Breen, R., & Jonsson, J. O. (2005). Inequality of opportunity in comparative perspective: Recent research on educational attainment and social mobility. *Annual Review of Sociology, 31*, 223–243.
- Breen, R., Luijkx, R., Müller, W., & Pollak, R. (2009). Nonpersistent inequality in educational attainment: Evidence from eight European countries. *American Journal of Sociology, 114*(5), 1475–1521.
- Brüderl, J., & Diekmann, A. (1994). Bildung, Geburtskohorte und Heiratsalter: Eine vergleichende Untersuchung des Heiratsverhaltens in Westdeutschland, Ostdeutschland und den Vereinigten Staaten. *Zeitschrift für Soziologie, 23*(1), 56–73.

- Lüttinger, P., & König, W. (1988). Die Entwicklung einer international vergleichbaren Klassifikation für Bildungssysteme. *ZUMA Nachrichten*, 12(22), 1–14.
- Müller, W., & Mayer, K. U. (1976). *Chancengleichheit durch Bildung? Untersuchungen über den Zusammenhang von Ausbildungsabschlüssen und Berufsstatus*. Stuttgart: Klett.
- Schneider, S. (2008). *The international standard classification of education: An evaluation of content and criterion validity for 15 European countries*. Mannheim: University of Mannheim.
- Schroedter, J., Lechert, Y., & Lüttinger, P. (2006). *Die Umsetzung der Bildungsskala ISCED-1997 für die Volkszählung 1970, die Mikrozensus-Zusatzerhebung 1971 und die Mikrozensus 1976–2004* (ZUMA-Methodenbericht 2006/08). Retrieved from http://www.gesis.org/fileadmin/upload/forschung/publikationen/gesis_reihen/gesis_methodenberichte/2006/06_08_Schroedter.pdf
- Shavit, Y., & Blossfeld, H. P. (1993). *Persistent inequality: Changing educational attainment in thirteen countries*. Boulder: Westview Press.
- UNESCO (2006). *International standard classification of education 1997*. Retrieved from <http://www.uis.unesco.org/Library/Documents/isc97-en.pdf>
- UNESCO (2012). *Mapping of national educational qualifications: Germany*. Retrieved from http://www.uis.unesco.org/Education/ISCEDMappings/Documents/North%20America%20and%20Western%20Europe/Germany_ISCED_mapping.xls

Appendix

Table A1 Frequencies of CASMIN Transitions (Men)

CASMIN	Transition... from	to									
		1a	1b	1c	2b	2a	2c_gen	2c_voc	3a	3b	Total
No degree	1a	–	1,887	166	2,229	5	1,337	13	1	2	5,640
HOB	1b	–	–	1,393	251	30	54	7	0	5	1,740
HMB	1c	–	–	–	–	153	–	66	12	5	236
MROB	2b	–	–	–	–	1,794	482	57	21	7	2,361
MRMB	2a	–	–	–	–	–	–	370	106	30	506
ABI	2c_gen	–	–	–	–	–	–	550	226	731	1,507
ABIMB	2c_voc	–	–	–	–	–	–	–	239	145	384
FH	3a	–	–	–	–	–	–	–	–	36	36
UNI	3b	–	–	–	–	–	–	–	–	–	–
Total		–	1,887	1,559	2,480	1,982	1,873	1,063	605	961	12,410

Note. Calculations based on 12,410 CASMIN transitions of men coded in the *Education* file of the Scientific Use File of NEPS Starting Cohort 6 (Data: NEPS SC6 1.0.0, own calculations).

Table A2 Frequencies of CASMIN Transitions (Women)

CASMIN	Transition... from	to									Total
		1a	1b	1c	2b	2a	2c_gen	2c_voc	3a	3b	
No degree	1a	-	1,674	98	2,597	3	1,457	11	1	3	5,844
HOB	1b	-	-	900	388	16	93	7	3	1	1,408
HMB	1c	-	-	-	-	52	-	16	3	4	75
MROB	2b	-	-	-	-	2,098	497	33	26	16	2,670
MRMB	2a	-	-	-	-	-	-	131	71	20	222
ABI	2c_gen	-	-	-	-	-	-	800	216	681	1,697
ABIMB	2c_voc	-	-	-	-	-	-	-	108	125	233
FH	3a	-	-	-	-	-	-	-	-	19	19
UNI	3b	-	-	-	-	-	-	-	-	-	-
Total		-	1,674	998	2,985	2,169	2,047	998	428	869	12,168

Note. Calculations based on 12,168 CASMIN transitions of women coded in the *Education* file of the Scientific Use File of NEPS Starting Cohort 6 (Data: NEPS SC6 1.0.0, own calculations).

About the authors

M. Munz
STR Coding, Nuremberg, Germany.

J. Skopek
European University Institute (EUI), Florence, Italy.
e-mail: jan.skopek@eui.eu

Disclosing the National Educational Panel Study

Tobias Koberg

Abstract

The National Educational Panel Study surveys a vast amount of information about individuals as well as organizational units, such as schools and universities. In addition to the voluminous questionnaire, the panel structure significantly increases data size with every new wave. Because of this data abundance, the de-anonymization and (re-)identification of singular units seem to be a major problem when disseminating data to third parties. While de-anonymizing institutions per se is undesired and goes against scientific ethics, it also aids in the task of tracking individual persons. Such an event, however, is believed to be a significant problem in terms of data privacy, federal law, and respondents' agreement. This obviously concerns national statistics institutes and federal data centers, which hold vast amounts of data, often without any firsthand agreement of the affected citizens. For this reason, researchers have rather sparsely reflected on motives of possible attackers and instead focused on anonymization (and even perturbative methods) and access technologies. This may not hold for the social sciences; however, instead of remaining astounded by the anxiety fed by a hardliner's understanding of data privacy, which leads to irrational procedures of data anonymization and seriously limits and harms scientific research, one should ask: Who would be interested in such an attempt? Which realistic options really exist for such an attacker? And what benefit may be gained? By discussing these trivial but difficult-to-answer questions, it is possible to find a realistic and moderate way to secure individual content while maintaining a solid scientific database. This article tries to shed light on which realistic disclosure risks the NEPS has to deal with and which are merely theoretical constructs. After a brief summary of definitions and common statements, possible assault scenarios are discussed and benchmarked. This process explicitly involves the comparison of expected gain with necessary expenses. The main part of this work explores the framework established and the NEPS' Sci-

entific Use Files (SUF) with regard to feasible re-identification and potential profit. The text concludes by presenting the developed anonymization methods applied.

1 Introduction

The National Educational Panel Study (NEPS) aims to provide data on more than 60,000 individuals in six different cohorts ranging from infants to adults. These data are intended to be disseminated as micro data, that is, they are not only to be publicized as aggregated reports or tables, but also to be accessible to researchers as data files containing one record per respondent. A vast amount of detailed information is collected in this process of surveying, resulting in up to 2,000 individual attributes of each real person, household, or institution.

Of course, the collection, archiving, and dissemination of personal data always have to satisfy laws and regulations to secure data in the best possible way. As the German Data Protection Act (*Bundesdatenschutzgesetz vom 20.12.1990*, BDSG) states in § 3a:¹

Data reduction and data economy:

Personal data are to be collected, processed and used, and processing systems are to be designed in accordance with the aim of collecting, processing and using as little personal data as possible. In particular, personal data are to be aliased or rendered anonymous as far as possible and the effort involved is reasonable in relation to the desired level of protection.

However, although the NEPS strictly complies with this setting, this is not the major impulse for an elaborate data security process. As the NEPS was designed as a longitudinal study, it is crucial that individuals (regardless of their age) do not drop out prematurely. One main cause of participation break-off is a loss of confidence in the study process and data usage. Therefore, enforcing trust in the honesty of the NEPS by taking data protection and privacy issues seriously is a key aspect in our data-processing system.

In addition to guarding individual information, prohibiting identification of certain institutions has long been known to be indispensable. While knowledge about the participation of specific schools or universities is a data security breach in its own right, the aim also enhances the de-anonymization of individual attendants. This is in no way acceptable.

Derived from these thoughts, a substantial amount of resources are essentially invested in data protection issues. This article shows all considerations and realizations

1 English translation retrieved from http://www.gesetze-im-internet.de/englisch_bdsge/englisch_bdsge.html (cited 29.07.2015).

of data protection mechanism worked out by the NEPS. Besides own considerations and beliefs, the process has also been orientated towards approaches from other data-producing facilities.

To define the proceeding and systems setup, some light is first shed on different statements and definitions regarding data privacy and the integrity of anonymity. Subsequently, conventional assault scenarios are observed that consider hazards for individual anonymity when obtaining NEPS data. Focus is placed on assault attempts made to disclose private data and the specification of prerequisites necessary for this task to succeed.

Afterwards, the NEPS' multidimensional security structure, a portfolio approach which rigorously tightens the net of data protection, is introduced and discussed.

The following section concludes with further thoughts and illustrates the current anonymization methods applied. Their aim is to prevent full data usability by also maintaining strict data protection and individual privacy needs.

Final remarks are given as a summary of results and an outlook on further NEPS data dissemination projects, which may need additional and new security measures.

2 Framework

To establish our own data-protection issues and disclosure-control techniques, known considerations and methods in the field of data security and anonymization were consulted. The evolved framework combines the most useful and efficient measures to protect sensible data. But what is sensible? Derived from several sources (see, for example, Hundepool et al., 2012; Skinner & Elliot, 2002), data (particularly the information behind them) can be classified in one of the following four categories:

Primary identifiers, which is information that makes it immediately possible to identify individuals or institutions. This might be names and addresses or phone numbers and email addresses. For the remainder of this article, it is fundamental to keep in mind that all primary identifiers are separated and secured at the data-collecting institutes contracted by the NEPS and never leave their custody. Those institutes replace all primary identifiers with an identification number (ID) before sending data to the NEPS data center. This is generally referred to as *pseudonymization*. During the successive process of data editing, this ID is also substituted with another, rendering impossible a direct merge from data to primary identifiers. It is important to note at this point that all further actions regarding data protection are conducted on data material *without* primary identifiers.

Quasi identifiers denotes information that may be used to track and identify individuals or equivalents, even when no primary identifier is at hand. This may be, for example, a combination of date of birth, place of residence, and sex. As realized later

on, with enough quasi identifiers available, the grade of detail of a single identifier is of no importance. As a result, date of birth remains a quasi identifier in the above example, regardless of whether it is stored as a full date (month/day/year) or only as an embracing decade.

Sensitive data refers to attributes of a more private or delicate character (e.g., medical history, income, or sexual orientation), which may also be seen as the earning one gains when a de-anonymization succeeds. Obviously, sensitive data are no quasi identifier as quasi identifiers must be commonly known (to allow linkage), but sensitive data has been secured to be not commonly known.

Other information not classifiable in one of the above definitions neither assists the task of de-anonymization attempts nor contains information that would make it worthwhile. For the challenge of anonymization, information in this category can be safely neglected.

It must be noted, though, that for surveys in social sciences (and therefore in the NEPS), all collected data material must initially be marked as *personal data*, a term that often is stressed, especially in legal expressions. A straightforward interpretation of personal data means the total of all information pinned to an individual. However, in the context of statistical disclosure, this is usually synonymously used to describe the first three of the above categories.

To protect individuals, the term of *de facto anonymity* is defined by BDSG² at § 3(6):

“Rendering anonymous” means the modification of personal data so that the information concerning personal or material circumstances can no longer or only with a disproportionate amount of time, expense and labour be attributed to an identified or identifiable individual.

This implies the *privilege of sciences*, which states that for scientific purposes, datasets may be anonymized such that de-anonymization must not remain completely impossible, but no reasonable equilibrium between costs and effort remains.

To attain this goal, several anonymization measures are available—statistical modifications or alterations of the existing data set. Modification may be achieved through non-perturbative methods like aggregation of specific codes (top-/bottom-coding), variable or cell suppression, substitution of specific values (e.g., micro aggregation), and so forth. Perturbative methods modify data by adding noise to disseminated information, thereby obfuscating precise values to complicate direct matching. This

2 Retrieved from http://www.gesetze-im-internet.de/englisch_bdsd/englisch_bdsd.html (cited 29.07.2015).

may be applied before releasing sensitive numerical variables such as income. To maintain a high scientific usability of data material, the NEPS does not use perturbative methods for anonymization. Please note that data security methods like restricted access or contractual commitment are not anonymization, but rather, protection methods.

The NEPS' cardinal purpose is to assemble and prepare data for dissemination to the scientific community. As the NEPS does not primarily collect data for its own scientific projects and research, the targeted data volume and scope is not precisely defined. Because of this, the grade of importance of certain data cannot be thoroughly determined. As every mutilation of data may harm or even prevent specific scientific analysis, perils lie in the execution of anonymization. It may seem easy to remove specific information from data files in order to avoid a discussion of disclosure risks, but this more often might seriously inflict damage than safety. That said, there is of course no excuse for sloppy data protection. If information permits individual re-identification, it must be efficiently secured.

To achieve these contrary aims equally, focus should be set on evaluating sensitive information *prior* to data modification and to discussing under which circumstances information becomes sensitive.

2.1 Assault Scenarios

In the common literature, two cardinal assault scenarios are realized: *Single attack* (“*Einzelangriff*”) and *mass catch* (“*Massenfischzug*”) (see, for example, Müller, Blien, & Knoche, 1991). The first, a single attack, denotes the possibility of identifying single individual units by using some external, selective data knowledge. For example, it may be possible for a close friend or a neighbor of the designated person to use his or her very private knowledge to detect the target in a data file. An abstraction of this idea applies to celebrities, for whom everyone may be seen as a neighbor in the above sense. How inevitable this task actually is becomes clear when overlapping neighbor and target. An individual may always locate him- or herself in a dataset, especially in such an extensive study as the NEPS, where one's life is meticulously tracked retrospectively and through subsequent panel waves. The following example illustrates this.

Example: K-Anonymity of Dichotomized Basics File

Investigation is done using only the *Basics* Data File of our Starting Cohort 6 Scientific Use File (SUF), version 1-0-0 (Download).³ The dataset comprises 67 variables

3 This paper uses data from the National Educational Panel Study (NEPS): Starting Cohort 6—Adults, doi:10.5157/NEPS:SC6:1.0.0. From 2008 to 2013, NEPS data were collected as part of the Framework Programme for the Promotion of Empirical Educational Research funded by the German Fed-

from 12 different subjects. After removing technical variables (describing count and duration of spell episodes) and information that is mostly redundant (e. g., different occupational statuses, like ISEI and EGP), the remaining 38 variables were dichotomized (e. g., high/low, old/young, poor/rich etc.), which means that variable content was reduced to one of three attributes: 0, 1, or missing value (NA). Altogether, this is an extremely severe alteration of data material which no doubt precludes an enormous amount of research topics. But even with this degenerated anonymization method, there are $K = 3^{38}$ different combinations of values, also called attribute sets (see Elliot, Manning, & Ford, 2002). As the Basics File only holds the most recent information for every respondent, it contains one single line per respondent, a total of $n = 11,649$ rows. That said, plain permutational probability for at least two respondents holding the same attribute set can be calculated as a zero close $p = 5 \cdot 10^{-11}$. The empirical distribution of this finding is by no means more promising. K-anonymity (Sweeney, 2002), which describes the minimal number of cases in a data file sharing the same attribute set, was computed. It turns out that 10,743 (92.2 %) individuals hold a unique combination. This means that if enough knowledge about an individual has been accumulated, even at a very general level (i. e., only a basic clustering in one of two categories is necessary), an individual re-identification is not preventable.

Result

This gives rise to a belief that any complete and flawless anonymization also corrodes the material for analytical purposes in such a way that reasonable scientific research is no longer possible. In fact, creating a thoroughly anonymized data file would simply result in no data file at all. This would most certainly contradict the NEPS' basic concept, which was established exactly for the purpose of disseminating data to researching scientists. This circumstance has even been acknowledged by Germany's Federal Commissioner of Data Production (Schaar, 2010), who mentions that "it goes without saying that the traditional method of rendering data anonymous and deleting individual statistics based on a type of *stage model* is not compatible with a method that links microdata" (p. 637).

This does *not* mean, however, that no anonymization procedures protecting individual units were conducted while disseminating the NEPS' Scientific Use Files. As it happens, quite the contrary is true. Despite the fact that all information collected was done so with full agreement and knowledge of the data owner, an amazing comparability is possible with techniques used in official governmental register surveys. The following two have been selected for illustration.

Table 1 HIPAA Ad. Simplification

A	Names	J	Account numbers
B	Geographic subdivisions	K	Certificate/license numbers
C	Dates related to individual (except year)	L	Vehicle identifiers and serial numbers
D	Telephone numbers	M	Device identifiers and serial numbers
E	Fax numbers	N	Web universal resource locators (URLs)
F	Electronic mail addresses	O	Internet protocol (IP) address numbers
G	Social security numbers	P	Biometric identifiers
H	Medical record numbers	Q	Photographic images
I	Health plan beneficiary numbers	R	Any other unique identifying number

When searching for official instructions and setpoints, one may find the often-cited⁴ United States Health Insurance Portability and Accountability Act of 1996 (HIPAA⁵), which reveals that “(health) information [is] not individually identifiable” in § 164.514 (b) (2i) if 18 identifiers are removed (see Table 1).

These HIPAA regulations were established to secure the medical data of patients for scientific purposes. As medical data are extremely sensitive information and very private property, they are by no means comparable with the NEPS’ surveyed data.

Notably, the NEPS does not provide information on any of these identifiers except for B and C, which hold slightly modified information.⁶

Panning over to more national regulations, let us take a look at the anonymization routine of the 1987 Census of the Federal Republic of Germany and its Public Use File (PUF, see Crößmann, 2009). The NEPS does not try to compete with procedures applied there for four main reasons:

- Census data are generated from a governmental register, collected without agreement from individuals. The NEPS survey is, in its basic property, a voluntary affair.
- Besides the given voluntariness, there is another major factor that satisfies classification of anonymous data: the availability of data only for a sample of the total population.

4 See, for example, Kushida et al. (2012).

5 Online at <http://www.hhs.gov/ocr/privacy/> (cited 23.12.2014).

6 Geographic localization (B) is possible up to the first five digits of AGS (“Amtlicher Gemeindegliederung”); dates (C) are provided by month (which is actually very necessary when working with spell data).

- The procedures describe anonymization to create a PUF (available to everyone). The NEPS disseminates Scientific Use Files, which are only accessible to researchers.
- Data in census PUF are defined as absolutely anonymous, that is, re-identification of individuals is completely impossible. The NEPS strives for *de facto anonymity* (see above).

Despite these huge differences, the NEPS routines for data protection are actually not very far off from the ones applied in the above-mentioned census, which are listed in the following overview:

Age of data: Census PUF were released more than 20 years after realization. This is, exceptionally, an anonymization measure that the NEPS cannot compete.

Sampling: The PUF only contains a five percent sample of the complete material. The NEPS actually *is* a sample survey.

Regional information: Only the federal state is available. The NEPS releases more detailed regional attribution but recognizes regional information as protectable.

Sorting: Data material was used to prevent the identification of regionally associated cases. The NEPS operates identically as our data files are sorted by a system-free identification number.

Coarsening: Values are aggregated so that each cell comprises at least 10,000 cases. The NEPS does not cover a strategy of minimal cell population. However, this aggregation of variables is executed to protect singularities in the univariate distribution of attributes only. Official authorities silently accept the fact that trying to aggregate multidimensional combinations of attributes is far from the reasonable limits. When aggregation becomes necessary, the NEPS also only regards one dimension.

Mass catch

The second assault scenario, denoted as *mass catch*, differs from the first as it does need an actual external database (of any kind whatsoever), that is, a secondary source of information which holds data for many individuals. An attacker may combine these datasets by adapting attributes available in both sources to create an enhanced and all-encompassing data file in which identification (and especially the mapping of primary data to known individuals) is extensively possible. The attacker then uses the created data file to tag individuals most valuable for his or her undertaking and to exploit them. This is a very hypothetical consideration and does not take into account the state of the following topics (especially for surveys in the social sciences):

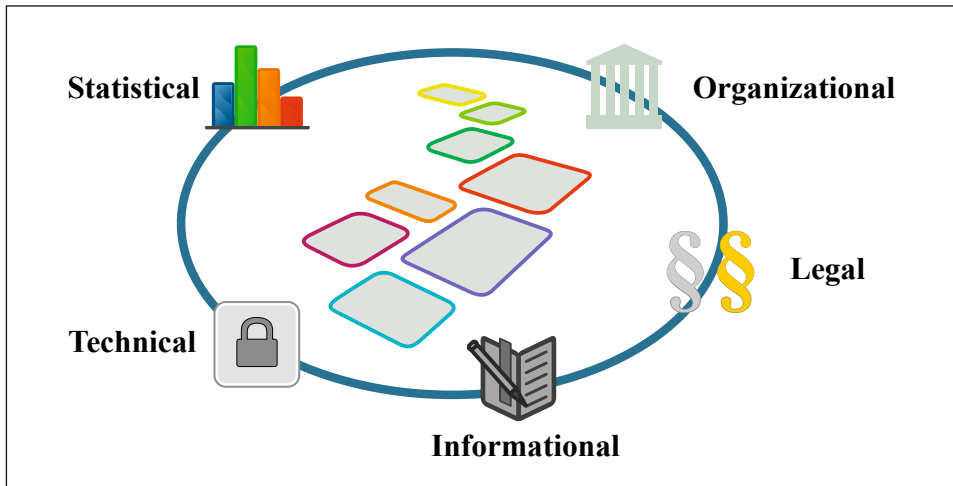
- NEPS data are only disseminated to researchers for scientific purposes. It is incomprehensible why members of the scientific community should have any desire for malicious tasks, risking their careers and reputations for pointless activities.
- The existence of such a database is highly disputable. It would have to contain primary identifiers (like names, addresses, phone numbers, etc.) while simultaneously holding attributes which may be used for merging. This database also must be accessible publicly, or at least by an attacker of dubious allegiance. Furthermore, this database must lack information included in our primary data source; otherwise, an (aspired) merge would be obsolete.
- The usability of information gained by a successful attack is by no means obvious. As previously stated, the NEPS only surveys voluntary data, primarily from data holders themselves. Thus, its value for a criminal action is basically void.
- Keeping the above two points in mind, the *usage vs. costs* criterion as given by definition of *de facto anonymity* is not proportionate.
- The origin of information on being a human being (likely reporting about him- or herself) entails another pitfall: Collected information may be imprecise because of errors in measurement. A linkage of two data sources, by incidental quasi identifiers rather than by consistent IDs, however, requires an accurate equality of variables. Thus, if not excessively considered and quantified, measurement error destroys direct concordance and hence prevents de-anonymization.

Reflecting on these ideas, it is immediately obvious that a mass catch is hardly operable as well as pointless. Its absent practical feasibility makes it a harmless risk. Prevention against it, however, would crucially hamper scientific research.

2.2 Portfolio Approach

So far, anonymizing data material seems to be a very raw and inefficient method of implementing data protection. On the other hand, not securing personal information at all is not reasonable and would very soon result in the project's failure. Other ways to secure data have been developed and established (for example, remote execution systems), but all of them feature some deficits, which makes it very hard to solely rely on them. Fortunately, exclusive usage of methods is by no way obligatory. Our system, adapted from ideas of Julia Lane (Lane, Heus, & Mulcahy, 2008), consists of the consideration of merging multiple, completely diverse approaches instead of abusing one single mechanism to its outage. We call it our *portfolio approach* (see also Figure 1). As a matter of fact, the following five modules achieve a sophisticated degree of data security and disclosure control in their combination by maintaining an outstanding availability of personal yet sensible information.

Figure 1 The five approaches of the high-level multidimensional data-protection system



Organizational approach

The organizational approach limits the possible user community to a trustworthy and respectable base. Only scientists associated with a scientific institute and with a justified research interest are granted access to the NEPS SUF. Usage is not intended to be available for the general public.

Prior to data access, a data use agreement has to be signed to guarantee the direct and personal responsibility, integrity, and respectability of the data recipient. In this agreement, besides the obligatory notion of name, contact information, and (connection to) institution, a researcher has to accurately describe his research project, including its purpose and duration. Additional data users can also be listed but must emerge from the same institution. In fact, this is the only way for students who are not employed at a scientific institute to obtain SUFs—to be listed as an additional user in a data use agreement duly signed by a fully fledged member of the scientific community.

These limitations affect confidence and acceptance shown by respondents in a positive way because debacles such as public release or commercial use of our scientific data are precluded.

Legal approach

The data use agreement comprises additional legal limitations to which researchers have to conform. The signer assures a scientific purpose, limited usage duration, and commitment to comply with legal stipulations regarding the data protection law. Moreover, transmission of data to third parties and the intentional re-identification of individuals or institutions, as well as of other misconduct, are strictly prohibited.

A breach of those regulations leads to a termination of the data use agreement, harsh fines, exclusion from further NEPS data usage, as well as to the dissemination of the researcher's name amongst the scientific community and other research data centers (a modern kind of proscription). Therefore, when violating terms of the data use agreement, one risks not only high fines but also severe consequences for his or her reputation.

Informational approach

Another major ambition of the NEPS is to provide as much documentation and user training as possible. This also includes detailed information about data protection and anonymization necessities, thereby sensibilizing users to adhere to these issues. In emphasizing data security strongly, the NEPS believes in the trustworthiness of its users (referred to as *safe settings*, *safe people*). Beyond this, the full spectrum of (meta-) information and documentation minimizes the risk of (possibly unintentional) unauthorized data usage as users do not have to stumble through our data material.

Technical approach

NEPS data are disseminated to the user by three different access modes: *OnSite*, *RemoteNEPS*, and *Download*. The first two approaches do not physically deliver the data to users; rather, the data stay in the NEPS' protection system. Because of this, it is possible to offer more sensible data in these modes as this system is highly secured and supervised, for example, import and export are only accomplished through NEPS staff.

Statistical approach

Regardless of former security measures, there might still be some need to modify data material to maintain individuals' anonymity. In addition to assuring anonymity of respondents, context persons, and institutions, the NEPS also tries to preserve the data basis in its entirety to not restrain scientific analysis. Nonetheless, the risk of re-identification is minimized by modifications like top coding, aggregation, and suppression. The remainder of this article examines the methods conducted in the statistical approach.

3 Analyzing the NEPS

To elaborate on the methods for generating anonymized and safe data material, one first has to screen the subject matter, that is, the given data material. To do so, an attempt is made to classify already-collected as well as expected future data (i. e., SUFs), by means of data structure, enclosed sensible information, and exploitability. Then, the underlying anonymization concept and key aspects of SUF-specific techniques are clarified.

Because the first release of Starting Cohort 6 (SC6, Adults) preceded all other NEPS surveys by almost a full year, the invention of anonymization procedures also began for this specific cohort. Furthermore, as all SUFs contain data collected from adults (as both primary targets and context persons), it seems a good idea to begin with anonymization techniques for this population as these considerations affect all surveys. Since one case only is made up by one single individual person, data structure is relatively simple. The major complexity lies in the availability of biography spell data, which is tremendously individual but almost not anonymizable without completely destroying any scientific usability. However, using spell data as quasi identifiers would require such a broad background knowledge about certain individuals that such an exploit need not be considered seriously. Moreover, the wisdom gained from spell data is meager. Besides this vita history, there is only little information surveyed worth exploiting and therefore required to be protected. The key aspect for the anonymization of adults is detected as a way to prevent the spotting of individuals in the data by using unique features in the whole population. Therefore, attributes are modified that only apply to a small subpopulation (e. g., very rare mother tongue) or that may be known by a wide group of people around the individual (e. g., number of employees). Additionally, regional localization is blurred to an extent in which cells cover a large enough number of residents.

In Starting Cohorts 2, 3, and 4 (SC2, SC3, SC4), not only students, but also parents, teachers, and headmasters were surveyed. This increases complexity, and additional knowledge about individuals is given, although information from these studies is intentionally only collected as context to students. As the main focus of the NEPS lies on educational systems, sensible information consists mainly of educational outcomes: grades, competency tests, and so forth. The only imaginable exploit of this data material seems to be at a very individual level, for example, when one seeks information about relatives or neighbors. This is very selective. However, another re-identification purpose seems more reasonable: Revealing identities of institutions because a comparison of different schools may in fact be a plausible scenario. Thus, the emphasis of anonymization methods was set on the protection of information that describes the setting of and in the institution.

Starting Cohort 5 (SC5, First-Year Students) is, at least in terms of anonymization considerations, in some way a hybrid of the former two. Because the focus of this survey lies on the individual development of respondents, the same measures as used in SC6 should be applied. Nevertheless, information was collected that may disclose the attending university, which reproduces the above-mentioned scenario of institutional comparison.

Starting Cohort 1 (SC1, Early Childhood) was sampled in an individual context (similar to SC6). Therefore, the same anonymization measures as conducted to SC6 seem reasonable. These have to be expanded by methods adopted in SC2 to SC5 when an institutional context (e. g., nurseries) becomes visible.

3.1 Specifications⁷

To ensure the best possible confidentiality protection of individuals and individual micro data, the NEPS complies with strict international standards. In order to operationalize these standards, they are abstracted to the following two criteria:

- Disseminated data is transferred to so-called *de facto anonymous data*. Identifiable information is coarsened or cut off and kept securely to minimize risk of statistical disclosure.
- The use of data is strictly confidential and for statistical purposes only. The closed contract only grants access to members of the scientific community. This contract has a vast amount of legal stipulations, one of them being a large fine that applies to the realization of intentional re-identification. Therefore, disseminated data are highly protected by law, which allows for a more flexible range of available data.

Concerning the latter point, regarding legal regulations, the NEPS has made a huge effort to offer as much analysis power of data as possible. This *paradigm of information esteem* reveals the fact that there are few conducted measures of statistical disclosure control. Moreover, if there really was a need for modification, only non-perturbative methods were used.

3.2 Onion-Shaped Model

The NEPS grants users three different modes of data access: *OnSite*, which stands for the opportunity to use the secured infrastructure made available at the NEPS in Bamberg, *RemoteNEPS*, which is a progressive remote access technology that provides a virtual desktop, and finally, *Download*, which indicates the possibility of collecting data via a secure web portal. These given access modes were created to allow for anonymization routines for a subtle differentiation of information. The three resulting levels of anonymization are as follows:

- Data provided *OnSite* are generally not anonymized further. However, even these data are rendered *de facto anonymous* for no disclosure risk to persist. All information contained remains completely sane. Although users have to deal with limited possibilities of data access (i. e., supervised import and export of their results), they are free to work with all data available at the NEPS in a secure environment.
- Access via *RemoteNEPS* is considered equivalent to *OnSite* hence, most of the data remains complete.

⁷ The rest of this section was originally published as documentation of our Scientific Use Files (Koberg, 2011).

- As *Download* is assumed to be the most hazardous access mode,⁸ some additional anonymization techniques are applied to the dataset.

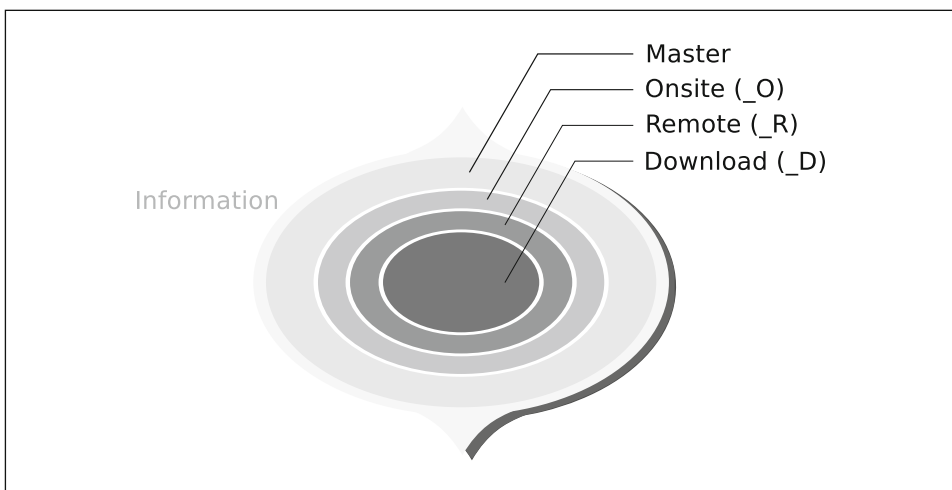
Obviously, this approach results in three different versions of all involved datasets. To enable a consistent structure, these data files always contain the entire set of variables; it is their content that differs through the three levels.

As there is normally no need to resign aggregated variables in higher levels (i. e., OnSite or RemoteNEPS), these variables are defined as a surplus to original variables in the OnSite-version. Stepping down to RemoteNEPS, the content of related variables that are too sensitive for this level is overwritten with an exclusive missing code—an operation that we define as *purging*. Note that system missing values are not affected, thereby allowing users to differ between value existence and nonexistence. This still is a valuable piece of additional information. The same applies to Download.

While there is no explicit documentation of this fact, it should remain clear that this procedure leads to accumulation, that is, purged content under RemoteNEPS is therefore neither included in RemoteNEPS nor in Download.

This onion-shaped model provides both an ease of (1) the use of different sensitivity models (e. g., preparing an analysis using the Download dataset and conducting it afterwards using the OnSite-data) and of (2) documentation since the subject of documentation is the most sensitive level (OnSite), with RemoteNEPS and Download levels being a subset of these data.

Figure 2 Onion-shaped model defining different anonymization levels



⁸ “Hazardous” in terms of downloaded content is no longer under physical control of the NEPS.

The fourth layer, *Master*, which is depicted below, contains every material that is needed during data processing by the NEPS, but it is not meant to be usable for the scientific community.

Technically, this model takes the form of a single letter suffixed to dataset and variable names. All datasets available OnSite only are marked with an additional *_O*, those available via RemoteNEPS with *_R*, and Download files with *_D*. The same procedure applies when it comes to variable differentiation. A variable that is only available OnSite is suffixed with *_O*. In RemoteNEPS-access or Download, this variable is still present but purged. If there is an alternate version (mainly with coarsened content) for RemoteNEPS (suffix *_R*) or Download (suffix *_D*), these can be used. As previously stated, these are already integrated in the OnSite version.

3.3 Conducted Measures

Keeping usability and the paradigm of information esteem in mind, only very few alterations are actually done to the dataset. These modifications always account for the fact that information may never be lost completely, but aggregated into coarse categories or variables. Please note that all information is still available somewhere and that only RemoteNEPS and (mainly) the Download version are constrained in this matter. In fact, usually merely about less than 10 % of the whole dataset volume is modified.

Table 2 provides an overview of the conducted measures. For a complete list of all variables modified by anonymization procedures, a look at the corresponding SUF's data manual supplement for anonymization techniques should be sufficient. An explanatory overview is given below.

Table 2 Availability of Sensitive Data

Topic	OnSite	RemoteNEPS	Download
International <i>(e. g., nation states, national languages)</i>	Full data	Full data	Collapsed
String variables	Anonymized	n/a	n/a
Institutional	Full data	Full data	n/a
Regional geographical information	NUTS-3	NUTS-3	NUTS-1
Number of employees	Full data	Full data	Top coded
Macro indicators	Accessible	n/a	n/a

Countries and languages

All information corresponding to (international) localization, nationality, and languages is only available in full OnSite or via RemoteNEPS. Variables comprised in the Download SUF are aggregated into larger categories.

Open-ended strings

All string variables containing actual text are purged in the RemoteNEPS version. Information remains accessible OnSite. However, all text entries are reviewed by staff to ensure that absolutely no re-identificational material is included.

Institutions

For Starting Cohorts 2 to 5, a special focus of anonymization is directed to protect institutional data, that is, information about Kindergarten and schools as well as about educators and teachers. This includes the complete data file *xInstitution* as well as basic structural details about Kindergarten groups and school classes. Furthermore, personal information about educators and teachers is treated more securely. Detailed information about these subjects can be found from RemoteNEPS onwards.

Regional information

In SC6 and SC1, regional information below the federal districts in Germany is only available via RemoteNEPS (i. e., NUTS-3, Download version contains NUTS-1⁹). This regards places of birth as well as work, school, and residence. Where data has been surveyed in a school context (SC2 to SC5), protection of the institution requires more conservative rules. Therefore, the Download version only comprises an indicator for West Germany and East Germany (including Berlin). Federal Districts can be found from RemoteNEPS onwards. Besides these regional keys, additional macro indicators are available OnSite (see below, e. g., Microm data).

Number of employees

Considering self-employed persons, information about the number of salaried employees is censored to prevent the easy identification of large entrepreneurs. Therefore, related variables are topcoded at 20 employees. Again, this information is still available via RemoteNEPS and OnSite.

Macro indicators

Additional information including structural topography and macroeconomic measures has been made available only OnSite. In SC2, SC3, and SC4, Microm and infas Geodata are available for the residence as well as for the location of the institution. In SC1, SC5, and SC6, these datasets are only available for the respondent's residence.

9 NUTS-1: Federal States (“Bundesländer”); NUTS-3: administrative districts (“Kreise”).

4 Conclusion and Outlook

When disseminating personal data to third parties, preserving the anonymity of individuals is an absolute necessity. However, as anonymization measures have mainly been developed on behalf of federal statistical agencies in their effort to make census data accessible, their proceeding is very restrictive. In social sciences, data material is not as delicate and assault scenarios are not as realistic as described above. The NEPS' data protection setup disarms sceptics whose only weapon of (re-)identification prevention is data anonymization. The setup shows that data protection and the respect of personal information does not have to be contrary to scientific usability. The capacious portfolio approach combines different data security measures so that actual data modification may be somehow disregarded. Additionally, by providing three different access modes, the NEPS has the possibility to make more sensible content available under surveillance.

As the NEPS' popularity increases, the need to disseminate collected data to further parties arises. In addition to offering SUFs to full-fledged researchers, boosting the acceptance of our data at universities by providing campus files would exceptionally enrich our program. Of course, because many of the security approaches from the portfolio approach cannot be kept up in this context, more current data modifications have to be executed. This would result in the establishment of certain (absolute) anonymity, by which the perils of individual disclosure would dissolve.

References

- Blossfeld, H.-P., Roßbach, H.-G., & von Maurice, J. (Eds.). (2011). Education as a life-long process: The German National Educational Panel Study (NEPS) [Special issue]. *Zeitschrift für Erziehungswissenschaft*, 14. Wiesbaden: VS Verlag für Sozialwissenschaften.
- Crößmann, A. (2009). *Konzept zur Anonymisierung der Volkszählung der Bundesrepublik Deutschland im Jahre 1987 zur Verwendung als Public-Use-File (PUF)*. Retrieved from http://www.forschungsdatenzentrum.de/bestand/volkszaehlung/puf/1987/fdz_vz_1987_puf_anonymisierungskonzept.pdf
- Elliot, M. J., Manning, A. M., & Ford, R. W. (2002). A computational algorithm for handling the special uniques problem. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10, 493–510. doi:10.1142/S0218488502001600
- Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E. S., Spicer, K., & de Wolf, P.-P. (2012). *Statistical disclosure control*. Hoboken, NJ: Wiley.
- Koberg, T. (2011). *Starting Cohort 6: Adults (SC6), SUF-Version 1.0.0, Data Manual [Supplement E], Anonymization Procedures* (Technical Report). Bamberg: University of Bamberg, National Educational Panel Study.

- Kushida, C. A., Nichols, D. A., Jadrnicek, R., Miller, R., Walsh, J. K., & Griffin, K. (2012). Strategies for de-identification and anonymization of electronic health record data for use in multicenter research studies. *Med Care, 50*, 82–101. doi: 10.1097/MLR.0b013e3182585355
- Lane, J., Heus, P., & Mulcahy, T. (2008). Data access in a cyber world: Making use of cyberinfrastructure. *Transactions on Data Privacy, 1*(1), 2–16.
- Müller, W., Blien, U., & Knoche, P. (1991). Die faktische Anonymität von Mikrodaten (Schriftenreihe Forum der Bundesstatistik, 19). Wiesbaden: Statistisches Bundesamt.
- Schaar, P. (2010). Data protection and statistics—A dynamic and tension filled relationship. In German DataForum (RatSWD) (Ed.), *Building on progress, expanding the research infrastructure for the social, economic and behavioural sciences* (Vol. 2, pp. 629–642). Opladen: Budrich UniPress.
- Skinner, C. J., & Elliot, M. J. (2002). A measure of disclosure risk for microdata. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 64*(4), 855–867.
- Sweeney, L. (2002). K-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10*(5), 557–570.

About the author

T. Koberg
Leibniz Institute for Educational Trajectories (LifBi),
Bamberg, Germany.
e-mail: tobias.koberg@lifbi.de

String Coding in a Generic Framework

Manuel Munz, Knut Wenzig and Daniel Bela

Abstract

For many questions in the social sciences that are supposed to be answered with survey data, reliable and detailed information about occupations is crucial. As classifications for occupations are very extensive and complex, it is not feasible to simply present a full scheme to the respondent. To overcome this issue, an open string is queried from the respondent and later converted to an appropriate entry in a chosen classification. This task can be handled using a generic coding framework, which is illustrated in this article. The raw material with the strings-to-code (reported occupations) and covariate information has to be prepared and delivered to the process itself. The selected coding scheme has to meet several requirements, such as discriminatory power, completeness, and adequacy. The NEPS's coding framework can be adapted to a larger set of variables: The interface for exporting content-to-code from the NEPS dataset files is used beyond the coding of occupational information. Every NEPS survey developer who is urged to classify his or her string variable(s) is provided with spreadsheets ready for the related workflow. When finished, the NEPS Data Center re-imports these spreadsheets into the dataset. Several further mechanisms have been integrated into this process to ensure high data quality.

1 Motivation for Coding String Information

Questions in which the respondent can state the answer in an open format, which is referred to as surveyed string information, can be located at several places within a survey instrument. Sometimes, a researcher cannot provide entire and closed categories in the answering scheme because he or she may not know all the possible entries of this desired list. Even if the researcher were aware of the completeness of such a

list, the list would be too long to be applied sensibly in the framework of the survey. It would take too much time to read out a long list during a Computer Assisted Telephone Interview (CATI)—and the respondent could get annoyed and, in the worst case, refuse to participate further. For example, no one can determine all the possible jobs or occupations that all respondents may state. It is furthermore possible for the respondents to misunderstand the questions, and they could state anything different, such as school types, branches, or industries, as well as even more deviant entries. Even if the instrument developer is able to construct a full list of all occupations and other possible entries, this list will very soon grow to be enormous, rendering it impracticable in the survey.

Therefore, the implementation of open questions is a possible way to avoid these problems about the unknown completeness and the non-manageable character of such a long list. The National Educational Panel Study (NEPS) collects all information in an open text (string) format so that the respondents can basically state anything they want. A practicable solution for dealing with this kind of entry is the coding of the information for the further processing. Generally, coding describes the process of the assignment of a code from a selected category scheme (classification) to the string information. We use examples from occupational coding to practically illustrate the theoretical deliberations described in this article. This is the most complex and elaborate string coding procedure that has been implemented in the NEPS up to now.

Several reasons for the coding of string information at the stage of the data edition have been able to be defined.

1.1 Reduction of complexity

The most obvious reason for the coding of string information at the stage of the data edition is the reduction of the complexity of the available string information. This kind of string information is very heterogeneous, and rules to integrate this information into an analysis cannot be easily defined. It is hard to handle openly stated string information because a researcher does not have full information about all possible entries. He can, for example, define sub-strings to cover and integrate most of the string information in data-edition commands, but there is a leftover in most cases.

It is not possible to define a full set of sub-strings to convert the string information to a manageable recoded version of the given information for several reasons: There could be (1) other strings that also have *auto* in their entry but that do not mention different branches and occupations (e.g., *automation*) or (2) synonyms that are not covered by the defined sub-string (*car* vs. *automobile*). (3) Not all possible spelling errors (*atuo* instead of *auto*) can easily be covered. As a result, it is far more practicable to code all unique strings one after another (cf. Table 1).

We can also use coded string information to derive multiple classifications and scales. Most of the scales and classifications (e.g., ISEI, SIOPS, EGP, cf. Ganzeboom

Table 1 Example on Branches Related to the String "auto"

car company	car assurance
automobiles	car production
car industry	automotive retail
car showrooms	car production (Bentley)
automobile industry	auto industry suppliers
car industry at Ford	automob.
car parts	atuo industry
car rental agency	automotive industry

& Treiman, 1996) are derived from a common classification, such as the International Standard Classification of Occupations 1988 (ISCO-88). We cannot derive the classifications on the raw string information; the strings have to be treated by coding them.

1.2 Avoidance of artifacts in research results

Given that every data user has to encode string data manually, the result should be that coded variables are as heterogeneous as the scientific community is. As data users may have diverse expertise in their specific field, this may result in artifacts in research results if coding is not performed according to common standards. Since the NEPS possesses vast knowledge in the area of coding string information (and transcoding other classifications), centrally performing this task helps to avoid heterogeneity and arbitrariness from the start.

1.3 Data protection

Last but not least, we have to code nearly every piece of string information to ensure compliance with data-protection issues. Coding leads to a coarsening of the string information and makes it nearly impossible to re-identify persons or institutions.

2 Choosing the Right Classification and Coding the Right Way

2.1 Preliminary Ideas Before Coding: Which Classification Should I Choose?

One important decision is the question of which classification is the best to represent the stated string information. There are different requirements that a classification has to fulfill. The most important ones are presented in the following sections.

Exactness and discriminatory power

These criteria describe some key features of good scales. The classification has to replicate the diversity of string information as exactly as possible. This means that there should be as much information transferred to the code as possible. Overall, maximizing the exactness and minimizing the loss of information could be considered initial motivation.

The criterion of discriminatory power is also a main issue for the coding. Each piece of string information should be explicitly assigned to no more than one code of the desired classification. One (1:1) or many (m:1) strings have to be converted to only one possible code in the desired classification (cf. Figure 1).

Figure 1 Optimal string-to-code assignment scheme (1:1 or m:1)

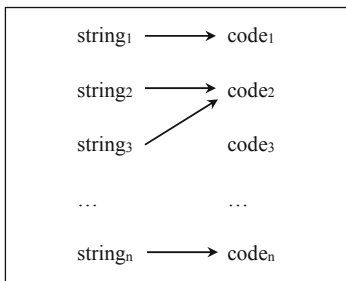
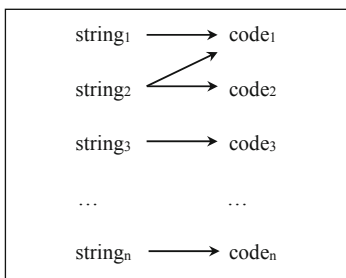


Figure 2 Assignment problems in the string-to-code assignment scheme (1:n)



It is clear that every string is assigned to only one code in the classification. In doing so, unique assignments (string₁ being assigned to code₁) are supposed to be distinguished from so-called “mergers” (string₂ and string₃ being assigned to code₂) and “splits” (one string being assigned to several codes), which are generally not desirable (cf. Ganzeboom & Treiman, 2010: 6).

Problems occur as soon as splits are involved because there is no way to derive the desired codes of the classification properly due to a lack of further information in the string.

The string₂ is coded via code₁ or code₂ of the classification. This assignment problem cannot be directly resolved. However, these hurdles can be overcome either by using multi-assignment schemes that define rules and conventions, by using the most common denominator in the classification, or by the application of auxiliary variables (cf. Figure 2).

High level of differentiation and high level of detail in the classification

We furthermore prefer a classification that features a high level of differentiation and detail. We can preserve more information content of the original string by using a highly detailed classification.

Some classifications have different layers and sections with different levels of exactness. It is important that a high discriminatory power also be transferred into the different layers of the classification. The coder has to be able to switch between the different levels without problems caused by a lack of discriminatory power. An element of a very fine category must not be part of several categories of an overlying level (technically, another split would be produced).

For example, a classification concerning occupations could contain different levels of exactness to code occupations, ranging from Level 1, which provides a very scarce description of occupations, to Level 3, which provides highly descriptive occupational titles.

Switching between the different levels of exactness represented by Level 1 through Level 3 is easy. Elements of a more exact level belong only to one value of an overlying level. For example, nurses do not belong to *Occupations in business administration* or to *Occupations in the financial service* (cf. Table 2).

Completeness of the classification

Another attribute of well-fitting scales is the completeness of the classification. All possible and sensible elements of the information (e. g., occupations, branches, etc.) should be represented in the classification.

Table 2 An example on the Level of Exactness of Classifications

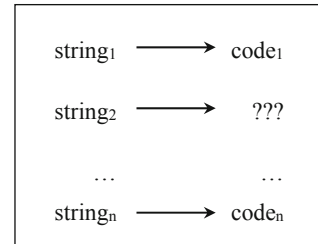
Level 1	Level 2	Level 3
Occupations in the tertiary sector	Occupations in business administration	Bureau clerk
		Wages clerk
		Industrial business management assistant
		Financial advisor
	Occupations in the financial service	Banker
		Tax accountant
	Occupations in the public health sector	Nurse
		Dental technician
		Childcare provider

As can be seen in Figure 2, string_1 is assigned to code_1 , while string_2 has no equivalent in the classification. This information cannot be coded initially. However, codes can be assigned by applying rules and conventions, by trying to find the most common denominator at an overlying level, or by using more information by considering auxiliary variables.

There are several more attributes and requirements of classifications, such as a possibility to transcode into other classifications (compatibility), an international application and knowledge of the classification, and so on. However, the options mentioned thus far should be seen as the most important characteristics.

The next step after choosing the proper classification is determining how to code the string information properly.

Figure 3 An example on the completeness of classifications



2.2 The Role of the Classification's Theoretical Framework

As previously mentioned, the initial aim of coding is to represent as much informational content of the string information as possible (representation of the string, input section). Additionally, the information content of the string information could be increased by surveying further information in so-called *auxiliary variables*.¹ These variables can enhance the accurateness of the string information and therefore lead to an improvement in the representation of the input section. Rules and conventions can also enhance this kind of representation.

The selection of the appropriate classification should depend on the method's discriminatory power and whether it is appropriate to the research objectives.²

During coding, it is critical to know the classification's underlying theoretical framework. These implications should now be illustrated further by transferring these *mechanics of coding* to the process of coding occupational information.

At the NEPS Data Center, we choose to code the information on occupations into the system of the Documentary Code Number (Dokumentationskennziffer/DKZ). As this appends 3 more digits to the Classification of Occupations 2010 (Klassifikation der Berufe 2010/KI dB2010), the classification is more differentiated than the KI dB2010 yet still relies on its theoretical framework.³

1 For example, also surveying the differentiated kind of self-employment increases the assignment accurateness for creating the EGP classes.

2 For example, internationally comparative research should be based on internationally standardized classifications.

3 This helps to derive more classifications (cf. Figure 6).

The developers of the KldB2010 treat occupational information as multi-dimensional information (cf. Paulus, Schweizer & Wiemer, 2010: 7 ff). The two main dimensions of occupational information are skill specialization (horizontal dimension) and skill level (vertical dimension) (cf. Paulus, Schweizer & Wiemer, 2010; Elias & Birch, 1994: 1 ff. (ISCO-88)). Both dimensions span a two-dimensional space with the skill level being the ordinaly scaled vertical dimension and with the skill specialization being the nominally scaled horizontal dimension.

It is only possible to code properly if one knows the values of all dimensions of a given piece of string information. Therefore, it is necessary to be aware of the number and the content of the dimensions that the information needs to be classified into.

Several auxiliary variables can serve as support to help assign the string information properly. There can be an enquiry after the first question about the current occupation in which the target person can specify the main activity in the mentioned occupation, or the differentiated occupational position linked with the mentioned occupation can be used to estimate the skill level more precisely. Other questions, such as the related industry, the size of the enterprise, the number of subordinates, etc., are possibly helpful. Table 3 shows some the sources of information and their contribution to the two dimensions, which have to be mapped.

Here, string information and optional auxiliary variables serve as sufficient indicators to code properly. Nevertheless, rules and conventions have to be defined for complex string information. These rules have to be implemented to ensure uniformity of the coding process when problems with coding or unknown proceedings come up.

Table 3 Determining the Dimensions via Auxiliary Variables

Information	Dimension	Example
String information	Horizontal and vertical	Horizontal: gardener, cooks, etc./vertical: ... helper/vertical & horizontal: medical doctor ¹
Differentiated occupational position	Vertical	Unskilled occupation
Information on industry	Horizontal	Differentiation between handicraft and industry
Number of subordinates	Vertical	Supervisory occupation?

¹ Some occupations can only be practiced by completing a study at a university.

3 The NEPS Coding Framework

The NEPS Data Center developed and implemented a complete integrated framework for the data edition process in the course of the generation of a Scientific Use File (cf. Bela, this volume). One part of this framework is the question of how to manage string information. We process all pieces string information nearly identically: Unique strings are exported, coded, and finally, re-imported into the source data. This process is implemented as a fixed part of the data-edition process.

3.1 Export and Import of String Information

Once all relevant string variables in the datasets to be processed have been identified, they have to be exported from the data file. This is implemented such that the information is saved to a spreadsheet format, making it easily processible with any kind of software or manual process. This export process follows a highly standardized procedure and is implemented in fixed Stata programs to ensure data quality. More information on the technical aspects of this procedure is documented in Bela in this volume.

In addition to the string to be coded, more auxiliary variables can be requested for export. To validly identify which information has to be exported, only a few parameters are needed. The example query in Table 4 illustrates this information: A variable “ts23201” out of dataset “C:\test\spEmp.dta” is coded into a target variable “ts23201_g10.” It should be automatically filled with information from a reference list called “reference.xlsx” and be provided to the coder “name@example.com.”

In NEPS data edition, these exports are controlled by a spreadsheet that contains this information. The edition scripts automatically export string variables accordingly and generate spreadsheets with coding information. All duplicate entries are thereby removed to avoid redundancy and to save resources (cf. Table 5). The advantages of this de-duplication are that (1) up to 25 % of all strings are saved and (2) every identical string gets the same code and treatment, (3) which ensures the important 1:1 or a 1:n relationship between source data and target data for merging the coded information back to the data.

As previously mentioned, we maintain reference lists that are separately stored for some items surveyed in open-text format, such as languages, countries, and list of studies. If an element of these lists is identical to the string, a code is assigned automatically and integrated into the exported spreadsheet file. Supervision can check the automatically assigned code once more to ensure data quality. This approach is well suited for less complex classifications with rather few elements that the respondents are familiar with.

After strings have been coded by the responsible staff, the spreadsheets are—again, fully automatically—re-imported to the datasets. This happens in an iterative proce-

Table 4 Parameters for the Export of String Information

Information	Content	Example
File path	directory in data-processing project	C:\test
File name	name of dataset with string variable	spEmp.dta
Source variable	name of string variable	ts23201 ¹
Target variable(s)	name of coded variable(s)	ts23201_g10
Auxiliary variable(s)	list of AVs	ts23204 ts23240 ²
Reference list(s)	access to list of reference for automatic pre-coding	reference.xlsx
Coder	supervisor of coding unit for the specific variable	name@example.com

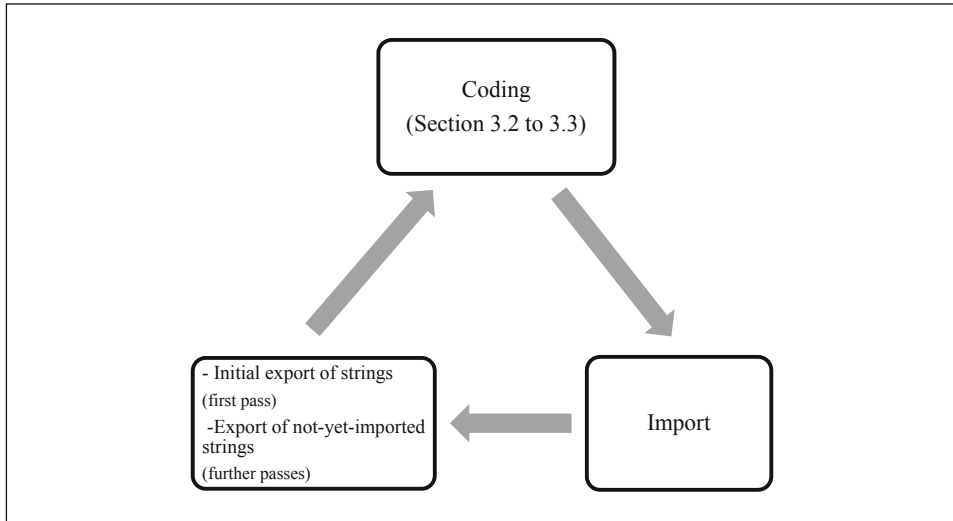
1 ts23201: open string information on occupation.

2 ts23205: differentiated occupational position; ts23240: open information about related industrial sector.

Table 5 The Export of Raw Strings

Original data				Exported data				
String	AV ₁	AV ₂	...	String	AV ₁	AV ₂	...	Code
bookkeeper	1	-55	...	bookkeeper	1	-55	...	34
gardener	6	2	...	gardener	6	2	...	56
gardener	6	2	...	gardener	3	4	...	12
gardener	3	4	...	mechanic	8	2	...	90
mechanic	8	2	...	social scientist	0	3	...	78
mechanic	8	2	...	teacher	-55	4	...	34
social scientist	0	3
teacher	-55	4	...					
...					

Note: AV = auxiliary variable.

Figure 4 Iterative work flow for export and import of string data

dure. First, coded information is imported. Afterwards, string information without any code is exported again, and the process starts over (cf. Figure 4). This is repeated until no uncoded information remains.

3.2 Computer-Generated Suggestions

When coding heterogeneous string information into very complex classifications, automatic coding is not practicable at first sight. As an alternative, one could try to suggest one or more possible codes using a computer-based approach. Afterwards, there should be an intellectual decision as to which of these suggestions (or any other solution) should be chosen.

There is a vast amount of material for which string coding is necessary, especially in the area of occupational coding. For instance, SUF version 1.0.0 of Starting Cohort 6—Adults contains nearly 200,000 coded occupations. However, the classifications to be coded are highly complex and multidimensional. Different techniques were used very early on to support coding via computational power.

The aim of the suggestion strategy is to enrich the raw material with one or more suggestions in a form

- that enables the coders to see the relevant information to be coded,
- that makes it possible to choose a suggestion or, if none of the suggestions is appropriate, to enter another solution, and

Table 6 Sheet with the Exported Data, Suggestions, and Information on the Sources of the Suggestion

String	Code	Industry	D. o. p.*	Distance
Verkäufer (Einzelhandel)		Einzelhandel	einfache Tätigkeit	
Einzelhandel; einfache Tätigkeit	-----			--
-----	-----			--
Verkäufer/in	B 62102-101	Einzelhandel	einfache Tätigkeit	0
Kaufmann/-frau—Einzelhandel	B 62102-100	Einzelhandel	einfache Tätigkeit	2
Verkäufer/in	B 62102-101	Einzelhandel	einfache Tätigkeit	2
Verkäufer/in—Einzelhandel	B 62102-117	Einzelhandel	einfache Tätigkeit	2
Verkaufssachbearbeiter/in	B 61122-101	Einzelhandel	einfache Tätigkeit	4
Einzelhandelskaufmann/-frau	B 62102-109	Einzelhandel	einfache Tätigkeit	4
Filialleiter/in, Verkaufsstellenleiter/in	B 62194-101	Einzelhandel	einfache Tätigkeit	4
Fachverkäufer/in—Sportartikel	B 62212-105	Einzelhandel	einfache Tätigkeit	4

* Differentiated occupational position.

- that can be converted to the structure that is expected by the import step (cf. Section 3.4).

The spreadsheet example in Table 6 has the required features:

- The first line shows the string to be coded.
- The second line shows the context information available for this case.
- There are a number of suggestions that follow.

Each line contains the information of the string and related auxiliary information. The coder is presented a linear transformation of the originally exported spreadsheet, which is enriched by suggested codes. After the coding process (Section 3.4), exactly one row additionally contains the decision of the coder, and the sheet can be re-collapsed to its original structure and then finally be imported (Section 3.4). This technique allows for using widely available office software as a front-end for the coders. The sheets, which are edited by the coders, could be used to evaluate and improve the quality of the suggestion.

The sources of the suggestions of the occupational include three items:

- coded material of previous coding jobs,
- the official classification, and

Table 7 Consolidated Source of the Suggestions

Reference string	Code	Source
Verkäufer (Einzelhandel)	B 62102-101	search keyword
Verkäuferin (Einzelhandel)	B 62102-100	previous coding job
Verkäufer Einzelhandel	B 62102-101	previous coding job
Kaufmann/-frau—Einzelhandel	B 62102-100	classification

- search keywords provided by the Federal Employment Agency, which are used, for example, in the agency’s online services.

These sources were consolidated to a simple table with three columns, in which the first column contains the text (e. g., the already-coded response, the term of the classification, or the search keyword) used as a reference for determining suggestions. The second column contains the corresponding code, and the last column contains the source.

In the next step, the distances between the string and the suggestions are calculated. The records of the source files with the smallest distances to the string serve as suggestions, which are arranged in the structure shown in Table 6. The programming of this step is done in R⁴ (R Development Core Team, 2011) using the “adist” command from the “utils” package, which calculates a generalized Levenshtein distance. As a large amount of distances have to be calculated, a sheet with 600 strings needs up to 6 hours of computing time. It would be possible to optimize this (e. g., by parallelization), but at this time, the suggestion procedure does not (yet) appear to be the bottleneck of the whole process.

3.3 Manual Coding

There are two general manual coding procedures at the NEPS: coding with and without computer-generated suggestions.

Manual coding without computer-generated suggestions

We mainly run the procedure described in Section 3.2 on up-to-date occupational and vocational-training information. All other string information collected in NEPS interviews and questionnaires (e. g., further education courses, sports, industrial sectors) is not yet pretreated with the suggestion process as this is a preliminary proce-

4 <http://www.r-project.org>

ture that still is to be generalized. Thus, there are no suggested codes available for the coders. Nevertheless, some string information⁵ is automatically coded during the export of raw strings as described in Section 3.2. Afterwards, every coding unit (NEPS Data Center, other NEPS departments) assigns the codes of selected classifications to the exported string information. Every coding unit maintains its own rules and conventions for handling string information.

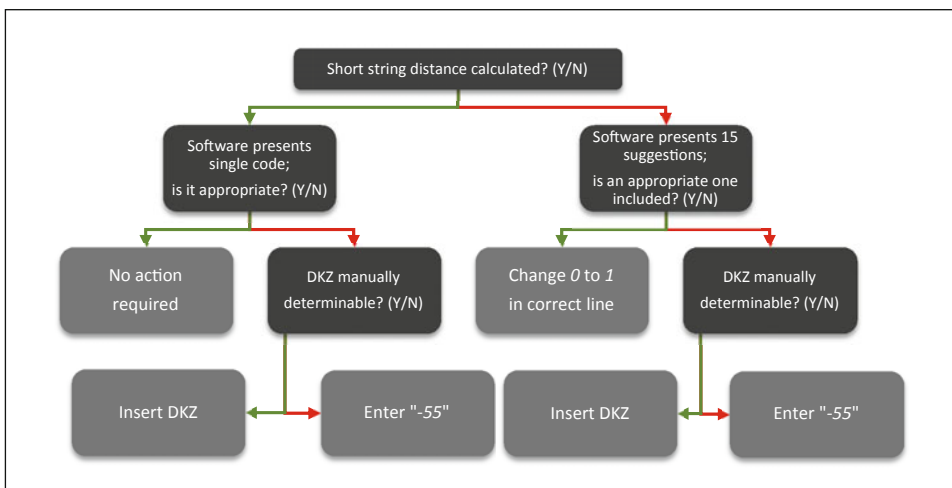
After this step, every coding unit has to send back the coded strings to the NEPS Data Center for further treatment (see Section 3.4 f).

Manual coding with computer-generated suggestions

The manual handling of strings with computer-generated suggestions is a bit different from the process described in Section 3.2. Coders find proposed codes for the related strings as described in Section 3.2. Dependent on the shortest distance detected by the suggestion system, the software presents either 15 suggestions or the single suggestion with a distance of approximately 0. From there on, coders have to evaluate the suggestions intellectually by following a decision tree (cf. Figure 5). In both cases, they end up either selecting a suggestion to be appropriate, manually entering a true code (that has not been presented by the software), or classifying the string observation as “not determinable” (code: -55).

In the multiple-suggestion case, up to 15 suggested codes with the lowest string distance are integrated into the proposal set for the judgment.

Figure 5 Multi-stage decision tree for coding and editing cases with suggested codes (left branch: “yes”-decision; right branch: “no”-decision; -55: code for “not determinable”)



5 Languages, countries, subjects of studies.

The software could not find any perfectly matching entry in our dictionary, so the first 15 closest-matching items were integrated in the proposal with the suggested codes. As described above, the coder either has to accept a suggestion by changing the “0” to a “1” in the fourth column or has to enter a new code into the marked cell near the string. If neither option is appropriate, the coder can also enter the missing value “-55 (not determinable)” into this cell. In the presented example, the second suggestion (code “B 62102-101”) would designate the appropriate choice and is consequently marked with “1.”

In the case that the suggestion algorithm does find a string match with a distance of almost zero,⁶ only the nearly perfect match is presented to the coder (cf. Table 9). If the suggestion is correct, no further input is required. Otherwise, the coder can also enter a manually queried code or the missing code “-55 (not determinable).” In the presented example, the suggestion is correct, and no further input is required.

The last example in Table 10 shows a multiple-suggestion scenario in which no suggestions are appropriate codings for the originally observed information. The string “München” represents a city instead of an occupation. This perfectly illustrates the need for intellectual interpretation of the proposals: The calculation of string distances is a technical solution with no chance for interpreting the output. The software cannot replace the human coder, who has to enter the missing value “-55 (not determinable).”

3.4 Derivation of Further Classifications

The next step contains the derivation of further classifications that are mainly based on the coded strings and other auxiliary variables. It is useful to offer further scales to maximize the data power. Every coding unit delivers the how-to-instruction⁷ for the generation of the derived variables. Occupational information and related AVs enable the generation of up to 15 different variables⁸ (see *Figure 6*).

We can derive the various variables about occupations by using the code of the Documentary Code Number (DKZ2010) and the related transcoding schemes (e. g., DKZ2010 to DKZ88: Bundesagentur für Arbeit, 2013b; ISCO-08 to ISCO-88: ILO, 2013). Some transcoding schemes had to be modified due to transcoding problems caused by 1:n or m:n⁹ assignments. The possible approaches to solve these problems were presented in Section 2.1. All derived variables receive the identification

6 Such an assignment is made if the distance is lower than or equal to 5% of the string length, that is, a string with 20 characters allows an absolute string distance of 1 character.

7 For example, in terms of a Stata do-file.

8 For example, due to data-protection issues; however, the NEPS offers a maximum of 12 variables.

9 The code of the target classification cannot be identified directly because of multiple possible assignments.

Table 8 Coding Proposal with Multiple Computer-Generated Suggestions (Abbreviated Version of Table 6)

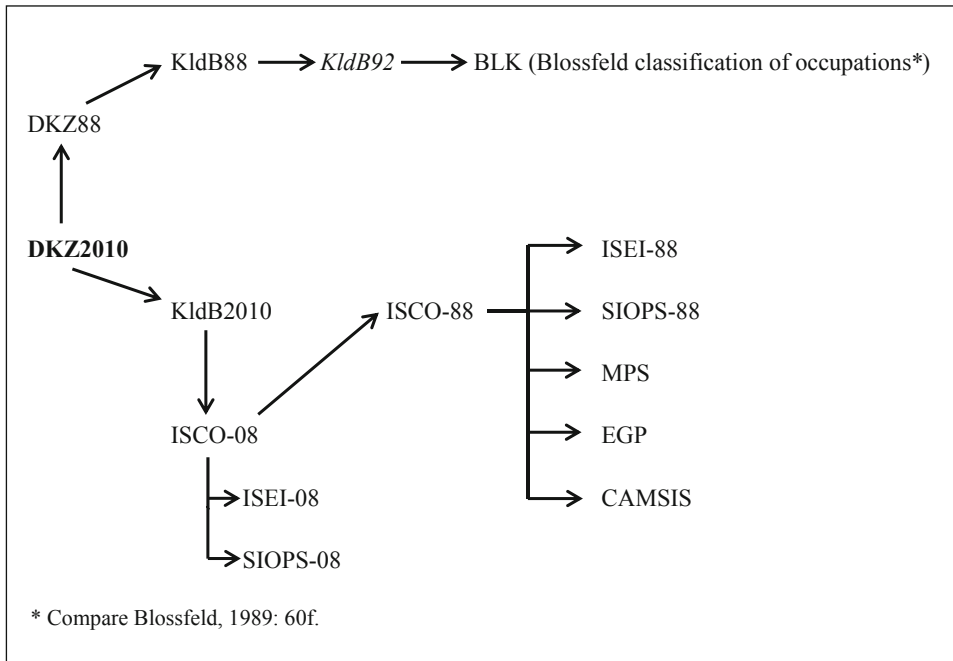
String Information	Distance	Code	Rating
Verkäufer (Einzelhandel)			3
Einzelhandel; einfache Tätigkeit	–	-----	--
-----	–	-----	--
Verkäufer/in	0	B 62102-101	1
Kaufmann/-frau—Einzelhandel	2	B 62102-100	0
Verkäufer/in	2	B 62102-101	0
Verkäufer/in—Einzelhandel	2	B 62102-117	0
...

Table 9 Coding Proposal With Only One Match

String Information	Distance	Code	Rating
Milchwirtschaftliche Laborantin			3
Molkereiwirtschaft; qualifizierte Tätigkeit	–	-----	--
-----	–	-----	--
Milchwirtschaftliche/r Laborant/in	0	B 41212-106	2

Table 10 Multiple-Suggestion Scenario in Which No Suggestions Are Appropriate

String Information	Distance	Code	Rating
München			3
-----	–	-----	--
Postzusteller/in	17	B 51322-101	0
Fachkraft—Brief- und Frachtverkehr	17	B 51322-104	0
Physikochemiker/in	19	B 41384-103	0
Hauswirtschafter/in	23	B 83212-111	0
...

Figure 6 The NEPS transcoding scheme for occupational information

suffix “_g#” (with # as a number for each derived variable per string variable). Some scales demand additional specific information that is not coded in the KldB2010.¹⁰

4 Outlook

The coding process in this generic framework will be enhanced in the future. We consider several possible improvements for the future.

4.1 Export of Strings

All string information could be standardized before the export of the strings. Every letter of the string should be transformed into a lowercase letter, and possible spaces at the beginning and the end of the string should be deleted. Furthermore, German

¹⁰ For example, CAMSIS (cf. Prandy & Jones, 2001) needs additional information about the sex and the occupational position.

umlauts and the sharp “s” (“ß”) should be transformed, as well.¹¹ Other possible standardization measures are possible. This would decrease the amount of strings by an additional 5–10 %.

4.2 Creating New Dictionaries

We have dictionaries for occupations, vocational training information, languages, countries, and a list of fields of study. These lists serve as reference material for computer-generated suggestions and for the automatic coding of already-coded strings. The number of lists could be increased by integrating lists for information about branches/industries, further education/courses, lists of sports, and every other piece of accumulated string information that should be coded. As these lists increase over time, less information would have to be coded manually.

4.3 Computer-Assigned Suggestions

Moreover, clear assignments (with a negligible distance) can be treated as coded cases that won't be delivered to the coder. Supervision should nevertheless check the assignment for inconsistencies. The investment for the manual coding would thereby decrease over time due to a growing dictionary and an increasing ratio of clear assignments. Therefore, a set of information (string and auxiliary variables) and a threshold for the accepted distance would have to be defined, and the source of the suggestions would have to be extended to these auxiliary variables. If the material that is to be coded comes with this set of information and the best hit is under the defined distance, the suggestion could be considered to be already coded.

References

- Blossfeld, H.-P. (1989): *Kohortendifferenzierung und Karriereprozess: Eine Längsschnittstudie über die Veränderung der Bildungs- und Berufschancen im Lebenslauf*. Frankfurt (Main), New York : Campus Verlag.
- Bundesagentur für Arbeit (2011): *Klassifikation der Berufe 2010: Definitiver und beschreibender Teil (Band 2)*. Nürnberg: Bundesagentur für Arbeit.
- Bundesagentur für Arbeit (2013a): *Gesamtberufsliste_der_BA.xlsx*. Retrieved from <http://download-portal.arbeitsagentur.de/files/download?fid=1654>

11 For example, “ß” to “ss” and “ä” to “ae.”

- Bundesagentur für Arbeit (2013b): *Umsteigetabellen_Implementierung_KldB2010.zip*. Retrieved from <http://download-portal.arbeitsagentur.de/files/personDetail.do?sortierfeld=&doNext=detailAnzeigen&dateiId=1154&breadcrumb=list>
- Elias, P., & M. Birch (1994): *Establishment of community-wide occupational statistics. ISCO 88 (COM)—A guide for users*. Retrieved from University of Warwick, Institute for Employment Research website: <http://www2.warwick.ac.uk/fac/soc/ier/research/classification/isco88/englishisco.doc>
- Ganzeboom, H. B. G., & Treiman D. J. (1996). Internationally comparable measures of occupational status for the 1988 International Standard Classification of Occupations. *Social Science Research* 25(10), 201–239.
- Ganzeboom, H. B. G., & Treiman D. J. (2010). *Occupational status measures for the new International Standard Classification of Occupations ISCO-08; with a discussion of the new classification*. Retrieved from <http://www.harryganzeboom.nl/isol/isol2010c2-ganzeboom.pdf>
- ILO (2013). *Correspondence table ISCO-08—ISCO-88*. Retrieved from <http://www.ilo.org/public/english/bureau/stat/isco/docs/corrtab08-88.xls>
- Paulus, W., Schweizer, R., & Wiemer S. (2010). *Klassifikation der Berufe 2010. Entwicklung und Ergebnis*. Retrieved from <http://statistik.arbeitsagentur.de/Statischer-Content/Grundlagen/Methodenberichte/Arbeitsmarktstatistik/Generische-Publikationen/Methodenbericht-Klassifikation-Berufe-2010.pdf>
- Prandy, K., & Jones F. L. (2001). An international comparative analysis of marriage patterns and social stratification. *International Journal of Sociology and Social Policy*, 21(4-6), 165–183.
- R Development Core Team (2011). *R: A language and environment for statistical computing*. Retrieved from <http://www.R-project.org/>

About the authors

D. Bela
Leibniz Institute for Educational Trajectories (LifBi), Bamberg.

M. Munz
STR coding, Nuremberg, Germany.
e-mail: manuel.munz@str-coding.de

K. Wenzig
German Institute for Economic Research (DIW Berlin), Berlin.

Visualizing Life Courses With the TrueTales View

Ralf Künster

Abstract

Collecting complete educational and employment histories, so-called life courses, is a key objective of the National Educational Panel Study (NEPS). The TrueTales View is a practical tool for assessing the quality of life-course data and for gaining an idea of the analytic potential of this kind of data by integrating information on various episodes into complete life courses and by relating them to varying time references and group attributes.

The TrueTales View serves two main tasks: First, the application enables users to create status-distribution charts that display the status development of selected respondent groups along a timeline. By creating status-distribution charts from monthly based data, the TrueTales View has proven to be a fast and easily manageable device for explorative analysis of the life courses collected from the respondents of NEPS Starting Cohort 6—Adults. Second, the TrueTales View is useful for editing data by displaying individual life courses as bar charts and in table form. Visualizing individual life courses facilitates the necessary process of data editing in an intuitive and therefore instantaneously comprehensible way by supporting the consideration of the quality, consistency, and complexity of the collected data.

1 Introduction

Collecting complete educational and employment histories, so-called life courses, is a key objective of the National Educational Panel Study (NEPS). A main goal is to test hypotheses of causal relations between educational attainment and other life domains. Since causes always chronologically precede their effects, causal relations can be observed primarily with the help of longitudinal data.

The participants of the first panel wave of the Starting Cohort 6—Adults (SC6) were between 23 and 66 years old at the date of the interview. Starting with the date of first school enrollment, the information on each activity of the 11,649 respondents is captured in episode form with monthly precision. We collected more than 135,000 episodes, which amounts to a total of about 6,200,000 reported months. Detailed status information is available for each episode (e.g., theoretically more than 100 variables for the status *employed*), and the amount of information will increase in the future due to the panel design. For the respondents of this starting cohort, we expect at least 100,000 new monthly records each year.

Owing to the characteristics of life-course data, the questions that can be answered with the data are strongly linked to the time reference to which the data is set. Age effects, cohort effects, and period effects may be observed independently depending on whether the information is related to the respondents' age or to historical time.

In addition, life-course data in one domain (e.g., educational and occupational history) can be related to events directly linked to this domain (e.g., successful completion of the first vocational training program, the end of a first unemployment spell), to events in other life domains (e.g., marriage or birth of the first child in the family-formation history), or to historic events (e.g., the German reunification). Each alteration of the time reference leads to a new perspective on the longitudinal data because the data are linked to each other in a new way.

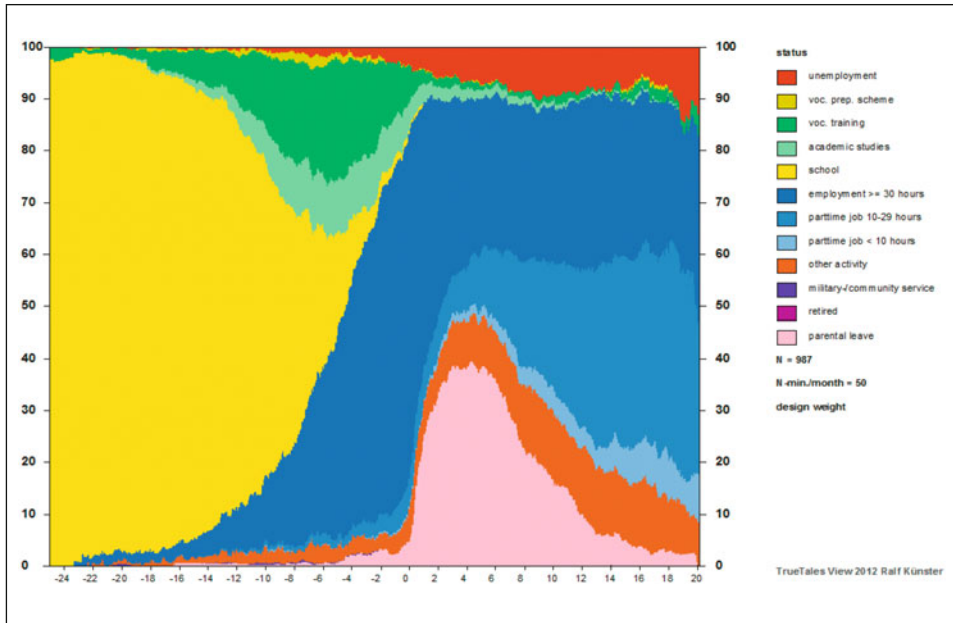
Obviously, information from life-course data increases exponentially compared with cross-sectional data.¹ Moreover, even simple tasks like conducting a frequency count for a certain time point require the user to be able to aggregate and merge the longitudinal files. Considerable data-management skills are indispensable for retrieving the desired information from life-course data.

I developed the TrueTales View to provide a good practical tool for assessing the quality of life-course data and to get an idea of the analytic potential of this kind of data by integrating the information on the various episodes into complete life courses and by relating them to varying time references and group attributes.

The visualizing tool was first created for the German Life History Study—GLHS (Max-Planck-Institute for Human Development—Berlin) in 2004 (Hillmert, Künster, Spengemann, & Mayer 2004; Matthes, Reimer, & Künster 2005; Matthes, Reimer, & Künster 2007; Matthes, & Reimer 2007). Later on, it was modified for use in different contexts of other longitudinal surveys (*Competence and Context* (Cocon)—Jacobs Center Zurich, *BIBB-Transition Survey 2006*—Federal Institute for Vocational Education and Training Bonn (Rohrbach-Schmidt 2010), *Working and Learning in a Changing World* (ALWA)—Institute for Employment Research Nürnberg (IAB)

1 Most statistical methods for analyzing life-course data use only a small amount of the available information by limiting the focus to a single transition from a given state to a target state. An exception is sequence analysis, which is capable of including information from all episodes of the collected life courses and searches for typical life-course patterns (Brzinsky-Fay, Kohler, & Luniak 2006).

Figure 1 Status-Distribution chart (women, age 35–44, living in West Germany, excluding Berlin)

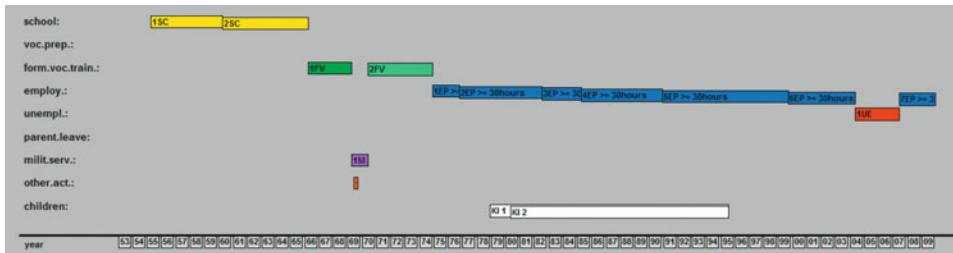


(Drasch, & Matthes 2009; Matthes, Drasch, Erhardt, Künster, & Valentin 2012), *The Entwined Life Courses of Academic Couples*—Social Science Research Center Berlin (WZB) (Rusconi, & Solga 2011)). The core idea of visualizing life courses has been used in these surveys to collect life courses, to edit life-course data, and to explore and analyze the data.

The current version of the TrueTales View serves two main purposes: First, the TrueTales View enables users to create status-distribution charts, which display the status development of selected respondent groups along a timeline (Figure 1). Through its creation of status-distribution charts from monthly based data, the TrueTales View is a fast and easily manageable tool for explorative analyses of the life courses collected from the respondents of NEPS Starting Cohort 6.² It allows for changing the design parameters of the charts, such as the order of the displayed states, and for selecting different references of the timeline (age, historical time, first child's date of birth, etc.).

2 This paper uses data from the National Educational Panel Study (NEPS): Starting Cohort 6—Adults, doi:10.5157/NEPS:SC6:1.0.0. From 2008 to 2013, NEPS data were collected as part of the Framework Programme for the Promotion of Empirical Educational Research funded by the German Federal Ministry of Education and Research (BMBF). As of 2014, the NEPS survey is carried out by the Leibniz Institute for Educational Trajectories (LifBi) at the University of Bamberg in cooperation with a nationwide network.

Figure 2 Visualization of an individual life course (a fictional example)



Second, the TrueTales View is useful for editing data by displaying individual life courses as bar charts (Figure 2) and in table form. Visualizing individual life courses facilitates the necessary process of data editing in an intuitive and therefore instantaneously comprehensible manner. Furthermore, it helps to consider the quality, consistency, and complexity of the collected data.

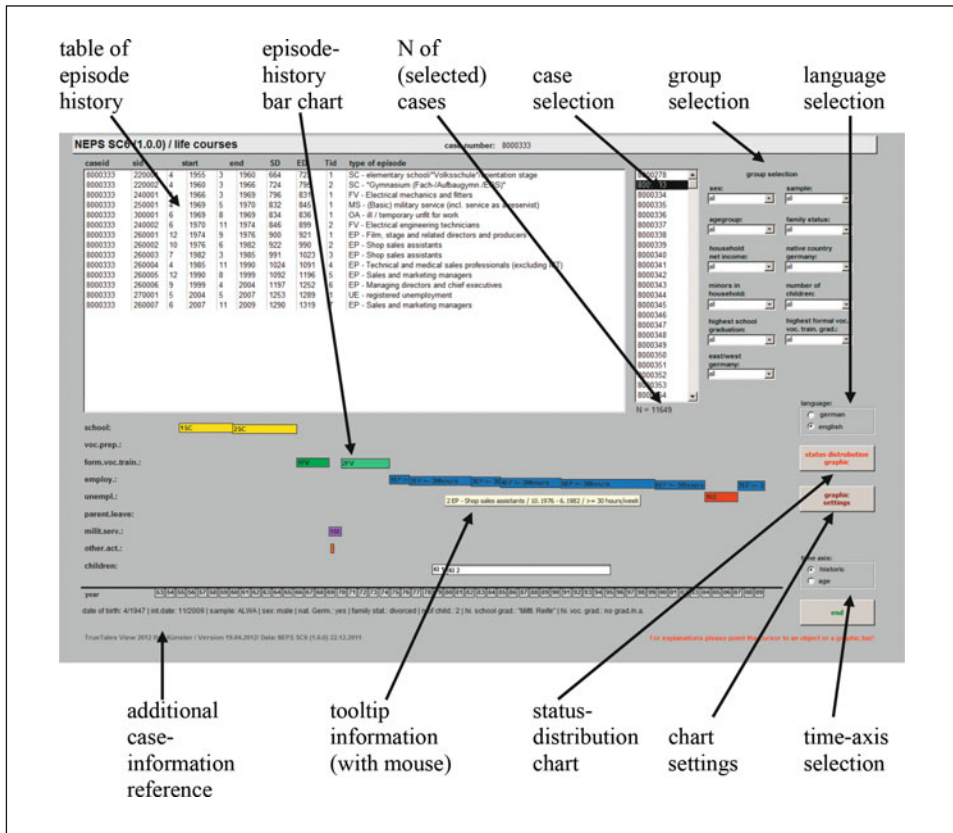
The TrueTales View is programmed with VBA (*Visual Basic for Applications*) and retrieves the data from a database. The application and data are stored in the same file. In order to use the TrueTales View, the installation of MS-Access 2000 or a more recent version of MS-Access is required.

2 The Visualization of Individual Life Courses

Life courses are composed of episodes spent in different states of a certain life domain. An episode is defined by the start and end date of a certain status a person holds during a given period. In the NEPS context, the following states of an individual's educational and occupational history are observed: school attendance, preparatory vocational training schemes, vocational training and academic studies, employment, unemployment, parental leave, military or community service. Periods not assigned to one of these states are covered by the category *other activities*. This category includes periods of homemaking and family care, illness, retirement, and other episodes of non-participation in education and the labor market. Start and end dates of episodes are measured with monthly precision. Unique episode numbers are added to each episode to provide for persons performing more than one activity of the same status simultaneously (like having two concurrent jobs). The intention of the NEPS survey is to collect complete and consistent life courses from our participants without any chronological gaps.

Whether this goal has been met can be assessed using the individual-case screen of the TrueTales View (Figure 3), which appears after the start of the program and offers an extensive impression of the quality of each life course. This part of the program is

Figure 3 Individual case screen (a fictional example of a life course)



particularly intended for editing the life-course data immediately following the data-collection process, not for analyzing individual cases.

The individual-case screen is divided into three functional sectors:

- a table containing the collected episodes of a case;
- the graphical visualization of the life course of this case as a bar chart; and
- the selection options in the right sector of the screen, in which individual cases, groups of cases, and program functions can be selected.

In the tabular representation, the episodes are sorted by date. In addition to the start date, end date, and episode number, there is information about the type of state and a more detailed description of the activity.

The graphical visualization of an individual life course consists of the different episodes plotted as bars along the time axis. Each type of state is plotted on a different

horizontal level and marked with a special color. The bars have captions to identify the episode related to them. A tooltip linked to each bar provides additional information about the episode represented by the bar.³ At the bottom of the bar chart, periods of living together with a biological or adopted child are displayed. This allows for observing the influence of having children on the respondent's educational and occupational career.

The reference of the time axis can be switched between historical time and the respondent's age. The graphical visualization primarily allows for a fast and easy uncovering of inconsistencies within life courses by reducing the complexity of the data to a visual pattern.

There is also an option to switch the program language between English and German.

Beneath the bar chart, more information related to the respective case is displayed in an abbreviated form (date of birth, date of interview, sample, sex, native- or foreign-born, marital status, number of children, highest school degree, highest vocational-training- and academic degree).

Each case of the survey is accessible through a list of case numbers. Drop-down lists allow for a group selection of cases. The available selection criteria are sex, sample, age group, marital status, household net income (grouped), native- or foreign-born, minors in the household, number of biological and adopted children, highest school degree, highest vocational-training- and academic degree, and resident in the East Germany or West Germany.⁴ The choice made here reduces the selectable case numbers of the case-selection list. The number of cases resulting from this selection is displayed below the case-selection list and is updated after each selection. The chosen selection also has an immediate impact on the creation of the status-distribution charts, which are described below.

3 Creating Status-Distribution Charts

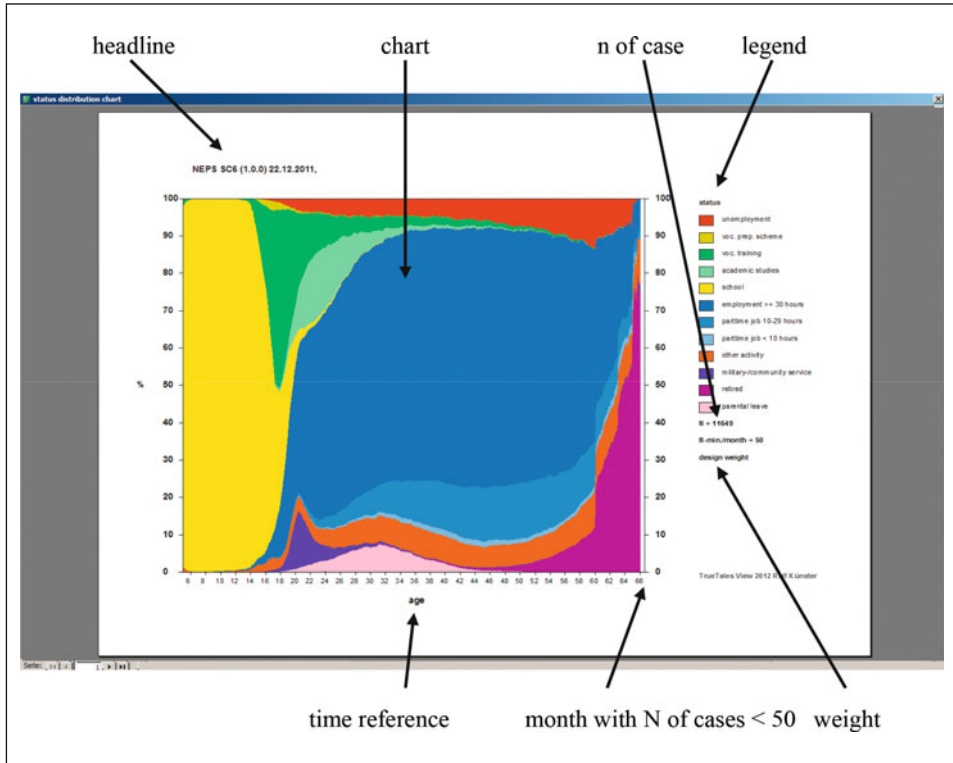
Status-distribution charts describe the time-varying status composition of a sample along a timeline. As a precondition, which kind of status the person had must be determinable for each month of the observation period and each person in the sample. We thereby need exact or at least well-estimated dates for each episode.⁵ To create a status-distribution chart, a person's status has to be unique at a certain point of time.

3 This kind of extended allocation of information through pointing the mouse to an object is available for nearly every object on the screens of the program.

4 For time-dependent criteria, such as living in East or West Germany, the situation at the date of the interview was taken.

5 Not every person can remember and date all the spells of their life course in their exact chronological order. Thus, there is often diffuse or missing date information, which is complemented by plausible date estimates during the data-collection process of the adult starting cohort.

Figure 4 Status-Distribution chart



Thus, a person is not allowed to be in two or more different states at the same time. In case of such concurrent episodes, a priority order has to (and can) be defined by the program user by indicating which of the concurrent states should be included in the calculation of the chart.

For purposes of differentiation, the state *employment* has been split up into *full-time employment with 30 hours per week and more*, *part-time employment with 10 to fewer than 30 hours per week*, and *part-time employment with fewer than 10 hours per week*. As retirement constitutes an important phase in life, it is not subsumed under *other activities*, but rather forms its own state, *retired*.

Percentages of the various states within a selected population are calculated for each month of the observation period. These monthly rates, which add up to 100%, are plotted as stacked bars along a timeline. The timeline reference should be chosen according to the objective of the analysis. Age or historical time are common references, but crucial events like the birth of the first biological child may also be used to calibrate the timeline. If the number of cases drops below 50 for a certain month, no distribution is displayed in the chart for this month.

The above-mentioned group-selection criteria are also used for creating the status-distribution charts and for limiting the number of cases included in the calculation to those that match the criteria.

The resulting figure illustrates changes in the sample's status composition over time (Figure 4). The example above contains all available cases of the adult starting cohort, revealing clearly and immediately, in addition to other things, the age span during which participation in formal vocational training, military service, and parental leave is most frequent. Moreover, it shows the life period in which the transition into retirement takes place, as well as the extent of this transition, illustrating that the risk of being unemployed increases with age. As a result, the observer is able to monitor the crucial events of each life phase. Later on, there are some examples of comparing status-distribution charts for different groups, which highlight the advantages of visualizing life-course data in this way.

4 Chart Settings

The “Chart Settings” menu (Figure 5) provides a set of options affecting the design, content, and interpretation of the status-distribution charts. Some settings have a direct impact on the calculation of the distributions (e.g., changing the time reference, selecting a different weighting measurement, or changing the priority order of the states). Other options are helpful for shaping the chart so it can be interpreted and compared more easily.

Here is a short description of the available options:

- The time reference can be set to historic time, age, birth of the first child, first marriage, first divorce, end of the first formal vocational training period, or end of the first unemployment episode. Changing the time reference structures the data in a completely new way and displays how status distribution changes after the occurrence of a major event, such as the birth of the first child.
- By using the *weight* option, it is possible to choose from among all available weighting factors or unweighted cases. The weighting is necessary “to account for sampling design and systematic nonresponse in the sample” (Aßmann, & Zinn 2011).
- The option to limit or extend the starting and ending point of the time axis is helpful when charts of different age groups are to be compared. However, the comparison should cover the same life period for these age groups (e.g., the period between age 20 and 30).
- Changing the ‘priority in case of concurrent episodes’ allows for defining the priority order of the states. This is relevant for persons holding two or more different states during the same period of time. In this case, it is necessary to decide which of the different states should be used for calculating the status-distribution chart.

Figure 5 Chart settings

changing chart settings

time reference:
 age
 birth of 1 child
 1 marriage
 1 divorce
 end of 1 Training
 end of 1 unemployment

start of time axis at: end of time axis at:
 Please specify in month from the 0-point of the time axis!

weight:
 calibrated with MZ2008
 calibrated with MZ2009

priority in case of parallel episodes:
 unemployment
 retired
 parental leave
 voc. prep. scheme
 military-/community service
 voc. training
 academic studies
 employment >= 30 hours
 school
 sideline job: 10-29 hours
 sideline job < 10 hours
 other activity
 Select an episode type with a mouse click and shift it with the arrow buttons.

high

 low

display order:
 unemployment
 voc. prep. scheme
 voc. training
 academic studies
 school
 employment >= 30 hours
 sideline job < 10 hours
 sideline job < 10 hours
 other activity
 military-/community service
 retired
 parental leave
 Select an episode type with a mouse click and shift it with the arrow buttons.

up

 down

sort legend by:

 save data to file:

For explanations please point the cursor to an object!

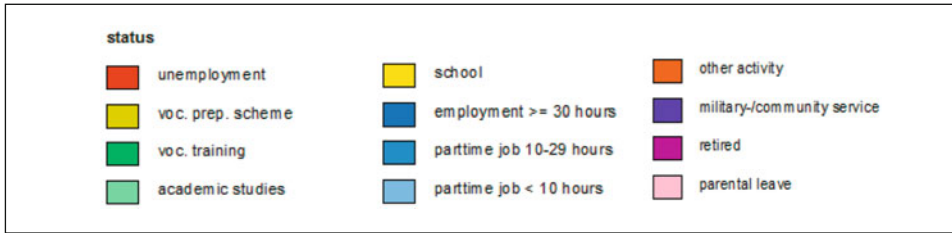
If, for example, the analytic interest focuses primarily on part-time jobs with fewer than 10 working hours per week, this state should be given a higher priority than the other states. Otherwise, it will be mostly concealed by states with higher priority.

- The 'display order' option defines the horizontal order in which the states are plotted in the chart. The highest entry in the list is plotted at the upper fringe of the chart, the lowest entry is plotted at the lower fringe, and the other states are plotted in relation to their order in the list. The distributions of the states that are plotted at the fringes are especially easy to interpret and should therefore be reserved for the states that are in the main focus.
- The state specifications of the legend can be sorted either in the order of their priority or in the order of their display order. The former is the only way to include information about the selected priority order in the chart, whereas the latter is much more intuitive because it refers to the states in their visualized order.
- A table of the percentage-status distribution can be stored in an external ASCII file. The data is then available for further processing with other applications (e.g., Excel).

5 Four Examples for Using Status-Distribution Charts

The examples on the next pages are meant to provide an idea of the illustrative power that status-distribution charts can develop beyond statistical measures. We examine comparisons of the distribution of different respondents groups, which show that there are considerable differences in these groups' participation in diverse activities.

Concerning the following examples, all charts refer to the same legend, which is displayed only once in order to leave more room for the actual comparisons.

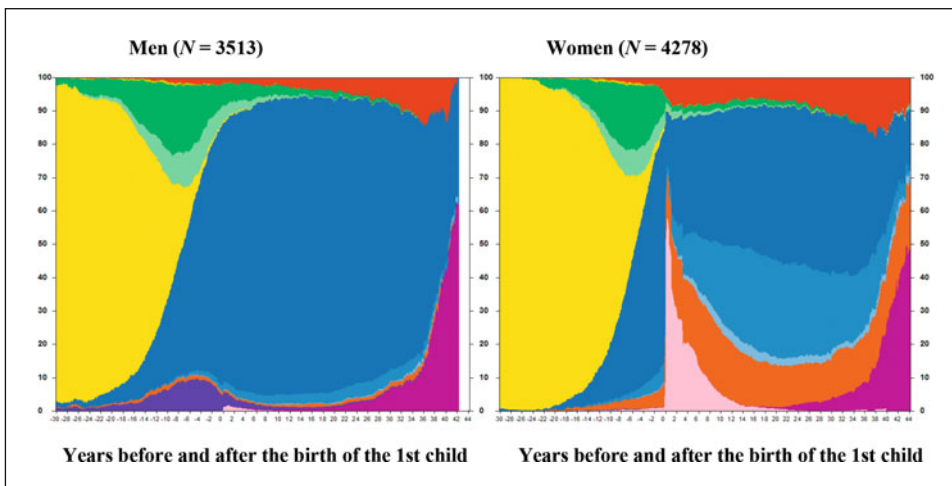


5.1 Comparison of Men and Women With Respect to a Certain Event (Date of Birth of the First Child)

The 0-point of both charts is adjusted to the birth of the first child. While men’s (occupational) life courses are virtually unaffected by the birth of a child, women’s life courses change dramatically. After a certain period of maternity leave, many women do not return to full-time employment. We observe a considerable increase in unemployment, part-time employment, and *other activities* (which mostly mean home-making after the end of parental leave). The women’s full-time employment rate does not recover to the level of the men’s employment rate, even 20 to 30 years after the birth.

A chart for childless women, which is not presented here, would reveal that the shape of their employment distribution is much closer to that of the men and that higher education and childlessness are highly related among women.

Figure 6 Comparison of men and women with timeline reference ‘date of birth of the first child’



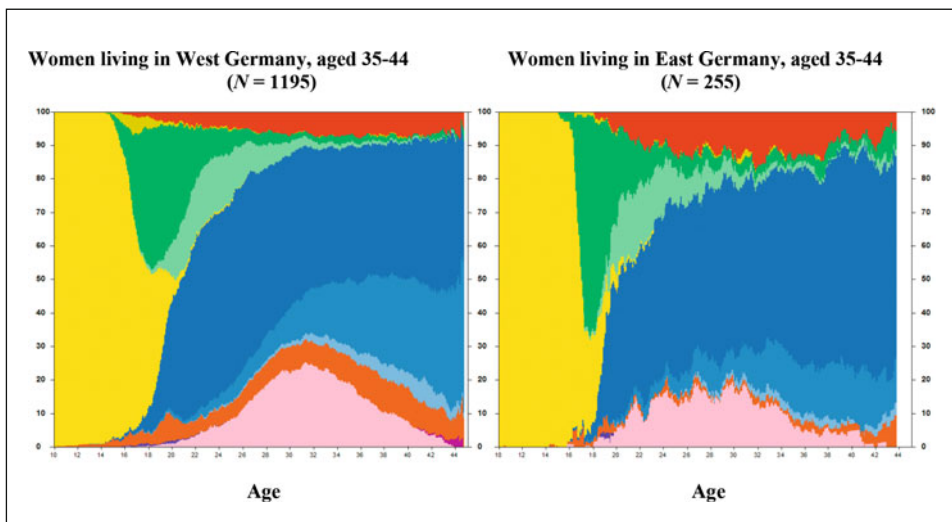
5.2 Comparison of Women Regarding Their Actual Place of Residence

If we compare women living in West and East Germany when they are between 35 and 44 years old, it is striking that women living in West Germany take parental leave at the average age of about 32, whereas women living in East Germany experience this episode mainly at the age of 28 to 29. Thus, the latter have their first child approximately 3–4 years earlier.

The effect of having children on the further employment history also seems to be different. While up to 50 % of West-German women at the age of 43 dedicate themselves to part-time employment or *other activities*, only up to 25 % of their East-German counterparts are in one of these two states at the same age. Conversely, more than 60 % of East-German women work full-time compared with only 40 % of West-German women. The result is clear: Even more than 20 years after the fall of the wall, there are marked differences between women living in East and West Germany with respect to their labor force participation.

The significantly higher rate of unemployment among East-German women is, on the one hand, probably due to their lower tendency to switch to a state of occupational inactivity. On the other hand, the fall of the wall and the decline of the East German economy has had a major impact on these women's employment careers (a look at the status-distribution chart of this group using the time reference *historical time* would confirm this assumption) (Diewald, Goedicke, & Mayer 2006; Mayer, & Solga 2010).

Figure 7 Comparison of women at the age of 35–44, living in West or East Germany, with timeline reference 'age'



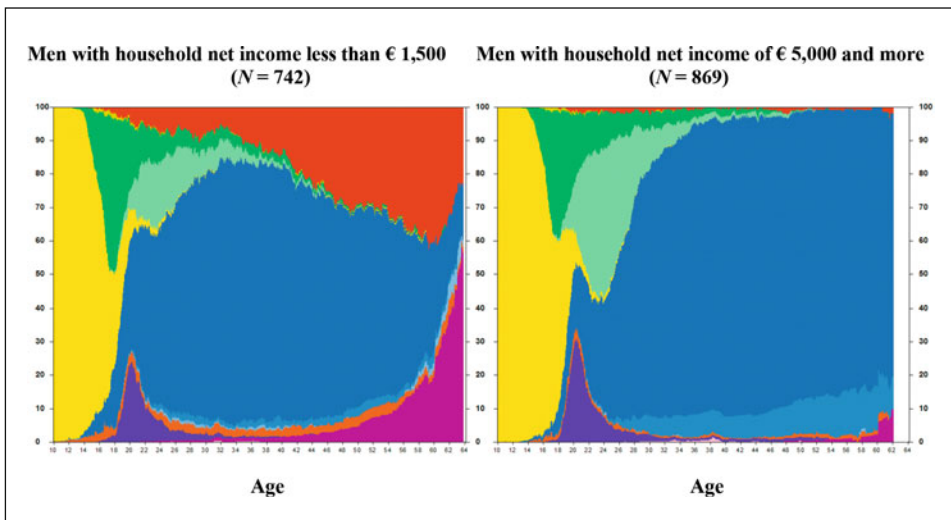
Another surprising finding is the notably higher rate of East-German women's attendance of vocational-training programs to become a skilled worker at the age of 17–18 (ca. 65 %) compared with their western counterparts (40–45 %).

5.3 Comparison of Men Regarding Their Actual Household Income

The perspective of this example begins with a given discrepancy in the size of the household net income of two groups of men and turns back to find the source of this discrepancy. Obviously, one of the main reasons for a high income is the previous amount of educational participation and attainment. *High-earning men* display a much higher level of university attendance. In contrast, the period of vocational training among *low-earning men* takes place earlier in life, covers a shorter period, and is less academic. Moreover, the overall attendance rate is lower. 'Low-earning men' face a higher risk of becoming unemployed in an early phase of their occupational career, a risk that considerably increases with age. They are also more likely to retire mid-career. Extensive unemployment experiences and early retirement are both responsible for a low household income.

Surprisingly, men with a high household income show a relatively high rate of part-time employment compared with the average of all men. The reason for this remains to be discovered.

Figure 8 Comparison of men with a household net income of up to € 1,500 versus more than € 5,000

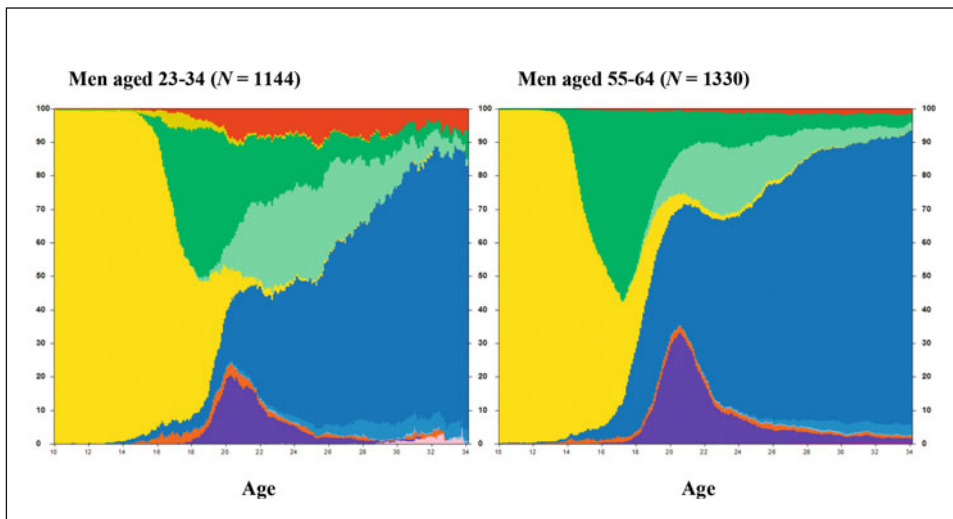


5.4 Comparison of Men From Different Cohorts for the Same Age Period

To a remarkable extent, the male members of the younger cohort more often opted for university/higher education and less often for vocational training as a skilled worker. Apparently, taking parental leave is still uncommon for all men, even in the younger cohort. The younger cohort completed military and community service considerably less often and was exposed to a higher risk of unemployment than the older cohort. The high level of unemployment among the younger cohort is an especially striking indication of the fact that establishing oneself in the labor market seems to be much more difficult today than it was for the older cohort. This points to negative labor market developments.

These four examples illustrate that there is a lot to gain by visualizing life-course data as status-distribution charts. Differences in the distributions between groups and in the impact of historical or individual events are clearly visible and provide an opportunity to form hypotheses about the causes of these effects. On the other hand, it should be mentioned that status-distribution charts average individual life-course dynamics by summing up the monthly status attendance from often strongly heterogeneous life courses. Therefore, status-distribution charts only tell a part of the whole story.

Figure 9 Comparison of men aged 23–34 versus 55–64, starting with school enrollment up to the age of 34



6 Further Development and Availability of the TrueTales View

As the NEPS is designed as a panel survey, we will continue to collect information on the life courses of the adult starting cohort in annual intervals. The TrueTales View will be updated promptly with the latest data.

A similar tool is likely to be created for other NEPS starting cohorts, such as Starting Cohort 4 for Grade-9 students passing into vocational training, for whom rich information about the life courses is also available.

There are plans to expand the status-distribution charts to life histories in other domains, such as family-formation history, including partnerships, changes of marital status, and spells of living together with children.

Another development will be the expansion to include grouping variables in a way that enables users to choose from nearly the whole range of available cross-sectional data and to combine these variables according to their own needs. Whether this is also possible for time-dependent variables and date variables, which would then calibrate the time axis, remains to be evaluated.

It has not yet been decided whether the fully functional version of the TrueTales View will be made available to the scientific public together with Scientific-Use-File data from Starting Cohort 6—Adults. There are some concerns about data-protection problems in connection with the visualization of the individual life courses. Therefore, a program version will be offered in which access to the visualization of individual life courses is prohibited while all functionalities to create status-distribution charts remain. The tool will be available in its full functionality for all researchers using the NEPS data on-site and perhaps also via remote access at the NEPS Data Center at the University of Bamberg.

It is obvious that other surveys collecting life-course data would also profit from a tool like the TrueTales View. In fact, the tool has already been applied in other contexts. Unfortunately, the TrueTales View doesn't adapt automatically to all kinds of life-course data due to different data structures and diverging variable concepts. Data quality and precision is another crucial aspect that has to be assessed prior to adapting the TrueTales View.

References

- Aßmann, C., & Zinn, S. (2011): *Starting Cohort 6: Adults (SC6), SUF-Version 1.0.0, Data Manual [Supplement C]*. (Technical Report). Bamberg: University of Bamberg, National Educational Panel Study.
- Blossfeld, H.-P., Roßbach, H.-G., & von Maurice, J. (Eds.). (2011). Education as a life-long process: The German National Educational Panel Study (NEPS) [Special issue]. *Zeitschrift für Erziehungswissenschaft*, 14. Wiesbaden: VS Verlag für Sozialwissenschaften.

- Brzinsky-Fay, C., Kohler, U., & Luniak, M. (2006): Sequence Analysis Using Stata. *The Stata Journal*, 6(4), 435–460.
- Diewald, M., Goedicke, A., & Mayer, K. U. (2006): *After the fall of the wall: East German life courses in transition*. Stanford: Stanford University Press.
- Drasch, K., & Matthes, B. (2009): *Improving retrospective life course data by combing modularized self-reports and event history calendars* (IAB-Discussion Paper 21). Nuremberg: Institut für Arbeitsmarkt und Berufsforschung.
- Hillmert, S., Künster, R., Spengemann, P., & Mayer, K. U. (2004): Projekt Ausbildungs- und Berufsverläufe der Geburtskohorten 1964 und 1971 in Westdeutschland: Dokumentation. Teil 1–9. In *Materialien aus der Bildungsforschung* (Vol. 78). Berlin: Max-Planck-Institut für Bildungsforschung.
- Matthes, B., Drasch, K., Erhardt, K., Künster, R., & Valentin, M. A. (2012): *Arbeiten und Lernen im Wandel. Teil IV: Editionsbericht* (FDZ Methodenreport, No. 03/2012). Nürnberg: Institut für Arbeitsmarkt und Berufsforschung.
- Matthes, B., Reimer, M., & Künster, R. (2007): Techniken und Werkzeuge zur Unterstützung der Erinnerungsarbeit bei der computergestützten Erhebung retrospektiver Längsschnittdaten. *Methoden, Daten, Analysen. Zeitschrift für empirische Sozialforschung*, 1(1), 69–92.
- Matthes, B., Reimer, M., & Künster, R. (2005): *TrueTales: Ein neues Instrument zur Erhebung von Längsschnittdaten* (Arbeitsbericht 2 des Projekts “Frühe Karrieren und Familiengründung: Lebensverläufe der Geburtskohorte 1971 in Ost- und Westdeutschland”). Berlin: Max-Planck-Institut für Bildungsforschung.
- Mayer, K. U., & Solga, H. (2010): Lebensverläufe im deutsch-deutschen Vereinigungsprozess. In P. Krause, & I. Ostner (Eds.): *Leben in Ost- und Westdeutschland: Eine sozialwissenschaftliche Bilanz der deutschen Einheit 1990–2010* (pp. 39–56). Frankfurt am Main: Campus.
- Reimer, M., & Matthes, B. (2007): Collecting event histories with TrueTales: Techniques to improve autobiographical recall problems in standardized interviews. *Quality and Quantity. International Journal of Methodology*, 41(5), 711–735.
- Rohrbach-Schmidt, D. (2010): *BIBB Übergangsstudie 2006: Version 1.0* (BIBB-FDZ Daten- und Methodenberichte No. 1). Bonn: Bundesinstitut für Berufsbildung.
- Rusconi, A., & Solga, H. (Eds.). (2011). *Gemeinsam Karriere machen: Die Verflechtung von Berufskarrieren und Familie in Akademikerpartnerschaften*. Opladen: Verlag Barbara Budrich.

About the author

R. Künster
Social Science Research Center Berlin (WZB),
Reichpietschufer 50, 10785 Berlin, Germany.
e-mail: ralf.kuenster@wzb.eu