NEPS
National Educational Panel Study

BOOK OF ABSTRACTS

6th International NEPS Conference

June 7–8, 2021
online

LIfBi
LEIBNIZ INSTITUTE FOR
EDUCATIONAL TRAJECTORIES

Table of Contents

Call for Contributions for Special Issue

Large-Scale Assessments in Education

November 30, 2021 | LIfBi

Participants of the 6th International NEPS Conference are invited to submit a full paper of their presentations to a special issue in the journal Large-Scale Assessments in Education. The focus of the special issue coincides with the theme of the NEPS conference and seeks contributions on neglected problems and innovative solutions for the analysis of psychological constructs across heterogeneous contexts and life stages as evaluated in longitudinal studies. The planned special issue is not limited to the NEPS, but is also open for contributions from different longitudinal studies in education.

The deadline for full paper submissions is November 30, 2021.

The detailed call-for-papers is available at:

https://largescaleassessmentsineducation.springeropen.com/neps

The editors Christian Aßmann, Timo Gnambs, Marie-Ann Sengewald, Tanja Kutscher, and Claus H. Carstensen are looking forward to your contributions.

Preconference Workshop

Log Data Analysis: Introduction, indicator construction and strategies for analysis

June 7, 2021 | virtual | LIfBi

Zoom Link: https://zoom.us/j/99843912306

HAHNEL, CAROLIN
Leibnitz Institute for Research and Information in Education (DIPF)

Technology-based assessments can provide many advantages, for example, by allowing the collection of data on the course of item processing (so-called process data) in addition to the recording of "traditional" outcome data such as item responses. Such process data includes log data (i.e., events, event-related attributes, and timestamps) that are accumulated in log files during computer-based testing. While log data was still considered a "by-product" in the last decade, its strategic collection and deliberate analysis is steadily increasing. The possibilities of analyzing log data are manifold, but can be quite challenging due to the amount and structure of data, lack of conceptual considerations or insufficient documentation. The workshop will provide an introduction on how to work with log data.

Participants will gain insight into the preparation and analysis of log data using the log data from a previous study. They will construct and analyze log data indicators themselves in hands-on exercises using the R package LogFSM. Finally, different strategies of analysis (i.e., log data indicators as dependent or independent variable in regression analyses, methods of machine learning, latent variable models) are briefly reviewed in an overview.

Workshop language is English.

Schedule of presentations

June 8, 2021 | virtual | LIfBi

9:00-9:15	**Welcome and Opening of the NEPS Conference** (Christian Aßmann)			
9:15-10:00	**Suzanne Jak (University of Amsterdam, NL)** Opening lecture: Modeling cluster-level constructs with individual-level measures			
10:00-10:15	BREAK			
10:15-11:15	**Morning Sessions**			
	Symposium 1	Session 1	Session 2	Networking Platform
11:15-11:30	BREAK			
11:30-12:30	**Midday Sessions**			
	Session 3	Session 4	Session 5	Session 6
12:30-13:00	LUNCH BREAK			
13:00-13:45	**NEPS Publication awards and short presentations of the award winners** (Laudator: Christian Aßmann)			
13:45-14:00	BREAK			
14:00-15:30	**Matthias von Davier (Boston College, U.S.)** Keynote lecture: Longitudinal modeling with discrete and continuous latent variables			
15:30-16:00	BREAK			
16:00-17:30	**Afternoon Sessions**			
	Symposium 2	Session 7	Session 8	Networking Platform

Welcome and Opening of the NEPS Conference

Tuesday | June 8, 2021 | 09.00 am – 09.15 am

Zoom Link: https://zoom.us/j/92345923960

We welcome every participant of this year's NEPS Conference. This series of conferences is hosted by the Leibniz Institute for Educational Trajectories (LIfBi) in Bamberg. It brings together scientists from different disciplines and from different stages of their academic careers to discuss current projects and findings in educational research.

This year, the main focus of the sixth international NEPS Conference will be on the advancement of methodological practices in educational large-scale assessments. We will discuss neglected problems and innovative solutions for the analysis of psychological constructs across heterogeneous contexts and life stages.

The organizing team of this year's conference, Christian Aßmann, Timo Gnambs, and Marie-Ann Sengewald are looking forward to an appealing program on current trends in the modeling and analysis of complex data structures. We wish all of us an enjoyable time full of inspiring presentations and lively discussions.

NEPS Publication awards and short presentations of the award winners

Tuesday | June 8, 2021 | 13.00 am – 13.45 am

Zoom Link: https://zoom.us/j/92345923960

This award is given out for excellent scientific works based on data from the National Educational Panel Study (NEPS) by the Leibniz Institute for Educational Trajectories (LIfBi). An interdisciplinary scientific jury consisting of members of the LIfBi senior management and the nationwide NEPS Network selected the prize winners from all nominated publications.

We congratulate this year's winners for their outstanding work.

Opening lecture: Modeling cluster-level constructs with individual-level measures

Tuesday | June 8, 2021 | 09.15 am – 10.00 am

Zoom Link: https://zoom.us/j/92345923960

JAK, SUZANNE
University of Amsterdam, NL

Researchers frequently use the responses of individuals in clusters to measure constructs at the cluster level. For example, student's evaluations may be used to measure the teaching quality of instructors, patient reports may be used to evaluate social skills of therapists, and residents' ratings may be used to evaluate neighborhood safety. When multiple items are used to measure such cluster-level constructs, multilevel confirmatory factor models are useful. These models allow for the evaluation of the factor structure at the cluster level (modeling the (co)variances among item means across clusters), and at the individual level (modeling the (co)variances across individuals within clusters). If the cluster-level construct, for example teacher quality, would be perfectly measured using the responses of students, all students evaluating the same teacher would exactly the same item scores. In that case, there will not be any systematic variance in the item scores within clusters (only sampling error), so there will be nothing to model at the individual level. In practice, individuals do not all provide the same responses to the items, leading to systematic variance (and covariance) to be explained at the individual level. The question then arises how the variance within clusters should be modeled. In this talk, I will review some of the interpretational difficulties related to existing two-level models for cluster-level constructs and I will discuss possible alternative options.

Keynote lecture: Longitudinal modeling with discrete and continuous latent variables

Tuesday | June 8, 2021 | 14.00 pm – 15.30 pm

Zoom Link: https://zoom.us/j/92345923960

VON DAVIER, MATTHIAS
Boston College, U.S.

Measuring change in discrete latent variables is not the most common approach to the analysis of longitudinal data. However, there are recent examples of model developments that showcase why these types of models may be of interest. An early example is the latent transition model, which examines latent class membership over time. Another example is the Saltus model, an approach to capture discontinuous development that can be either applied to cross-sectional or, with appropriate extensions, to longitudinal data. Recent examples of more general approaches concerning multiple discrete latent variables can be found in recent literature on diagnostic classification models.

The presentation will compare and contrast discrete and continuous latent variable models for longitudinal approaches, discuss where predictions and model fit may differ, and will provide examples and illustrate fields of application.

Networking Platforms

Tuesday | June 8, 2021 | 10.15 pm – 11.15 pm

Tuesday | June 8, 2021 | 16.00 pm – 17.30 pm

Zoom Link: https://zoom.us/j/96398414200

It would have been a pleasure for us to welcome all of you in Bamberg. A private exchange of views, or just a place to meet and chat is often missing in an online setting. Thus, we set up a Networking Platform with the Breakout-Rooms "Little Venice", "Upper Bridge", "Bamberg Cathedral" - representing famous landmarks in the historic city.

The Networking Platform is open during the allocated times and during the breaks, offering the opportunity for having a coffee break together with other participants or for continuing discussions.

Bamberg Cathedral

Little Venice

Upper Bridge

Morning Sessions

Tuesday | June 8, 2021 | 10.15 am – 11.15 am

Symposium 1

Chair: Tobias Deribo, Support: Jana Welling

Studying the missing data mechanisms behind filtering rapid guessing in cognitive assessments
(Deribo, Kröhne & Goldhammer)

Validating ability-related time components in reading tasks with unit structure
(Engelhardt, Kröhne, Hahnel, Deribo & Goldhammer)

Mixture IRT identification of rapid responding using response times in questionnaires
(Kroehne, Buchholz & Goldhammer)

Session 1

Chair: Eren Özberk, Support: Anna Scharl

Modeling the dynamics between maths and reading skills with continuous-time models
(Jindra, Sachse & Hecht)

Applying continuous-time modeling to PISA data: An illustration
(Lohmann, Zitzmann & Hecht)

Using subgroup discovery and latent growth curve modeling to identify unusual educational trajectories
(Kiefer, Langenberg, Lemmerich & Mayer)

Session 2

Chair: Anika Bela, Support: Eva Zink

Company? Yes please! Using the NEPS to analyse why apprentices with Abitur decline university
(Preböck & Annen)

The effect of teacher characteristics on students’ science achievement
(Sancassani)

How stable is the relationship between education and class in Germany? Empirical distributions, counterfactual worlds, and a configurational analysis of NEPS data
(Glaesser)

Studying the missing data mechanisms behind filtering rapid guessing in cognitive assessments

Tuesday | June 8, 2021 | 10.15 am – 11.15 am

Zoom Link: https://zoom.us/j/92345923960

DERIBO, TOBIAS
Leibniz Institute for Research and Information in Education

KRÖHNE, ULF
Leibniz Institute for Research and Information in Education

GOLDHAMMER, FRANK
Leibniz Institute for Research and Information in Education & Centre for International Student Assessment

The increased availability of time-related information as a result of computer-based assessment has enabled new ways to measure test-taking engagement (e.g., Molenaar, 2015). One of these ways is the distinction of solution and rapid guessing behavior (Schnipke and Scrams, 1997). Prior research has recommended response-level effort filtering to deal with rapid guessing behavior (Rios et al., 2017). As response-level effort filtering introduces missing data, this can lead to biased parameters if rapid guessing depends on the measured trait or (un-)observed covariates (e.g., Holman & Glas, 2005). Therefore, we utilized a model based on Mislevy and Wu (MW-Model, 1996) to investigate the assumption of ignorable missing data underlying response-level effort filtering. The model allows us to investigate and compare different approaches to treat rapid guessing in a single framework through parameterization. We applied the model to a subsample of test-takers from the German longitudinal National Educational Panel Study (Blossfeld, Roßbach & von Maurice, 2011) who participated in an unproctored online assessment in the year 2013 (N = 4960). Here we focus specifically on an assessment of Information and Communication Technology literacy (Senkbeil et al., 2013) and figural reasoning (Lang et al., 2014). We found an unconstrained MW-Model to show better model fit then possible alternatives. This may indicate that, at least for the studied assessments, the emergence of rapid guessing behavior depends on test-takers rapid guessing propensity and true responses on items they have rapid guessed on. The finding implies a violation of the assumption of ignorable missing data underlying response-level effort filtering. Further, ability estimates appeared sensitive to different treatment approaches for rapid guessing and model-based approaches exhibited higher convergent validity evidence compared to other treatments. The results illustrate the importance of studying underlying assumptions and sensitivity of estimates before deciding on and applying a specific treatment to rapid guessing.

Validating ability-related time components in reading tasks with unit structure

Tuesday | June 8, 2021 | 10.15 am – 11.15 am

Zoom Link: https://zoom.us/j/92345923960

ENGELHARDT, LENA
Leibniz Institute for Research and Information in Education

KRÖHNE, ULF
Leibniz Institute for Research and Information in Education

HAHNEL, CAROLIN
Leibnitz Institute for Research and Information in Education

DERIBO, TOBIAS
Leibniz Institute for Research and Information in Education

GOLDHAMMER, FRANK
Leibniz Institute for Research and Information in Education & Centre for International Student Assessment

The relation of processing times and item scores varies according to ability and is less positive/more negative for more able respondents (see time on task effect; Goldhammer et al., 2014). Theoretically, this can be explained by a difference in the extent to which different able persons use automated and controlled processes (dual-processing theory; Schneider & Shiffrin, 1977). Hence, persons using automated processes to greater extent can correctly solve tasks in less time. The goal of the present study is to identify time indicators in unit-structured tasks which show this time on task effect. Time indicators affected in first line by a different use of automated and controlled processing, such as the time for reading a text (Hypothesis 1a), should show the time on task effect; but not time indicators which depend predominantly on motivational or strategic aspects, such as revisiting the text (Hypothesis 1b). *N* = 506 test-takers worked on five reading units (Gehrer et al., 2013). Latent variables for time indicators and ability were modeled in Mplus (Muthén & Muthén, 2015) and their relations analyzed using multigroup analyses (low, middle, high ability). As expected, the relation of ability and the longest time period on the text page (Hypothesis 1a) was less positive for more able readers (low: *β* = 0.706***, medium: *β* = 0.359n.s., high: *β* = 0.265n.s.). As expected, the time used for text revisits was not less positive for more able readers (Hypothesis 1b: low: *β* = 0.214n.s., medium: *β* = 0.469n.s., high: *β* = 0.805***). Results support that differences in the extent of automated and controlled processing become visible in the longest time period on the text page, but not in the time used for text revisits. Still, investing time for revisiting the text seemed to be an effective strategy for test-takers with high reading ability.

Mixture IRT identification of rapid responding using response times in questionnaires

Tuesday | June 8, 2021 | 10.15 am – 11.15 am

Zoom Link: https://zoom.us/j/92345923960

KRÖHNE, ULF
Leibniz Institute for Research and Information in Education

BUCHHOLZ, JANINE
Leibniz Institute for Research and Information in Education

GOLDHAMMER, FRANK
Leibniz Institute for Research and Information in Education & Centre for International Student Assessment

For surveys or questionnaires administered as item batteries, Kroehne and Goldhammer (2018) illustrate how item-level response times can be extracted from log-file data. If the resulting response time distributions are bimodal, rapid responding can be identified within questionnaires using time thresholds (Kroehne et al., 2019), similar to rapid guessing based on response times in cognitive tests (e.g., Wise et al., 2006). Rapid responding was found to be related to ability (Kroehne et al., 2019) and response style (Kroehne et al., 2020). However, the identification of rapid responding is laborious and requires visual inspection of response time distributions to define thresholds. Moreover, threshold identification for questionnaires is less efficient than for cognitive tests since the probability to solve an item is not available for questionnaire items (Soland et al., 2019). Therefore, this paper presents a new approach using mixture IRT models for trait and speed simultaneously that allows identifying disengaged rapid responders for item batteries without visual inspection of the response time distributions. Using data from the PISA 2015 context questionnaire from Switzerland, we illustrate how baseline models that allow either a linear or a curvilinear relationship of speed and trait can be compared using information criteria. Assuming that rapid responding is characterized by the absence of a (latent) relationship between speed and trait, mixture IRT models are used to identify classes of students with rapid responding. The models can be applied for selected scales separately or to pairs of adjacent Likert-type questionnaire scales. To validate the method the identified responses will be compared to rapid responses identified using time thresholds derived with visual inspection. The method based on response times extracted with LogFSM (Kroehne, 2021) can be applied to large-scale data since the mixture IRT model can be implemented using standard software (such as Mplus, Muthén & Muthén, 2017).

Modeling the dynamics between maths and reading skills with continuous-time models

Tuesday | June 8, 2021 | 10.15 am – 11.15 am

Zoom Link: https://zoom.us/j/94315936213

Unobserved stable between-subject differences and reversed causality are major threats to causal inference. Approaches like cross-lagged models with correlated random intercepts or dynamic panel models with fixed effects offer some protection but fall short of providing a model of the continuous time process that might underlie the relationship between the constructs of interest over time. As a consequence, studies usually report cross-lagged effects for one specific time interval, instead of estimating the underlying continuous-time process which can be used to calculate cross-lagged effects for any desired time interval length. Continuous time models were developed for this type of task and also can account for unobserved stable between-subject differences. They are thus well suited for many of the challenges in the analysis of longitudinal data. We demonstrate the potential of this approach for educational research by analysing the relationship between math and reading skills over time in the starting cohort 3 of the National Educational Panel Study. Our results show that students' mathematics and reading proficiencies are temporally related with maximal effects for a time interval of around 10 months, for which both (standardized) cross-lagged effects were positive and significantly different from zero. Our findings also strengthen previous studies, showing that a large share of the relationship between the two constructs is due to unobserved time-constant heterogeneity.

JINDRA, CHRISTOPH
Humboldt University Berlin

SACHSE, KAROLINE A.
Humboldt University Berlin

HECHT, MARTIN
University of Tübingen

Applying continuous-time modeling to PISA data: An illustration

Tuesday | June 8, 2021 | 10.15 am – 11.15 am

Zoom Link: https://zoom.us/j/94315936213

Longitudinal methods are an important tool for addressing a variety of research questions in the education sciences and psychology. Well-known examples are the development of cognitive abilities, the long-term effects of interventions and educational programs, and the permanent empirical monitoring of education systems via regular representative large-scale assessments. In the context of longitudinal studies, a prerequisite to ensure the validity of conclusions is the usage of appropriate statistical methods of longitudinal data analysis. Many common methods either do not take the dynamics of developmental processes into account (static models) or require equally spaced measurement occasions (discrete time dynamic models). The latter comes at the expense of flexible designs and cross-study comparisons, which are essential for the accumulation of knowledge in science. One approach that aims to overcome these limitations is continuous-time modeling. Continuous-time models use stochastic differential equations to estimate the underlying continuous processes and therefore also allow to investigate the dynamical development of relevant variables, such as the development of student achievement. The objective of the present study is to outline the features and potentials of applying continuous-time models to data from large-scale assessments. Specifically, we illustrate the application of continuous-time modeling using the PISA reading literacy scores of N = 59 countries between 2000 and 2018 (T = 7). Moreover, we demonstrate the predictive accuracy of a continuous-time model and compare it with a standard growth curve model. Overall, our results suggest advantages for the continuous-time approach but also point out some potential pitfalls and shortcomings of this type of model in the longitudinal analysis of large-scale assessments.

LOHMANN, JULIAN F.
Hector Research Institute of Education Sciences and Psychology, University of Tübingen

ZITZMANN, STEFFEN
Hector Research Institute of Education Sciences and Psychology, University of Tübingen

HECHT, MARTIN
Hector Research Institute of Education Sciences and Psychology, University of Tübingen

Using subgroup discovery and latent growth curve modeling to identify unusual educational trajectories

Tuesday | June 8, 2021 | 10.15 am – 11.15 am

Zoom Link: https://zoom.us/j/94315936213

Latent growth curve models are frequently used statistical models to analyze (educational) longitudinal data and to gain insights into interindividual differences in growth trajectories. A key research question in this context is whether certain subgroups with unusual developmental trajectories exist. In this presentation, we introduce a new framework that integrates the modeling of growth curves with state-of-the art pattern mining techniques, i.e., subgroup discovery and exceptional model mining, as an alternative to previous approaches such as growth mixture modeling (GMM) and structural equation model trees (SEM Trees). This allows for flexibly and efficiently finding the interpretable subgroups of interest in large datasets, where subgroups can be potentially determined by a multitude of covariates. We illustrate this approach in an empirical example using data from the National Educational Panel Study (NEPS): We investigated trajectories of university students' dropout intention during four years of studies and identified several subgroups that exhibit exceptional trajectories of dropout intention compared to the complement of the sample. In this example, subgroups were formed with respect to gender, age, desired final qualification, and educational background of one’s parents. This exploratory data mining approach can help educational researchers to better understand heterogeneity in growth trajectories using data from large-scale assessments and to generate hypotheses for future research.

KIEFER, CHRISTOPH
Bielefeld University

LANGENBERG, BENEDIKT G.
Bielefeld University

LEMMERICH, FLORIAN
University of Passau

MAYER, AXEL
Bielefeld University

Company? Yes please! Using the NEPS to analyse why apprentices with Abitur decline university

Tuesday | June 8, 2021 | 10.15 am – 11.15 am

Zoom Link: https://zoom.us/j/99664826211

PREBÖCK, TANJA
Otto Friedrich University, Bamberg

ANNEN, SILVIA
Otto Friedrich University, Bamberg

Young adults in the German educational system who choose vocational training after their Abitur have the opportunity to pursue a university degree afterwards. Several studies, such as Aynsley & Crossouard (2010), Jacob & Solga (2015) and Pilz et al. (2020) examine the aspects that affect this career choice. These studies are often limited to regional analyses or focussed on specific occupations or groups of apprentices. We aim to broaden the knowledge about this educational choice and the socio-structural embeddedness by using data from the German National Educational Panel Study (NEPS). We use the SC4-11.0.0 from NEPS to extend the recent regional and occupation specific findings of Pilz et al. (2020), comparing Abitur holders at the end of their vocational training who choose university versus those who stay in the labour market. Here, we focus on their sociodemographic background. We applied various statistical tests to analyse whether the two groups differ significantly regarding sociodemographic characteristics and/or framework conditions. Similar to the results of Pilz et al. (2020), our results suggest that gender represents a significant background variable with regard to choosing university following vocational training. Additionally, we replicated results showing both groups do not differ significantly regarding their number of siblings or migratory backgrounds. Contradictory, significant differences also exist regarding age, grade during the training and the parental educational degree. Additionally, a significant correlation between participation in cultural events with the ambition towards university was discovered, which was not revealed by Pilz et al. (2020). We intend to use NEPS in follow-on-research as its longitudinal structure enables us to reconstruct individual’s educational paths, which analytically allows us to separate within from in-between subject effects. Furthermore, we aim to focus on individual career paths and associated fields of occupation.

The effect of teacher characteristics on students' science achievement

Tuesday | June 8, 2021 | 10.15 am – 11.15 am

Zoom Link: https://zoom.us/j/99664826211

SANCASSANI,
PIETRO
Ifo Institute, Munich

Using data from TIMSS 2015, an international large-scale assessment of student skills, I investigate the effect of teacher characteristics on students' science achievement. My identification strategy exploits the feature that in many education systems different science domains (physics, biology, chemistry, and earth science) are taught by different teachers. The availability of students' test scores as well as teachers' questionnaires for each of these domains allows me to implement a within-student approach which controls for unobserved student heterogeneity. I find a positive and significant effect of teacher specialization in the specific science domain on students' results, equivalent to 1.7% of a standard deviation. Holding a Master's degree, pedagogical preparation and teaching experience have no significant effect. Teachers' experience has a negative impact on the extent to which students like to study a subject or find teaching engaging. To the best of my knowledge, this is the first study that exploits the within-student across-subject variation using subjects that belong to the same field. Furthermore, I focus on science achievement, a subject which has been less investigated in the related literature, thus enriching the existing literature on the teacher quality.

How stable is the relationship between education and class in Germany? Empirical distributions, counterfactual worlds, and a configurational analysis of NEPS data

Tuesday | June 8, 2021 | 10.15 am – 11.15 am

Zoom Link: https://zoom.us/j/99664826211

GLAESSER, JUDITH
University of Tübingen

In earlier work using British data a co-author and I employed an innovative approach to illuminate the well-documented relationship between educational qualifications and achieved social class positions. This approach contrasts the empirical evidence with an idealised counterfactual “meritocratic” model (allocation to available positions by qualifications alone based on the empirical marginal distributions), followed by analysis of the sufficient and necessary conditions for social class outcomes using Ragin’s configurational method, Qualitative Comparative Analysis (QCA), which employs Boolean algebra and truth tables. Here I use the same approach with NEPS data. Qualifications are, from one perspective, positional goods. Their value, ceteris paribus, changes with their scarcity. The meritocratic model allows me to specify, counterfactually, what qualifications would have represented necessary and sufficient conditions in the modelled meritocracy for reaching a salariat class position as the distributions of qualifications and of class positions changes over time. Thus, I created a meritocratic model for two NEPS cohorts (1944-1950 and 1976-1980) to explore changes in the potential value of qualifications over one generation. My comparisons showed that, while there were changes in the distributions of qualifications and class positions, a degree would have been a sufficient, but not necessary, condition for entering the salariat in both cohorts’ counterfactual worlds. In the empirical worlds, relations of sufficiency and necessity have the same general tendency as in the counterfactual worlds, but, as expected, they are not identical. Possible explanations for the deviations from the counterfactual worlds are explored in the second step which employs multi-value QCA, with highest post-school qualification, gender, and parental class as potential sufficient/necessary conditions for social class outcomes. The findings suggest a change in the role of gender as well as a large degree of stability for qualifications and parental class, highlighting the complex interplay of individual and contextual factors in class allocation.

Midday Sessions

Tuesday | June 8, 2021 | 11.30 am – 12.30 pm

Session 3

Chair: Timo Gnambs, Support: Jana Welling

Comparing original and marginal trend estimates in large-scale assessment studies: Analytical derivations and a simulation study
(Robitzsch & Lüdtke)

Analyzing PISA Mathematics 2000-2012: Evaluating the effects of the choice of calibration samples, item samples, estimation method, and linking method
(Heine & Robitzsch)

The impact of measurement error for causal effect analysis: An illustration with NEPS data
(Sengewald, Mayer)

Session 4

Chair: Ariane Würbach, Support: Anna Scharl

Challenges of representativeness in survey research: An evaluation of the ERiK Surveys 2020
(Schacht, Gedon & Gilg)

Adjusting to the survey: How within-survey interviewer experience relates to interview duration
(Pirralha, Haag, & von Maurice)

SES and gender bias in teacher assessments and their consequences for achievement inequality
(Olczyk & Schneider)

Session 5

Chair: Christoph Homuth, Support: Eva Zink

Labour market returns of ICT competences - status quo and methodological perspectives
(Thürer & Annen)

Hidden Figures – Profiles and potentials of returns to education in working German adults
(Reinwald & Annen)

A distributional analysis of gender gaps in wages and numeracy skills
(Battisti, Fedorets, & Kinne)

Midday Sessions

Tuesday | June 8, 2021 | 11.30 am – 12.30 pm

Session 6

Chair: Claus H. Carstensen, Support: Sören Reimers

Increasing test efficiency of the teacher knowledge survey module of TALIS 2024 through multidimensional adaptive testing
(Fink & Frey)

The performance of item selection algorithms in Mokken scaling for various sample size: An empirical example with intelligence scale
(Özberk)

CANCELLED: Distributed Leadership and Alignment Optimization: A comparative, cross-cultural perspective across 40 countries
(Eryilmaz)

Comparing original and marginal trend estimates in large-scale assessment studies: Analytical derivations and a simulation study

Tuesday | June 8, 2021 | 11.30 am – 12.30 pm

Zoom Link: https://zoom.us/j/92345923960

One major aim of international large-scale assessments (ILSA) like PISA is to monitor changes in student performance over time. To accomplish this task, a set of common items is repeatedly administered in each assessment. Linking methods based on item response theory (IRT) models are used to align the results from the different assessments on a common scale. The one-parameter logistic (1PL) and the two-parameter logistic (2PL) IRT models are employed as scaling models for dichotomous item response data in this work. The present article discusses two types of trend estimates for countries in ILSA. First, the original trend estimate for a country is defined by the achievement difference between the country means in the two assessments. Original trend estimation relies on item parameters that are obtained from data involving all countries. Second, the marginal trend estimate for a country is the achievement difference in two assessments that solely relies on country-specific item parameters. It is shown in an analytical derivation and through a simulation study that the performance of the two trend estimators depends on the extent of violations of measurement invariance (country-by-item interaction, assessment-by-item interaction, and country-by-assessment-by-item interaction) and the existence of unique items (i.e., items that are only administered in one assessment). Marginal trend estimates outperform original trend estimates in many conditions for the 1PL and the 2PL model.

ROBITZSCH, ALEXANDER
Leibniz Institute for Science and Mathematics Education, Kiel & Centre for International Student Assessment, Kiel

LÜDTKE, OLIVER
Leibniz Institute for Science and Mathematics Education, Kiel & Centre for International Student Assessment, Kiel

Analyzing PISA Mathematics 2000-2012: Evaluating the effects of the choice of calibration samples, item samples, estimation method, and linking method

Tuesday | June 8, 2021 | 11.30 am – 12.30 pm

Zoom Link: https://zoom.us/j/92345923960

HEINE, JÖRG-HENRIK
Technical University of Munich

ROBITZSCH, ALEXANDER
Leibniz Institute for Science and Mathematics Education, Kiel & Centre for International Student Assessment, Kiel

The international database of five PISA survey rounds is rescaled by concurrent item calibration. The analyses refer to the core domain mathematical literacy, which covers 179 items, where 8 and 35 trend items were administered in all rounds from 2000 to 2012 and 2003 to 2012, respectively. The data comprises responses of about 1:7 million students from 79 countries. Although PISA cannot be termed a panel or longitudinal study, the trend view comes more to the foreground in the reporting. Such a view is justified by the although cross-sectional, but representative sampling of fifteen-year-old students at the respective survey period and the relative continuity in terms of the type of data collected. Thus, PISA scores are typically viewed as comparable and valid trend indicators for educational systems that use a common metric to quantify student performance. The overarching question is to what extent analytic choices may influence the inference about two central outcome measures country ranking and development trends between and within countries. In particular, four key methodological factors are considered: (1) The selection of country sub-samples for item calibration differing at three factor levels. (2) The item sample, which refers either on the two sets of trend items or covers all items used within PISA. (3) The method of item calibration, using either the likelihood marginal maximization method as implemented in TAM (Robitzsch, Kiefer, & Wu, 2020) or an pairwise row averaging approach according to Choppin (1968) as implemented in pairwise (Heine, 2021). (4) The type of linking method, which is either concurrent calibration or separate calibration with successive chain linking. In total, 3 x 3 x 2 x 2 = 36 analyses are carried out. We expect the item selection and choice of linking method to have a greater effect on country ranking and trend estimates than the estimation method and sample selection used for item calibration.

The impact of measurement error for causal effect analysis: An illustration with NEPS data

Tuesday | June 8, 2021 | 11.30 am – 12.30 pm

Zoom Link: https://zoom.us/j/92345923960

Well-constructed achievement tests are a special strength of educational large-scale assessments (LSAs) and the students' proficiencies are frequently of main interest in subsequent analysis - like for the comparison of respondent groups. The estimation of covariate adjusted group differences is common practice for causal effect analysis in non-randomized LSA data. Yet, in standard procedures for covariate adjustment (i.e., analysis of covariance or propensity score methods) often fallible test scores, instead of the latent proficiencies itself, are used for modeling the outcome and relevant covariates. We illustrate the impact of measurement error in this setting, using an example from the National Education Panel Study (NEPS). Specifically, we investigate the effect of tutoring in mathematics on a subsequent math ability measure, while controlling for the previous math ability and additional covariates. The application is well-suited for illustrating theoretical conditions under which measurement error in test scores biases treatment effects. In addition, the NEPS provides item-level data, based on which a direct integration of latent variables for causal effect estimation is possible. For this, we provide an extension of the R package EffectLiteR that facilitates causal effect analysis based on structural equation models for categorical indicators. Effect estimates from an analysis with latent math proficiencies will be compared to effect estimates when using fallible tests scores either as a relevant covariate or the outcome. The impact of measurement error in practice will be discussed in relation to the theoretical conditions.

SENGEWALD, MARIE-ANN
Leibniz Institute for Educational Trajectories (LIfBi), Bamberg &
Otto Friedrich University, Bamberg

MAYER, AXEL
Bielefeld University

Challenges of representativeness in survey research: An evaluation of the ERiK Surveys 2020

Tuesday | June 8, 2021 | 11.30 am – 12.30 pm

Zoom Link: https://zoom.us/j/94315936213

At what point can we claim that survey data and results are representative of the entire population? And what actions can researchers take to improve representativeness? Despite the fact that these questions are fundamental to almost any research endeavour, they are rarely explicitly addressed. This presentation aims to address this gap and showcases the challenges of generalisability faced by many researchers by discussing the approach taken by the national study "Indicatorbased monitoring of structural quality in the German early childhood education and care system (in German: Entwicklung von Rahmenbedingungen in der Kindertagesbetreuung – indikatorengestützte Qualitätsbeobachtung; ERiK for short)". We first describe the ERiK survey concept, which consists of five cross-sectional surveys in 2020, covering the multiple stakeholder perspectives of directors and pedagogical staff in day-care facilities, family day-care workers, youth welfare offices and day-care facility providers. We then evaluate the quality of the ERiK data collection with regard to their representativeness, especially for the 16 German federal states, focusing in particular on selectivity due to varying sampling and participation probabilities. We use several different measurements for this assessment, including comparisons between actual and ideal sample size, the share of respondents of the total population and the response rate. Additionally, we discuss other possible sources of error at different stages of the survey process (summarised in the concept of the Total Survey Error) and develop appropriate weighting factors. We conclude that the ERiK surveys 2020 can be considered representative and therefore can be used to make generalised statements about the quality of child day-care in Germany. Finally, we review some limitations of the datasets.

SCHACHT, DIANA D.
Deutsches Jugendinstitut (DJI), Munich

GEDON, BENJAMIN
Deutsches Jugendinstitut (DJI), Munich

GILG, JAKOB J.
Deutsches Jugendinstitut (DJI), Munich

Adjusting to the survey: How within-survey interviewer experience relates to interview duration

Tuesday | June 8, 2021 | 11.30 am – 12.30 pm

Zoom Link: https://zoom.us/j/94315936213

It has been well established that interviewers are important actors in Computer Assisted Telephone Interviews (CATI). They are important because they are responsible for setting the interview pace or influence how much of a cognitive effort the respondent dedicates to answering. A significant finding regarding CATI interviewers is that not only are they heterogeneous regarding the time they spend in administering an interview but also that this time tends to shorten over the course of the fieldwork. There are several hypotheses discussed in the literature to explain this tendency. In particular, it is often argued that interviewers show a learning effect and optimize survey administration as they gain within-survey experience. This presentation examines the relationship between general survey experience, within survey experience and interview length using data from wave 1 of the parents CATI interviews from the National Educational Panel Study (NEPS), Starting Cohort Grade 9 (doi:10.5157/NEPS:SC4:11.0.0). We employ multilevel models that show considerable influence of the interviewers on the interview duration and find that interview duration decreases as the within-survey experience increases. This effect is robust even after controlling for various respondent, interviewer, and interview characteristics. We also find that, contrary to our expectations, higher item nonresponse rates are actually associated with longer interview duration. Even though other data collection methods are gaining in importance, interviewer administered surveys are still a primary method of collecting information in large-scale educational assessments. This presentation contributes to the methods literature by shedding light on the role of the interviewer experience and its effects on data quality. Results are discussed with a focus on interviewer training and supervision.

PIRRALHA, ANDRÉ
Leibniz Institute for Educational Trajectories (LIfBi), Bamberg

HAAG, CHRISTIAN
Leibniz Institute for Educational Trajectories (LIfBi), Bamberg

VON MAURICE, JUTTA
Leibniz Institute for Educational Trajectories (LIfBi), Bamberg

SES and gender bias in teacher assessments and their consequences for achievement inequality

Tuesday | June 8, 2021 | 11.30 am – 12.30 pm

Zoom Link: https://zoom.us/j/94315936213

OLCZYK, MELANIE
University of Leipzig

SCHNEIDER, THORSTEN
University of Leipzig

Various dimensions of educational success, such as student achievement, vary systematically by parental socio-economic status (SES) or gender. Discrimination by teachers may account for at least some of the observed inequalities. Research indicates, for instance, that teacher stereotypes can initiate bias in teacher judgments along with students' SES or gender. Differential teacher assessments and expectations can in turn be mediated through differential verbal and non-verbal behaviors and result in a self-fulfilling prophecy. Such a process can exacerbate intergroup inequality in school achievement. Based on NEPS Starting Cohort 2 data, we, first, investigate whether there are SES and gender bias in teacher assessments of students` language and mathematical skills at school entry. Second, we examine the effects of (potentially biased) teacher assessments measured in grade 1 on student achievement in grade 4. For research question 1, we regress teacher assessment on students' results from different tests and parents' report on school related behavior to reduce the risk of measurement error as well as omitted variables. The residuals of these regressions are used to identify biased assessment by SES and gender. For research question 2, value added models in a multi-level framework (students nested in classes) are estimated, whereas we do not only use information on test results from grade 1 as a predictor but also the residuals from the previous regression. The results are discussed against the background of the advantages and disadvantages of the methodological approach as well as by considering challenges like the change of teachers over time.

Labour market returns of ICT competences - status quo and methodological perspectives

Tuesday | June 8, 2021 | 11.30 am – 12.30 pm

Zoom Link: https://zoom.us/j/99664826211

THÜRER, SEBASTIAN
Otto Friedrich University, Bamberg

ANNEN, SILVIA
Otto Friedrich University, Bamberg

Information technologies are gaining importance in various occupational fields requiring an increased need for Information and Communication Technology (ICT) competencies (Kiener et al. 2019). Previous studies (e.g. Greenwood et al. 2011; Falck et al. 2016) confirm that ICT competences are relevant for individual labour market outcomes and provide some empirical evidence on how the labour market rewards ICT competences. Furthermore, STEM (science, technology, engineering, and mathematics) occupations are associated with macroeconomic benefits (e.g. Croak 2018). This presentation aims to contribute insights focusing on the wage effects of ICT competences and its interrelatedness with other wage determining variables (gender, age, migration status, education, occupation) by using data from the "Starting Cohort 6" of the German National Educational Panel Study (*N*=2,808). Using the dependent variable of log gross wages and the independent variable of individual ICT competence values, linear regression models included controls for gender, age and a quadratic in age, migration status, education (academic versus non-academic), and occupation (STEM versus Non-STEM). The regression models show that ICT competences and working in a STEM-occupation have a significant positive wage effect. Furthermore, our calculations reveal positive effects of age (B=0.60, p=0.00) and academic qualification level (B=0.20, p=0.00). Being female (B=-0.34, p=0.00) as well as having a migration background (B=-0.01, p=0.39) negatively affect individuals' wages. To test the interrelation of ICT competences with the above group variables, we calculated respective interaction models. Here, results show that being female together with high ICT competences has a weak positive effect (B=0.07, p=0.00) on wages. In further research, we aim to compare our models to Unconditional Quantile Regression models (Firpo et al. 2009), which would allow the estimation of the effect of a variable on individuals' positions in the wage distribution.

Hidden Figures – Profiles and potentials of returns to education in working German adults

Tuesday | June 8, 2021 | 11.30 am – 12.30 pm

Zoom Link: https://zoom.us/j/99664826211

Through profiles of adult learners, it is possible to establish a priority list for programs and investments which support learners that yield high value for society. We define potential as the difference in created value between similar individuals that either received or did not receive significant further education and training. The literature review exposes a set of variables that are either highly variant in their input, produce highly heterogeneous outputs or stay inconclusive when used separately (e.g. Hansson (2008), Ehlert (2017)). This suggests that we have to look at the value out of further education and training through the lens of case-clusters conducive to revealing combined influence. In line with the current findings, we retain the following cluster variables: gender, prior education, job mobility, wage (developments/levels). In a person-centred approach, we first perform k-medoids cluster analysis - combined with silhouette scores and in contrast to factor analyses - to identify smaller groups of cases with similar characteristics. Second, we rank the clusters according to their value-creating-potential. Finally, we create archetype profiles for the retained clusters to render them operable for future research and policy. We aim to find an optimal number of clusters for framing questions about continuing education. The clusters can be ranked according to their educational potential, and thus help to identify high and low potential individuals. Cluster membership correlates clearly with participation in further education and training and with common measures of returns to education. There is an optimal set of variables to explain educational potential. The demographic profiles of cluster centroids are useful for qualitative follow-up studies as well as decision-making in politics and corporations.

REINWALD, SIMON
Otto Friedrich University, Bamberg

ANNEN, SILVIA
Otto Friedrich University, Bamberg

A distributional analysis of gender gaps in wages and numeracy skills

Tuesday | June 8, 2021 | 11.30 am – 12.30 pm

Zoom Link: https://zoom.us/j/99664826211

BATTISTI, MICHELE
University of Glasgow

FEDORETS, ALEXANDRA
DIW, Berlin

KINNE, LAVINIA
Ifo Institute, Munich

This paper investigates gender-specific patterns in the distribution of numeracy skills among adults, measured by standardized tests in the international PIAAC survey. We report three patterns: first, we document that both wages and returns to skills are higher for individuals with higher numeracy skills. Second, we show that women with the highest numeracy levels experience lower returns to skills compared to men with the same numeracy levels. Lastly, we find that women in general have lower numeracy skills than men, especially at the top of the numeracy distribution. In order to shed light on potential determinants, we perform a decomposition of gender gaps across the numeracy distribution and find that especially children and fields of study play an important role. These findings contribute to the discussion about glass ceilings for women in the labor market and emphasize the role of potential labor-market discrimination of women. This study uses a variety of relevant methods to explore in detail the multiple facets of an international skill dataset such as PIAAC. Research in education-related labor economics often only uses mean comparisons when looking at skills or wages. Although meaningful in themselves, these comparisons can only provide a very broad picture of the dynamics behind observed associations. Instead, quantile regressions provide a much more detailed insight to the determinants and consequences of skill measures at different points of the skill distribution and can hence inform the policy debate in a much more meaningful way. Furthermore, combining these quantile analyses with Oaxaca-Blinder decomposition techniques, presents not only a relatively new approach in empirical economics, but also fully exploits the size and richness of the present dataset in order to determine potential channels leading to gender gaps in numeracy skills.

Increasing test efficiency of the teacher knowledge survey module of TALIS 2024 through multidimensional adaptive testing

Tuesday | June 8, 2021 | 11.30 am – 12.30 pm

Zoom Link: https://zoom.us/j/96398414200

The TKS assessment module will form an optional module for the next cycle of the Teaching and Learning International Survey (TALIS) in 2024. The TKS assesses general pedagogical knowledge on the three key dimensions *instructional processes*, *learning processes*, and *assessment*. The goals for the TKS assessment module are very high: While reducing the testing time by 50%, the reliability, which did not meet common reporting standards in the pilot study, has to be increased substantially. We will present possibilities to increase the measurement efficiency of the TKS assessment module by using multidimensional adaptive testing (MAT) and describe how a MAT design for the TKS assessment module is configured best. The potential of the proposed MAT design is illustrated by a Monte Carlo simulation study comparing MAT with a non-adaptive multi-unidimensional and a non-adaptive multidimensional test. Evaluation criteria are test information and reliability. The results demonstrate that MAT increases the measurement precision up to a range that is well suited for precise result reporting, while this is not the case for the two non-adaptive conditions. The talk ends with concrete and directly applicable recommendations for the MAT design, with which the TKS assessment in particular, but also other educational large-scale assessments can be transformed into innovative and highly efficient measurement instruments by using computers in the best possible way.

FINK, ARON
Goethe University Frankfurt

FREY, ANDREAS
Goethe University Frankfurt & Centre for Educational Measurement (CEMO) & University of Oslo

The performance of item selection algorithms in Mokken scaling for various sample size: An empirical example with intelligence scale

Tuesday | June 8, 2021 | 11.30 am – 12.30 pm

Zoom Link: https://zoom.us/j/96398414200

Automated item selection procedure (AISP) and a genetic algorithm (GA) are used to select the Mokken scale containing as many items as possible. However, the application of item selection algorithms to data can be challenging. Up to now, there is very limited knowledge about the sample size required for a Mokken scaling due to the lack of information on the distribution properties of Mokken scales has made sample size estimation difficult. Recent studies provided minimum sample size requirements for Mokken scale analysis for polytomous scored items and short scales (N=20). The present study examined the sample size requirements for 48 dichotomously scored items. Moreover, the study also investigated the performance of invariant item ordering for various sample sizes on real data set. Per element accuracy (PEA) index evaluates the impact of sample size around scale, item, and item pair scalability coefficients. The results showed that GA generally performed well in most conditions, followed by the AISP, showing a slightly lower accuracy rate. In contrast, the peculiarities slightly overestimated the item scalabilities for AISP*.*

ÖZBERK, EREN HALIL
Leibniz Institute for Educational Trajectories (LIfBi), Bamberg

CANCELLED: Distributed Leadership and Alignment Optimization: A comparative, cross-cultural perspective across 40 countries

Tuesday | June 8, 2021 | 11.30 am – 12.30 pm

Zoom Link: https://zoom.us/j/96398414200

ERYILMAZ,
NURULLAH
University of Bath

Distributed leadership (DL) is an increasingly used concept among researchers, policymakers, and educationalists worldwide. However, a limited study in the literature compares principal distributed leadership scale cross-culturally and posits to rank countries regarding their principal distributed leadership properties. The purpose of this study employs an alignment optimization approach to compare the latent mean of principal distributed leadership across 40 countries using The Teaching and Learning International Survey (TALIS, 2018) data. In this study, we used a relatively novel and recent approach known as Alignment Optimization (Asparouhov & Muthen, 2014; Muthen & Asparouhov, 2018) to compare latent means of distributed leadership scale from the perspective of principals. To the best of our knowledge, there is no alignment optimization study by employing principal distributed leadership scale) using The Teaching and Learning International Survey (TALIS, 2018), which is one of the most widespread multi-country assessments conducted by the Organization for Economic Cooperation and Development (OECD) across the world (OECD, 2019a). We choose Alignment Optimization over traditional measurement invariance approaches like partial invariance. The partial measurement invariance approach is a bit demanding, mainly once there are many groups or many items within each factor in the model since it needs manually adjusting the model as guided by the modification indices. This study revealed that Korea, Colombia, Shanghai (China) and Lithuania had the highest level of distributed leadership processes from principals' perspective, whereas the Netherlands, Belgium, Argentina, and Japan had the least principal distributed leadership perceptions. The finding of this study gives some valuable insights to policymakers, national governments, and educational systems to assimilate such that distributed leadership might be enhanced.

Afternoon Sessions

Tuesday | June 8, 2021 | 16.00 pm – 17.30 pm

Symposium 2: Using process data for investigating test-taking behaviour

Chair: Esther Ulitzsch, Support: Jana Welling

Under pressure: Measuring cognitive abilities under instruction-induced time pressure
(Alfers, Gittler, Ulitzsch & Pohl

How do examinees balance response time and accuracy on cognitive tests? Using mixture modeling to Investigate different test-taking strategies
(Beverly, Loken & Weissman)

A multilevel mixture IRT framework for modeling response times as predictors or indicators of response engagement in IRT Models
(Nagy & Ulitzsch)

Using sequence mining techniques for understanding incorrect behavioral patterns on interactive tasks
(Ulitzsch, He & Pohl)

Session 7

Chair: Marie-Ann Sengewald, Support: Anna Scharl

The achievement gap in reading competence: The effect of measurement non-invariance across school types
(Rohm)

Unfairness, rater bias and further applications for DIF-approaches in large-scale assessment
(Gürer & Draxler)

Measurement invariance: Dealing with the uncertainty in anchor item choice by Bayesian model averaging
(Schulze & Pohl)

Borrowing historical information for the analysis of large-scale assessments
(Kaplan)

Afternoon Sessions

Tuesday | June 8, 2021 | 16.00 pm – 17.30 pm

Session 8

Chair: Sina Fackler, Support: Eva Zink

Indicating information deficits at the start of university: A novel method to measure student’s level of informedness
(Mouton & Ertel)

Social origin - Key concept and blackbox?
(Siegel, Gröschl & Wohlkinger)

How can we measure young adults’ reading behavior: Is the title-recognition-test (TRT) a good alternative?
(Pfost, Schnabel & Locher)

Comparing the construct validity of trait estimates of different response formats in the measurement of learning strategies
(Tupac-Yupanqui, Heine, Schiepe-Tiska & Reiss)

Under pressure: Measuring cognitive abilities under instruction-induced time pressure

Tuesday | June 8, 2021 | 16.00 pm – 17.30 pm

Zoom Link: https://zoom.us/j/92345923960

ALFERS, TOBIAS
Freie Universität Berlin

GITTLER, GEORG
University of Vienna

ULITZSCH, ESTHER
IPN - Leibniz Institute for Science and Mathematics Education

POHL, STEFFI
Freie Universität Berlin

A broad range of professions and work environments often require a high level of cognitive skills in stressful and time-critical situations. To allow for valid assessments in terms of job-related competencies that closely reflects daily working life, cognitive skills also have to be administered under time pressured conditions. However, applying item-level or global time limits as standard procedures to operationalize time pressure introduce different measurement challenges: violation of the unidimensionality assumption, provoking guessing behavior and introducing missing values due to not reaching the end of a test. Additionally, comparisons of response behavior between stressed and non-stressed conditions are often neglected. These comparisons could provide substantial information about the change in performance when adopting cognitive skills in stressful situations. In an empirical study with over 1300 applicants for an air traffic controller training program we induced time pressure alternatively via a time pressured verbal statement within the instruction of a spatial ability test. For all participants an instruction-induced time pressure condition was administered after a condition without time pressure. Using a multidimensional extension of the hierarchical framework for modeling speed and accuracy by van der Linden (2007), we were able to a) examine the relation between latent variables for accuracy and speed within and between the two conditions and b) also obtain latent change scores for both accuracy and speed. Utilizing a combination of the proposed method of test administration and the chosen modeling approach allows for a detailed and differentiated analysis of inter- and intraindividual changes in response behavior between stressed and non-stressed conditions. Possible interpretations and implications for psychological and educational assessments will be discussed.

How do examinees balance response time and accuracy on cognitive tests? Using mixture modeling to Investigate different test-taking strategies

Tuesday | June 8, 2021 | 16.00 pm – 17.30 pm

Zoom Link: https://zoom.us/j/92345923960

The traditional joint model for response times (RT) and accuracy proposed by van der Linden (2007) assumes conditional independence, implying a fixed relationship between working speed and ability. This fixed relationship implies that working speed and ability are constant across the entire test. However, this assumption is typically violated in practice. One of the reasons for conditional dependencies is that examinees adjust their working speed, impacting their accuracy. Using data from a cognitive test, the relationship between RT and accuracy is explored using descriptive statistics and latent profile analysis. This study investigates whether examinees employ different response time-accuracy strategies during the test. From the descriptive analyses, we found a relationship between time allocation and performance across the test. This relationship is not constant; examinees switch their time allocation strategies. Time allocation strategies differ across performance groups. To understand this relationship, we explored multiple RT-accuracy patterns in the data using latent profile analysis. We found four different RT-accuracy relationships, which coincides with the descriptive analyses that many examinees switch their strategy during the test. This study's results reveal the complex relationship between time allocation and accuracy, suggesting that this relationship is not fixed across the test. Within-person differences in the time allocation and accuracy relationship cannot be accounted for in the traditional joint model.

BEVERLY, TANESIA
Law School Admission Council

LOKEN, ERIC
University of Connecticut

WEISSMMAN, ALEXANDER
Law School Admission Council

A multilevel mixture IRT framework for modeling response times as predictors or indicators of response engagement in IRT Models

Tuesday | June 8, 2021 | 16.00 pm – 17.30 pm

Zoom Link: https://zoom.us/j/92345923960

Disengaged item responses pose a threat to the validity of the results provided by large scale assessments. Several procedures for identifying disengaged responses on basis of observed response times have been suggested, and IRT models for response engagement have been proposed. We outline that response time based procedures of classifying response engagement and IRT models for response engagement are based on common ideas, and propose the distinction between independent and dependent latent class IRT modes. In all IRT models considered response engagement is represented by an item-level latent class variable, but the models either assume that response times reflect or predict engagement. Existing IRT models belonging to each group are summarized and extended to increase their flexibility. Furthermore, we propose a flexible multilevel mixture IRT framework in which all IRT models can be estimated by means of marginal maximum likelihood. The framework is based on the popular Mplus software, thereby making the procedure accessible to a broad audience. The procedures are illustrated on basis of publically available large scale data. Results showed that the IRT models provided slightly different adjustments of item parameters relative to a conventional IRT model, and provided different adjustments of individuals‘ proficiency estimates.

NAGY, GABRIEL
IPN – Leibniz Institute for Science and Mathematics Education

ULITZSCH, ESTHER
IPN – Leibniz Institute for Science and Mathematics Education

Using sequence mining techniques for understanding incorrect behavioral patterns on interactive tasks

Tuesday | June 8, 2021 | 16.00 pm – 17.30 pm

Zoom Link: https://zoom.us/j/92345923960

Interactive tasks designed to elicit real-life problem-solving behavior are rapidly becoming more widely used in educational assessment. Incorrect responses to such tasks can occur for a variety of different reasons, such as low proficiency levels, low metacognitive strategies use, or motivational issues. We demonstrate how behavioral patterns associated with incorrect responses can, in parts, be understood, supporting insights into the different sources of failure on a task. To this end, we make use of sequence mining techniques that leverage the information contained in time-stamped action sequences commonly logged in assessments with interactive tasks for a) investigating what distinguishes incorrect behavioral patterns from correct ones and b) identifying subgroups of examinees with similar incorrect behavioral patterns. Analyzing a task from the PIAAC 2012 assessment, we find incorrect behavioral patterns to be more heterogeneous than correct ones. We identify multiple subgroups of incorrect behavioral patterns, which point towards different levels of effort and lack of different subskills needed for solving the task. Albeit focusing on a single task, meaningful patterns of major differences in how examinees approach a given task that generalize across multiple tasks are uncovered. Implications for the construction and analysis of interactive tasks as well as the design of interventions for complex problem-solving skills are derived.

ULITZSCH, ESTHER
IPN – Leibniz Institute for Science and Mathematics Education

HE, QIWEI
Educational Testing Service, Princeton, USA

POHL, STEFFI
Freie Universität Berlin

The achievement gap in reading competence: The effect of measurement non-invariance across school types

Tuesday | June 8, 2021 | 16.00 pm – 17.30 pm

Zoom Link: https://zoom.us/j/94315936213

ROHM, THERESA
Leibniz Institute for Educational Trajectories (LIfBi), Bamberg

After elementary school, students in Germany are separated into different school tracks (i.e., school types) with the aim of creating homogeneous student groups in secondary school. Consequently, the development of students' reading achievement diverges across school types. Findings on this achievement gap have been criticized as depending on the quality of the administered measure. Therefore, the present study examined to what degree differential item functioning affects estimates of the achievement gap in reading competence. Using data from the German National Educational Panel Study, reading competence was investigated across three timepoints during secondary school: in grades 5, 7, and 9 (N = 7, 276). First, using the alignment method, measurement invariance across school types was tested. Then, multilevel structural equation models were used to examine whether a lack of measurement invariance between school types affected the results regarding reading development. Our analyses revealed some measurement non-invariant items that did not alter the patterns of competence development found among school types in the longitudinal modeling approach. However, misleading conclusions about the development of reading competence in different school types emerged when the clustered data structure (i.e., students being nested in schools) was not taken into account. Finally, the relevance of measurement invariance and accounting for clustering in the context of longitudinal competence measurement will be discussed.

Unfairness, rater bias and further applications for DIF-approaches in large-scale assessment

Tuesday | June 8, 2021 | 16.00 pm – 17.30 pm

Zoom Link: https://zoom.us/j/94315936213

DIF-analyses provide a strong instrument to find possible sources of unfairness in questionnaires. Recently several penalization approaches for DIF-detection in high-dimensional cases have been suggested. In this talk we will firstly stress the advantage of penalization when dealing with many potentially DIF-inducing variables and highlight the benefits of utilizing cmlDIFlasso therefore – a penalization approach based on the conditional Likelihood and the L1-penalty. Additionally, it will be shown how penalization approaches – and especially cmlDIFlasso – conveniently help to solve problems beyond unfairness, as e.g. investigating multi dimensionality, detecting rater bias depending on candidate properties and measurement of change in time. Data examples will be presented to illustrate the use and usefulness of DIF-approaches in these areas.

GÜRER, CAN
UMIT – Private University for Health Sciences, Medical Informatics and Technology

DRAXLER, CLEMENS
UMIT – Private University for Health Sciences, Medical Informatics and Technology

Measurement invariance: Dealing with the uncertainty in anchor item choice by Bayesian model averaging

Tuesday | June 8, 2021 | 16.00 pm – 17.30 pm

Zoom Link: https://zoom.us/j/94315936213

SCHULZE, DANIEL
Freie Universität Berlin

POHL, STEFFI
Freie Universität Berlin

A major goal of large scale assessments are comparisons on latent scales, e.g. differences in competence level between boys and girls, or comparisons across time. Measurement invariance (MI) is a prerequisite for such comparisons, meaning that item properties are consistent across groups or time. MI is, however, not always given. Partial MI modeling provides a solution using selected items for anchoring the scale. Different approaches for anchor item selection exist which usually require to make an assumption about MI (e.g. that MI holds for the majority of items). However, in most of the cases different assumptions may be plausible and a researcher cannot choose one of them with certainty. We argue for considering different possible solutions and depicting the uncertainty in anchor item choice in the results. Our approach consists of 1) identification of different possible anchor item sets, 2) estimating the parameter of interest using several models, each of which with a different set of anchor items, and 3) aggregating the results of all models with respect to a parameter of interest. For the last step, we propose a variant of Bayesian model averaging (BMA). This BMA approach allows for choosing model weights a priori, giving the researcher leverage to include expert knowledge on the plausibility of different anchor item sets. We derive the properties of the approach, give an applied example, and introduce an R package for this type of analysis. In conclusion, BMA facilitates the inclusion of uncertainty due to anchor item choice into the estimation of latent parameters of interest - a source of uncertainty typically neglected.

Borrowing historical information for the analysis of large-scale assessments

Tuesday | June 8, 2021 | 16.00 pm – 17.30 pm

Zoom Link: https://zoom.us/j/94315936213

KAPLAN, DAVID
University of Wisconsin

The elicitation of substantive prior information is a difficult problem for researchers using Bayesian statistical methods. In the context of research using large-scale educational assessments such as PISA, past cycles of these data sources could be used to develop informative priors for substantive research. However, large-scale assessments are usually generated from a multi-stage sampling design and these designs must be accounted for when borrowing information from historical data sources of similar design to inform current analyses. The purpose of this paper is to develop and demonstrate the use of Bayesian dynamic borrowing (Viele, 2014) for borrowing historical information with applications to large-scale educational assessments. Bayesian dynamic borrowing is a method for systematically incorporating prior historical data into current analyses where the strength of the borrowing depends on the heterogeneity among historical data and current data. A joint prior distribution over the historical and current data sets is specified with the degree of heterogeneity across the data sets controlled by the variance of the joint distribution. We demonstrate Bayesian dynamic borrowing in both single-level and multilevel models. We also discuss an alternative approach to historical borrowing referred to as the power prior (Ibrahim & Chen, 2000) and compare its performance to Bayesian dynamic borrowing in single-level and multilevel situations. Two case studies using data from the Program for International Student Assessment reveal the utility of Bayesian dynamic borrowing in terms of predictive accuracy. This is followed by two simulation studies that reveal the utility of Bayesian dynamic borrowing over simple pooling and power priors in cases where the historical data is heterogeneous compared to current data based on bias, mean squared error, total effective sample size, and predictive accuracy. In cases of homogeneous historical data, Bayesian dynamic borrowing performs similarly to complete data pooling and power priors.

Indicating information deficits at the start of university: A novel method to measure student's level of informedness

Tuesday | June 8, 2021 | 16.00 pm – 17.30 pm

Zoom Link: https://zoom.us/j/99664826211

MOUTON, DIVAN
Universität der Bundeswehr, Munich

ERTL, BERNHARD
Universität der Bundeswehr, Munich

Choosing a career path requires students to sift through a multitude of possible information sources on a wide range of study and vocational alternatives. Although most students find helpful information, some students are not optimally informed about studying and are more likely to be uncertain about their career choice. Conversely, a substantial number of German university students have claimed false study expectations as one of the most decisive reasons for dropout, which has been cited as an indirect indication of information deficits at the beginning of studies. To investigate the information deficit phenomenon this study uses a novel method to construct informedness groups and compared them to a single five-point item about how well-informed students feel about studying. The method ranked the highest level of perceived usefulness gained from any information source used for deciding on and planning their studies. A student who found at least one source (4) "very helpful" is Well Informed, Fairly Informed students found at least one source (3) "rather helpful", while the lower usefulness scores were Poorly Informed. Construct validity was tested using indicators of information deficits. This study found that students who do not find any source of information as optimally useful showed significantly poorer trends on indicators of information deficit, as compared to their better-informed counterparts. When compared to the single item, the informedness groups produced significant distinctions on indicators, especially at higher informedness levels. While the Well Informed group is significantly associated to students who indicate being (5) "very well" informed, the Fairly Informed group are associated with students who are either (2) "rather poorly" or (3) "partly" informed. The disaggregated version of informedness appears to prompt students to reflect on their level of informedness better than a single item. Results suggest that informedness groups are a valid indicator of information deficit.

Social origin - Key concept and blackbox?

Tuesday | June 8, 2021 | 16.00 pm – 17.30 pm

Zoom Link: https://zoom.us/j/99664826211

SIEGEL, FABIAN
Ludwig-Maximilians Universität, Munich

GRÖSCHEL, BENJAMIN
Ludwig-Maximilians Universität, Munich

WOHLKINGER, FLORIAN
Ludwig-Maximilians Universität, Munich

For centuries, social science looked upon the influence of the social class, socio-economic resources and family-cultural background on processes of inter-generational mobility. In educational research, this complex of issues (typically labeled as ‘social origin‘) could be described as a long-time favorite and one of the most influential factors on individual educational trajectories (Sirin, 2005). Although social origin is an empirically very successful and widely used concept that is taken into account by a variety of theoretical perspectives, the choice of indicators for this concept is rarely justified theoretically, but rather seems to represent a self-propelling practice (Thaning & Hällsten, 2020). Concerning the empirical success in large-scale assessments, the key concepts of social origin and the relationships within this subject area seem to be locked in a black box (Latour, 2002). On the other hand, it is indispensable for a cumulative research process to transparently present both the basic theoretical construction and the empirical consequences resulting from the choice of a particular indicator. Only in this way can a fit between theory, method, and data, which is appropriate to the complexity of the research object, be achieved (Merton, 1995). Therefore, this contribution aims to (re-)open the black box of social origin. We will discuss the theoretical implications of commonly used indicators provided by studies like NEPS, such as ISEI or EGP. In addition to this focus, a brief outlook on examples of methodical challenges will illustrate the broader perspective of the project. Our main objective is to raise awareness of the complexity of the 'social origin' construct and discuss its different methodological implications resulting from different theoretical premises. Among other things, this is supposed to revitalise the discourse for a more enlightened, theory-sensitive use of existing indicators or the development of innovative and more sophisticated operationalisations.

How can we measure young adults' reading behavior: Is the title-recognition-test (TRT) a good alternative?

Tuesday | June 8, 2021 | 16.00 pm – 17.30 pm

Zoom Link: https://zoom.us/j/99664826211

PFOST, MAXIMILIAN
Otto Friedrich University, Bamberg

SCHNABEL, VERENA
Otto Friedrich University, Bamberg

LOCHER, FRANZISKA
Pädagogische Hochschule St.Gallen

Prior research has shown that reading behavior, and especially the reading of fiction books, is closely related to reading comprehension (Mol & Bus, 2011; Pfost, Dörfler & Artelt, 2013). However, there is a large variety of instruments that intend to measure students' and adults' reading behavior, such as global and text-type-specific retrospective ratings (e.g., Locher & Pfost, 2019), reading diaries (e.g., Anderson, Wilson & Fielding, 1988) or indirect measures such as author- and title-recognition-tests (e.g., Cipielewski & Stanovich, 1992). The basic idea of a recognition test is that a person has to differentiate between book titles (TRT) or authors (ART) known and not known to him/her from a list of book titles respectively authors. Furthermore, in order to prevent persons from guessing, the list of book titles/authors is interfused with imaginary book titles/authors. Although recognition tests are widely used in education research (Mol & Bus, 2011), it was seldom questioned how such tests behave in comparison to other reading behavior measures. In this talk, we will present findings on two studies on university students (N = 91 and N = 35) that implemented different reading behavior measures and compared these measures. Our results show non-significant correlations of the TRT with global retrospective ratings of time spend reading. Correlations of the TRT with ratings on the number of books that had been read in the last twelfth months were significant. Furthermore, there was a minor positive, although not significant, relation of the TRT to the time spend reading measured by an electronic reading diary. Taken together, the TRT seems to cover a different facet of reading behavior as do other measures of reading behavior such as retrospective ratings. Future research might take a closer look on further variables such as film adaptions and its effect on the TRT-score.

Comparing the construct validity of trait estimates of different response formats in the measurement of learning strategies

Tuesday | June 8, 2021 | 16.00 pm – 17.30 pm

Zoom Link: https://zoom.us/j/99664826211

In international large-scale assessments, latent constructs are typically assessed via rating-scale response formats, such as Likert-type scales. Especially in international comparisons of diverse populations, critics point out response biases, such as acquiescence, socially desired responding and other distortions. However, the validity of self-report measurements is crucial for a proper interpretation of the results. An alternative might be the forced-choice response format. The objective of the present study is to investigate the construct validity of the two different response formats, forced-choice and rating-scale, to assess mathematical learning strategies in adolescent students in the Programme for International Student Assessment. This study examines whether these two response formats measure the same latent variables and how socially desirable responses are related to each response format. The sample consisted of ninth graders in Germany that participated in the PISA main study in 2012 (*n* = 5739). Mathematical learning strategies included control, elaboration and memorisation strategies and were assessed as forced-choice scale that were scaled by applying the Thurstonian IRT model as well as rating-scale that were scaled using the cumulative partial credit IRT model. Socially desirable response behavior was assessed using a rating-scale comprising distinct measures for self-deception and impression management. Results showed lower reliability estimates for the forced-choice response format compared to the rating-scale format. A multitrait-multimethod-correlation matrix showed low to moderate relations between the estimated traits of the three different learning strategies. The correlation for the social desirability scale was lower for the forced-choice than for the rating-scale format. The current findings suggest that it is still unclear whether the two response formats assess the same latent variables. They might assess different aspects of the latent variables or respondents understand them differently. Assessing learning strategies validly is essential to improve educational assessment practices.

TUPAC-YUPANQUI, ANA
Technical University of Munich

HEINE, JÖRG-HENRIK
Technical University of Munich

SCHIEPE-TISKA, ANJA
Technical University of Munich

REISS, KRISTINA
Technical University of Munich