## Contents

Greeting .................................................. 7

### General Information .................................. 9

- Directions ............................................. 10
- Child care .............................................. 11
- Lunch and Snacks ..................................... 11
- Internet access ....................................... 11

### Schedule ............................................. 13

- Tuesday, July 24 ...................................... 14
- Wednesday, July 25 .................................. 15
- Thursday, July 26 .................................... 19
- Friday, July 27 ........................................ 25

### Pre-Conference-Workshops ......................... 31

- WORKSHOP Intensive Longitudinal Data Analysis / DSEM ........................................ 32
- WORKSHOP Multilevel Structural Equation Modeling with Lavaan ............................. 33
- WORKSHOP Hypothesis evaluation using the Bayes factor ........................................ 34
- WORKSHOP Understanding SEM: Where do all the numbers come from? ..................... 35
- WORKSHOP Theory and analysis of conditional and average causal effects .................. 36

### Poster .................................................. 37

- Poster Session 1 ....................................... 39
- Poster Session 2 ....................................... 65

### Scientific Program Wednesday, July 25 ........... 87

- 10:15 ..................................................... 88
  - KEYNOTE The Statistics of Replication ................................................................. 88
- 11:30 ..................................................... 90
  - SESSION Causal Inference in Dynamic Models ..................................................... 90
  - SESSION Item Response Theory ................................................................. 95
  - SESSION Applied Statistics ................................................................. 100
  - SESSION Replication Crisis ................................................................. 104
STATE-OF-THE-ART Using incidental data for serious social research .............................. 109

16:30  
SESSION Latent Variable Analysis ............................................................................ 112
SESSION Multilevel Analysis .................................................................................. 118
SESSION Structural Equation Modeling .................................................................. 124
SYMPOSIUM New Developments in Mokken Scale Analysis ................................. 130

Scientific Program Thursday, July 26

08:30  
KEYNOTE Towards a deeper understanding of the effectiveness of interventions: New methods based on structural equation models and causal inference ......................... 138

09:30  
SESSION Large Scale Data .................................................................................. 140
SYMPOSIUM Differential Item Functioning (DIF) in Educational Settings: Methods, Simulations and Applications ............................................................... 145
SESSION Factor Analysis ...................................................................................... 150
SESSION Applied Statistics .................................................................................. 156

11:30  
SESSION Bayesian Statistics ................................................................................ 162
SESSION Causal Inference ................................................................................... 162
SYMPOSIUM Response Time Modeling in Psychometrics ...................................... 174
SESSION Experimental Design ............................................................................ 180

15:30  
STATE-OF-THE-ART The role of time in dynamic models of change .................... 185
STATE-OF-THE-ART Optimal Design ................................................................... 186

16:30  
SESSION Structural Equation Modeling ............................................................... 188
SESSION Latent Variable Analysis ...................................................................... 188
SESSION Item Response Theory ......................................................................... 199
SESSION Applied Statistics ................................................................................. 203

Scientific Program Friday, July 27

09:00  
KEYNOTE How to make missingness ignorable in longitudinal modeling ................. 208

10:00  
SESSION Item Response Theory ........................................................................ 210
SESSION Missing Data ......................................................................................... 215
SYMPOSIUM Applying Subjective Bayes to Real Life Data ..................................... 219
<table>
<thead>
<tr>
<th>Time</th>
<th>Session</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>11:30</td>
<td>SESSION Longitudinal Data</td>
<td>223</td>
</tr>
<tr>
<td></td>
<td>SESSION Latent Variable Analysis</td>
<td>228</td>
</tr>
<tr>
<td></td>
<td>SESSION Multilevel Modeling</td>
<td>234</td>
</tr>
<tr>
<td></td>
<td>SESSION Bayesian Statistics</td>
<td>240</td>
</tr>
<tr>
<td>14:00</td>
<td>SESSION Item Response Theory</td>
<td>245</td>
</tr>
<tr>
<td></td>
<td>SYMPOSIUM Challenges in Interdisciplinary Research Methodology: The Study of Complex Systems</td>
<td>250</td>
</tr>
<tr>
<td></td>
<td>SESSION Latent Class Models</td>
<td>256</td>
</tr>
<tr>
<td></td>
<td>SESSION Applied Statistics</td>
<td>260</td>
</tr>
<tr>
<td>15:30</td>
<td>STATE-OF-THE-ART Systematic observation of human behavior from a methodological perspective</td>
<td>265</td>
</tr>
<tr>
<td></td>
<td>STATE-OF-THE-ART Addressing Treatment Non-adherence in Randomized Experiments</td>
<td>266</td>
</tr>
</tbody>
</table>

**Index** 268

**Impressum** 274
Greeting

Dear friends and colleagues,

It is a great pleasure to welcome you in Jena at the 8th conference of the European Association of Methodology (EAM). It is a great honor for my university that the EAM gave us the opportunity to organize this conference. I hope that it will provide the opportunities for many inspiring meetings, presentations, and discussions.

I would like to thank the keynote speakers Sophia Rabe-Hesketh from Berkeley University, Larry Hedges from Northwestern University in Chicago, and the vice-president of the EAM, Axel Mayer from the University of Aachen, for their contributions. I am also looking forward a lot to the state-of-the-art talks offered by Maria Teresa Anguera, Heinz Holling, José Muniz, Daniel Oberski, Manuel Voelkle, and Steve West. However, there are many other contributions in symposia, paper sessions, and poster sessions, which promise that this conference will offer inspiring days. Thanks to all of you who contribute.

I would also like to thank the members of the scientific committee for their evaluations of the contributions. This helped a lot to prepare the symposia, paper sessions, and poster presentations.

Thanks are also due to the members of the local organization team, in particular Andreas M. Neudecker, Marcel Bauer, and Katrin Schaller, who shouldered most of the work associated with the organization of this conference.

Finally, I would also like to thank the sponsors of this conference: The Friedrich Schiller University of Jena, the Deutsche Forschungsgemeinschaft (DFG), who supported us with a generous grant, the publishers Hogrefe and Springer, as well as JMP from SAS.

Finally, I hope that the city of Jena with his numerous restaurants and pubs offers a nice atmosphere for your post conference discussions and that you will leave us with many new exciting ideas and new friends.

Rolf Steyer
Chair of local organization committee
General Information
Directions to the Conference Venues

Steigenberger Hotel (Conference Venue)

Address: Carl-Zeiss-Platz 4, 07743 Jena

The Steigenberger Hotel is located right next to the Uni-Campus and can be reached by tram number 5 (stop Ernst-Abbe-Platz) or by walking from the city centre. Current information on public transport in Jena is available on the website: www.jenah.de

IAAC (pre-conference-workshops)

Address: Humboldtstraße 8, 07743 Jena

The IAAC is located near the bus stop Am Steiger. To get there you can take the bus number 16 (direction Mühltal / Isserstedt) from one of the bus stops Löbdergraben or Teichgraben in the city centre. Current information on public transport in Jena is available on the website: www.jenah.de

Multimediazentrum MMZ (pre-conference-workshop Steyer)

Address: Ernst-Abbe-Platz 8, 07743 Jena

The Multimediazentrum is located on the Uni-Campus. It can be reached by tram number 5 (stop Ernst-Abbe-Platz) or by a short walk from the city centre.

Cafeteria zur Rosen (Welcome Reception)

Address: Johannisstraße 13, 07743 Jena

Paradiescafé (Conference Dinner)

Address: Vor dem Neutor 5, 07743 Jena

The Paradiescafé is located in the Paradies Park (Paradise Parc) which lies on both sides of the Saale river. You can enter the park e.g. from the train station "Paradies" or from the tram stop "Felsenkeller" (tram number 2). The Paradiescafe is located between these two locations.

Parking facilities

In the city centre you can find parking spaces e.g. in the underground car parks of the shopping centres Goethe-Galerie (entrance Ernst-Abbe-Straße, 7am–11pm) and Neue Mitte (entrance Kollegiengasse, 24 h) and at the so-called Inselplatz. From there it is only a short walk to the bus stops Löbdergraben and Teichgraben.
Child care

JuniKinder offers a child care program. You can find further details on the following website: [www.familie.uni-jena.de/en/Family+Office/JUni+Kids+_+flexible+child+care](http://www.familie.uni-jena.de/en/Family+Office/JUni+Kids+_+flexible+child+care).

Lunch and Snacks

In the coffee breaks we will offer you free snacks and drinks.

Canteens

For lunch there are three university canteens within walking distance. You can find their current offers on the following website: [www.stw-thueringen.de](http://www.stw-thueringen.de). The vegetarian canteen VegeTable and the canteen Ernst-Abbe-Platz are located directly on campus. There you can also find some tables outside. In a few minutes walk there is also the canteen Philosophenweg (Philosophenweg 20, 07743 Jena). On the back side of the canteen Philosophenweg there is a garden with tables and deck chairs where you can eat and enjoy the sun.

Other lunch possibilities

The highest restaurant density can be found in Jena’s popular Wagnergasse. Between Johannisplatz and IAAC you can find a great choice of different dishes and snacks. A few suggestions:

<table>
<thead>
<tr>
<th>Name</th>
<th>Where in the Wagnergasse?</th>
<th>What kind of food?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asia-Food</td>
<td>Corner Quergasse</td>
<td>Asian fast food</td>
</tr>
<tr>
<td>Café Wagner</td>
<td>Wagnergasse 26</td>
<td>Vegetarian and vegan cuisine</td>
</tr>
<tr>
<td>Fritz Mitte</td>
<td>Johannisplatz</td>
<td>Chips, currywurst and mayonnaise</td>
</tr>
<tr>
<td>Stilbruch</td>
<td>Wagnergasse 1–3</td>
<td>Regional and international kitchen</td>
</tr>
<tr>
<td>Sushi friends</td>
<td>Johannisplatz</td>
<td>Sushi</td>
</tr>
</tbody>
</table>

Internet access

At the conference Venue

You can login to the WLAN at the Steigenberger Hotel. You need to connect to the WLAN first and then login using the following credentials

Name: WLAN
Zimmernummer (Room Number): 2005
At the IAAC and the Multimediazentrum

Friedrich-Schiller-University takes part in DFN roaming of the DFN association and is thus integrated into the eduroam project. As a member of institutions participating in DFN roaming you can sign up without any further registration with the user ID you already have. To do so you only need a valid account at the FSU Jena or at your home institution. The concrete settings on your own device depend on the specifications of the FSU Jena. WLAN with the identifier (SSID = ServiceSetIdentifier) “Eduroam” is available throughout the university. Information on alternative internet access for participants without DFN-roaming-account will be given at the venue.

You need a computer with 802.11b / g WLAN interface, which handles 802.1X authentication with WPA2 / WPA | EAPPEAP | MSCHAPv2 or WPA2 / WPA | EAP-TTLS | Mastered PAP. Please note that Deutsche Telekom Root CA 2 must be registered as superior root certificate. In the case of questions or problems more information is available at: www.uni-jena.de/WLAN_der_FSU_Jena.

Particular information on the IAAC building

Since it is a building of chemistry we want to point out in advance that you should leave the building immediately in the case of a fire alarm. We also request you to refrain from drinking in the auditorium or sitting down on the small tables. In the seminar rooms it is permitted to drink water. Thank you for your understanding!

Sponsors
Schedule
<table>
<thead>
<tr>
<th>Time</th>
<th>Location</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>09:00</td>
<td>IAAC - SR1</td>
<td>WORKSHOP Hypothesis evaluation using the Bayes factor Part 1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Chair: Herbert Hoijtink</td>
</tr>
<tr>
<td>14:00</td>
<td>IAAC - SR1</td>
<td>WORKSHOP Hypothesis evaluation using the Bayes factor Part 2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Chair: Herbert Hoijtink</td>
</tr>
<tr>
<td></td>
<td>IAAC - SR2</td>
<td>WORKSHOP Multilevel Structural Equation Modeling with Lavaan</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Chair: Yves Rosseel</td>
</tr>
<tr>
<td></td>
<td>IAAC - SR3</td>
<td>WORKSHOP Intensive Longitudinal Data Analysis / DSEM</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Chair: Noémi Schuurman</td>
</tr>
<tr>
<td>19:00</td>
<td>Cafeteria zur Rosen</td>
<td>EVENT Welcome Reception</td>
</tr>
<tr>
<td></td>
<td>Multimediazentrum - HS E028</td>
<td>WORKSHOP Theory and analysis of conditional and average causal effects</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Chair: Rolf Steyer</td>
</tr>
</tbody>
</table>
Wednesday, July 25

09:00 - Saal Friedrich Schiller
OPENING Opening Ceremony

10:15 - Saal Friedrich Schiller
KEYNOTE The Statistics of Replication (Larry Hedges)
Chair: Heinz Holling

11:00 - Coffee Break (Hotellobby)
<table>
<thead>
<tr>
<th>Time</th>
<th>Session</th>
<th>Room</th>
<th>Title</th>
<th>Chair(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Item Response Theory</td>
<td>Salon Schlegel</td>
<td>A State-Space Approach for Student Growth Percentile Estimation</td>
<td>Frans Kamphuis</td>
</tr>
<tr>
<td></td>
<td>Item Response Theory</td>
<td>Salon Höffelerin</td>
<td>A General Nonlinear Model for the Identification of Mediators Without the No Confounder Assumption</td>
<td>Holger Brandt</td>
</tr>
<tr>
<td></td>
<td>Applied Statistics</td>
<td>Salon Novalis</td>
<td>Model-Based Reliability and Item Bias in PISA and PIACC</td>
<td>Jörg Blasius</td>
</tr>
<tr>
<td></td>
<td>Replication Crisis</td>
<td>Salon Hölderlin</td>
<td>Fabrication of Interview Data</td>
<td>Jörg Blasius</td>
</tr>
<tr>
<td></td>
<td>Replication Crisis</td>
<td>Salon Hölderlin</td>
<td>Transparency and Replicability in Cross-National Survey Research</td>
<td>Elena Damian</td>
</tr>
<tr>
<td></td>
<td>Item Response Theory</td>
<td>Salon Schlegel</td>
<td>Psychometric Evaluation of the D2 Test of Sustained Attention With the Rasch Model</td>
<td>Parya Baghaei</td>
</tr>
<tr>
<td></td>
<td>Item Response Theory</td>
<td>Salon Schlegel</td>
<td>Forecasting Clinical Outcomes by Combining Measurement and Prediction Models for Health Evaluations</td>
<td>Niels Smits</td>
</tr>
<tr>
<td></td>
<td>Item Response Theory</td>
<td>Salon Schlegel</td>
<td>Impact of Formal Educational Upgrading on the Likelihood of Leaving Unemployment</td>
<td>Johannes Jaenicke</td>
</tr>
<tr>
<td></td>
<td>Item Response Theory</td>
<td>Salon Schlegel</td>
<td>Exchanging Selection Rules from Cognitive Diagnosis to Traditional Item Response Theory</td>
<td>Miguel A. Sorrel</td>
</tr>
<tr>
<td>Time</td>
<td>Location</td>
<td>Event Description</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------</td>
<td>---------------------</td>
<td>----------------------------------------------------------------------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13:00</td>
<td>Lunch Break</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14:00</td>
<td>Foyer (Saal Friedrich Schiller)</td>
<td>Poster Session 1&lt;br&gt;<strong>STATE-OF-THE-ART</strong> Using incidental data for serious social research (Daniel Oberski) Chair: Axel Mayer</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
16:30  Saal Friedrich Schiller  
**SESSION Latent Variable Analysis**  
Chair: Tobias Koch  
Analyzing Different Types of Moderated Method Effects in Confirmatory Factor Models for Structurally Tobias Koch  
Repeated Measures ANOVA with Latent Variables Using the Latent Growth Component Approach Benedikt Langenberg  
DIF of Self-Assessment Items Across Different Levels of a Latent Variable: Positive Affect Ana Hernández  
Parameter Associations in Bivariate Dual Change Score Models: Implications for Simulation Studies Holly O’Rourke  

Salon Schlegel  
**SESSION Multilevel Analysis**  
Chair: Johannes Hartig  
Estimation of Random Group DIF Using Two- and Three-Level GLMMs Johannes Hartig  
Multilevel Models for Evaluating the Effectiveness of Instruction: ANCOVA vs. Change-Score Approach Carmen Köhler  
The Optimal Design of Cluster Randomized Trials With Outcomes at Individual and Cluster Level Mirjam Moerbeek  
Comparative Performance of Single Trial Multilevel Analyses of Event-Related Brain Potentials Juan Carlos Oliver-Rodríguez  

Salon Hölderlin  
**SESSION Structural Equation Modeling**  
Chair: Keith Widaman  
Unreliability Has Important Negative Effects: Correcting May Be Easier Than You Think Keith Widaman  
Multilevel SEM for Discrete Data With the Pairwise Likelihood Estimation Method Mariska Barendse  
An Alternative Estimation Method for Multilevel SEM Based on Factor Scores Ines Devlieger  
Omitted Cross-Loadings in Nonlinear SEM: A Monte Carlo Study Karina Rdz-Navarro  

Salon Novalis  
**SYMPOSIUM New Developments in Mokken Scale Analysis**  
Chair: Andries van der Ark  
Introduction to Mokken Scale Analysis Klaas Sijtsma  
Checking Assumptions in Two-Level Mokken Scale Analysis Letty Koopman  
Two-Level Mokken Scale Analysis: The State of the Art Andries van der Ark  
Using Mokken Scaling Techniques to Evaluate Educational Assessments Stefanie Wind
08:30  Saal Friedrich Schiller

KEYNOTE Towards a deeper understanding of the effectiveness of interventions: New methods based on structural equation models and causal inference (Axel Mayer)
Chair: Rolf Steyer
<table>
<thead>
<tr>
<th>Time</th>
<th>Location</th>
<th>Session</th>
<th>Symposium/Session</th>
<th>Chair/Author</th>
</tr>
</thead>
<tbody>
<tr>
<td>09:30</td>
<td>Saal Friedrich Schiller</td>
<td>SESSION Large Scale Data</td>
<td>Chair: Steffi Pohl</td>
<td>Item-Person Mismatch and Parameter Recovery Accuracy in Sparse Multi-Matrix Booklet Designs</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Anta Akuro</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Measurement Invariance of the Academic Performance for Fifteen Countries With the Alignment Method</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>José Muñiz-Fernández</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Treatment of Measurement Error and Missing Data Using Nested and Non-Nested Multiple Imputation</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Simon Grund</td>
</tr>
<tr>
<td></td>
<td>Salon Schlegel</td>
<td>SYMPOSIUM Differential Item Functioning (DIF) in Educational Settings: Methods, Simulations and Applications</td>
<td>Chair: Rudolf Debelak, Martin Tomasik</td>
<td>A Regularized Moderated Item Response Model for Assessing Differential Item Functioning</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Alexander Robitzsch</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Calibration of a Criterion-Referenced Computerized Adaptive Test in Higher Education</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Hanna Köhler</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Differential Item Functioning in the Context of Multistage Testing</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Sebastian Appelbaum</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>A Flexible Method for the Detection of Differential Item Functioning</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Rudolf Debelak</td>
</tr>
<tr>
<td></td>
<td>Salon Hölderlin</td>
<td>SESSION Factor Analysis</td>
<td>Chair: Florian Scharf</td>
<td>Orthogonal Versus Oblique Rotation in Temporal EFA for Event-Related Potentials</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Florian Scharf</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Common Factor Analysis and Principal Component Analysis: Competing Indeterminacies</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Keith Widaman</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>The Number of Factors in Exploratory Factor Analysis</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Max Auerswald</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>On the Influence of Processing Speed on Investigations of Structural Validity: A Simulation Study</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Karl Schweizer</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Thomas Schäfer</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Introducing Indigenous Methodology Into the Practice of the European Social Research</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Justyna Pilarska</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>The Impact of Test-Review Models on Improving Tests and Testing: The Case of Spain</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Hernández Ana</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Neets in French Labour Market: A Multidimensionnal and Fuzzy Approach</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Claire Bonnard</td>
</tr>
</tbody>
</table>
Thursday, July 26 (continued)

<table>
<thead>
<tr>
<th>11:00</th>
<th><strong>Coffee Break</strong> (Hotellobby)</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>11:30</th>
<th>Saal Friedrich Schiller</th>
<th>Salon Schlegel</th>
<th>Salon Hölderlin</th>
<th>Salon Novalis</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>SESSION Bayesian Statistics</strong></td>
<td><strong>SESSION Causal Inference</strong></td>
<td><strong>SYMPOSIUM Response Time Modeling in Psychometrics</strong></td>
<td><strong>SESSION Experimental Design</strong></td>
</tr>
<tr>
<td></td>
<td>Chair: Herbert Hoijtink</td>
<td>Chair: Rolf Steyer</td>
<td>Chair: Steffi Pohl</td>
<td>Chair: Volker Kraft</td>
</tr>
<tr>
<td></td>
<td>Computing Bayes Factors</td>
<td>Average Effects Based on</td>
<td>Disentangling Missingness</td>
<td>Statistical Power in Pooled</td>
</tr>
<tr>
<td></td>
<td>From Data With Missing Values</td>
<td>Regressions with Log Link: A New Approach with Stochastic Covariates</td>
<td>Due to Lack of Speed From Missingness Due To Quitting</td>
<td>Time Series</td>
</tr>
<tr>
<td></td>
<td><em>Herbert Hoijtink</em></td>
<td><em>Christoph Kiefer</em></td>
<td><em>Esther Ulitzsch</em></td>
<td><em>Marcel Elipe Miravet</em></td>
</tr>
<tr>
<td></td>
<td>Handling Ordinal Predictors</td>
<td>Ignoring Ignorability: Towards a Realistic Public Policy Evaluation</td>
<td><em>Response Time Models for Automated Test Assembly</em></td>
<td>Understanding (the) Power in</td>
</tr>
<tr>
<td></td>
<td>in Regression Models via</td>
<td><em>Trinidad Gonzalez</em></td>
<td><em>Benjamin Becker</em></td>
<td>Designed Experiments</td>
</tr>
<tr>
<td></td>
<td>Monotonic Effects</td>
<td></td>
<td></td>
<td><em>Volker Kraft</em></td>
</tr>
</tbody>
</table>
|        | *Paul Bürkner*                  |                          | Response Times and Latent | Degrees of Freedom Ap-
|        | Bayesian Estimation for         |                          | Response Style Classes in | proximations in Multilevel |
|        | Cases of Empirical Under-
|        | identification                 |                          | Noncognitive Measures   | Meta-Analysis of Standard-
|        | *Jonathan Helm*                 |                          | *Artur Pokropek*        | ized Single-Case Experiment |
|        | Operationalizations of Inac-
|        | curacy of Prior Distributions  | How to Model Production | A Finite-State Machine | *Laleh Jamshidi*       |
|        | in Simulation Studies           | in Psychology? A Bayesian | Approach to Extract Item |                       |
|        | *Milica Miocevic*               | Stochastic Frontier Struc-
|        |                                | tural Equation Model     | Response Times From Que-
|        |                                | *Rüdiger Mutz*           | stionnaire Item Batteries ||
|        |                                | Bias in Estimating Treatment |                          | *Ulf Kroehne*          |
|        |                                | Effects of Latent Non-normal Distributed Outcome Variables With Binary Indicators |


<table>
<thead>
<tr>
<th>Time</th>
<th>Location</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>13:00</td>
<td>Lunch Break</td>
<td>Foyer (Saal Friedrich Schiller)</td>
</tr>
<tr>
<td>14:00</td>
<td>Poster Session 2</td>
<td>Salon Schlegel</td>
</tr>
<tr>
<td></td>
<td>Poster Session 2</td>
<td>Executive Committee Meeting</td>
</tr>
<tr>
<td>15:30</td>
<td>Saal Friedrich Schiller</td>
<td>STATE-OF-THE-ART The role of time in dynamic models of change (Manuel Völkle)</td>
</tr>
<tr>
<td></td>
<td>Salon Schlegel</td>
<td>STATE-OF-THE-ART Optimal Design (Heinz Holling)</td>
</tr>
<tr>
<td></td>
<td>Chair: Noémi Schuurman</td>
<td>Chair: Mirjam Moerbeek</td>
</tr>
<tr>
<td>16:00</td>
<td>Coffee Break</td>
<td>(Hotellobby)</td>
</tr>
<tr>
<td>Time</td>
<td>Location</td>
<td>Session</td>
</tr>
<tr>
<td>-------</td>
<td>------------</td>
<td>----------------------------------------------</td>
</tr>
<tr>
<td>16:30</td>
<td>Saal Schiller</td>
<td><strong>SESSION: Structural Equation Modeling</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Salon Schlegel</td>
<td><strong>SESSION: Latent Variable Analysis</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Salon Hölderlin</td>
<td><strong>SESSION: Item Response Theory</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Salon Novalis</td>
<td><strong>SESSION: Applied Statistics</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Thursday, July 26

19:00 Paradiescafe

EVENT Conference Dinner
09:00  Saal Friedrich Schiller  
KEYNOTE: How to make missingness ignorable in longitudinal modeling (Sophia Rabe-Hesketh)  
Chair: Mirjam Meebeek
<table>
<thead>
<tr>
<th>Time</th>
<th>Location</th>
<th>Session/Subject</th>
<th>Title</th>
<th>Chair(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10:00</td>
<td>Saal Friedrich Schiller</td>
<td>Item Response Theory Chair: Christoph König</td>
<td>Reducing Sample Size Requirements of the 2PL With a Bayesian Hierarchical Approach Christoph König</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Salon Schlegel</td>
<td>Item Response Theory Chair: Christoph König</td>
<td>Instructional Sensitivity of Polytomous Test and Questionnaire Items Alexander Naumann</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Salon Hölderlin</td>
<td>Item Response Theory Chair: Christoph König</td>
<td>Estimation of a Multidimensional Item Response Model Using Bayesian Nonparametrics Felix Naumann</td>
<td></td>
</tr>
<tr>
<td>11:00</td>
<td></td>
<td>Coffee Break (Hotellobby)</td>
<td>Coffee Break (Hotellobby)</td>
<td></td>
</tr>
<tr>
<td>Time</td>
<td>Salon Location</td>
<td>Session</td>
<td>Chair</td>
<td>Title</td>
</tr>
<tr>
<td>-------</td>
<td>----------------</td>
<td>---------</td>
<td>-------</td>
<td>----------------------------------------------------------------------</td>
</tr>
<tr>
<td>11:30</td>
<td>Saal Friedrich Schiller</td>
<td><strong>Longitudinal Data</strong></td>
<td>Manuel Arnold</td>
<td>Individual Parameter Contribution Regression for Longitudinal Data</td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Salon Schlegel</strong></td>
<td></td>
<td>Generalized Continuous Time Models and the Continuous Time Rasch Model as an Example</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>The Association Between Depression and Education in UK Adolescents: A Cross-Lagged Panel Analysis</td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Salon Hölderlin</strong></td>
<td></td>
<td>On the Effect of Observations and Parameters on the fit of SEM Models With Large Sample-Sizes</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Evaluating Model Quality in Exploratory Bi-factor Modelling</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Analyzing Approximate Invariance From a Mixed-Method Ecological Approach to Validation</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>A Correlated Covariate Amplifies the Bias of a Fallible Covariate in Causal Effect Estimates</td>
</tr>
<tr>
<td></td>
<td>Salon Novalis</td>
<td><strong>Multilevel Modeling</strong></td>
<td>Wouter Talloen</td>
<td>Measurement Error and Unmeasured Confounding in Multilevel Mediation Models</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Same Same but Different?! Measuring of Local Sex Ratios</td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Bayesian Statistics</strong></td>
<td>David Kaplan</td>
<td>Assessing Structures of Prejudice in Europe with Multilevel Latent Class Analysis</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Detecting Selection Bias in Meta-Analyses with Dependent Effect Sizes: A Simulation Study</td>
</tr>
<tr>
<td>Time</td>
<td>Location</td>
<td>Session</td>
<td>Topic</td>
<td>Chair</td>
</tr>
<tr>
<td>-------</td>
<td>---------------</td>
<td>----------------------------------------------</td>
<td>----------------------------------------------------------------------</td>
<td>---------------------</td>
</tr>
<tr>
<td>13:00</td>
<td>Lunch Break</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14:00</td>
<td>Saal Friederich Schiller</td>
<td><strong>SESSION Item Response Theory</strong></td>
<td>Chair: Timo Bechger</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Salon Schlegel</td>
<td>SYMPOSIUM <strong>Challenges in Interdisciplinary Research Methodology: The Study of Complex Systems</strong></td>
<td>Chair: Hilde Tobi</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>DIF Methods in Dexter</td>
<td>Ivailo Partchev</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>The Great Dexperiment: Psychometrics With Observed Variables</td>
<td>Timo Bechger</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Bayesian Estimation of Item Response Models to Account for Learning During the Test</td>
<td>José H. Lozano</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>A Probabilistic IRT Model for the Joint Assessment of Objects and Persons in Fully Crossed Designs</td>
<td>Georg Hosoya</td>
<td></td>
</tr>
<tr>
<td>14:00</td>
<td>Salon Hölderlin</td>
<td><strong>SESSION Latent Class Models</strong></td>
<td>Chair: Ana Gomes</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Internet Use in the European Union: A Multilevel Latent Class Analysis</td>
<td>Ana Gomes</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Using Simulation Models to Measure Resilience</td>
<td>Guus ten Broeke</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>A Modest Step Toward Bringing Unity in Interdisciplinary Research</td>
<td>Jarl Kennard Kampen</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Innovation Modelling in Engineering and Scholastic Philosophy</td>
<td>Julia P. A. von Thienen</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mapping Validity in Modelling for Interdisciplinary Research</td>
<td>Hilde Tobi</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Time</th>
<th>Location</th>
<th>Session</th>
<th>Topic</th>
<th>Chair</th>
</tr>
</thead>
<tbody>
<tr>
<td>14:00</td>
<td>Salon Novalis</td>
<td><strong>SESSION Applied Statistics</strong></td>
<td>Chair: Alrik Thiem</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Interpretation of Main Effects for Moderated Regression Models</td>
<td>Julie Lorah</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Topic Modeling As a Type-Forming Process of Social-Ecological Education Research</td>
<td>Thomas Prescher</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Small Act, Huge Effect: Algorithmic Sources of Publication Bias in Political Science Research</td>
<td>Alrik Thiem</td>
<td></td>
</tr>
<tr>
<td>Time</td>
<td>Location</td>
<td>Event</td>
<td></td>
<td></td>
</tr>
<tr>
<td>--------</td>
<td>----------------</td>
<td>----------------------------------------------------------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15:30</td>
<td>Saal Friedrich</td>
<td><strong>STATE-OF-THE-ART</strong> Systematic observation of human behavior from a methodological perspective (M. Teresa Anguera) Chair: Daniel Oberski</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Schiller</td>
<td><strong>STATE-OF-THE-ART</strong> Addressing Treatment Non-adherence in Randomized Experiments (Stephen West) Chair: Steffi Pohl</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16:00</td>
<td></td>
<td><strong>Coffee Break</strong> (Hotellobby)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16:30</td>
<td>Saal Friedrich</td>
<td><strong>EAM Members Meeting</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Pre-Conference-Workshops
In this workshop participants will learn the basics of Dynamic (Multilevel) analysis of intensive longitudinal data, which is the basis for Dynamic SEM (DSEM). These models are also what is used in the background for creating dynamic network models in Psychology.

We start by covering what intensive longitudinal data is, and why these data and the single subject and multilevel dynamic models for analyzing them are of interest. The single subject analyses can be used to model the dynamics within a particular individual over time, such as the reciprocal effects of positive affect and worrying may have on each other. The multilevel extensions of these models can be used to model the individual within-person dynamics at level 1, and the individual differences in these dynamics at level 2. We will discuss the basics of univariate and multivariate single subject autoregressive models, and the multilevel extensions of this model. We discuss various software options (in and outside of R) and their benefits and limitations, and give a live demonstration of the models in the DSEM module of Mplus v8. This includes a hands on part in R (bring a laptop with R; or a laptop with Mplus v8 if the participant has a license).

After the basics we move on to discuss further extensions of the models, such as predicting individual differences in dynamic effects with other level 2 variables, using the individual differences as predictors of level 2 variables, and accounting for measurement error with latent variables. We end with a discussion of potential caveats and pitfalls, recent developments, and what developments to expect for these models and data in the future.
WORKSHOP Multilevel Structural Equation Modeling with Lavaan
Instructor: Yves Rosseel

Tuesday, July 24, 09:00 - 13:00
IAAC - SR2

The aim of this half-day workshop is to provide an introduction to multilevel structural equation modeling (SEM). After a brief history, and an overview of the different frameworks, we will focus on the two-level within/between approach that is most commonly used in the applied literature. Special focus will be given to the status/meaning of latent variables in a multilevel setting, and the distinction between observed and latent covariates. Several examples will be discussed, including the setup in lavaan. Finally, we will discuss several alternative approaches to multilevel SEM, and explain when they should be used.
This workshop will introduce the participants to null hypothesis significance testing and its role in the replication crisis. Subsequently, an alternative, hypothesis evaluation using the Bayes factor will be introduced. It will be elaborated what the Bayes factor is, how it can be applied and should be interpreted. There will be attention for Bayesian updating (an alternative for power analysis), Bayesian (conditional) error probabilities, limitations of the approach, and software with which the Bayes factor can be computed.
This half-day workshop will offer a glimpse behind the scenes of SEM software. In the first part, it will be demonstrated how the model parameters in a SEM are estimated. We will set up a small function (in R) that takes a parameter vector as input, and computes the value of a suitable discrepancy function. We will then exploit the built-in optimizers of R (optim, nlminb) to find the best fitting parameters. In the second part, we will discuss the concept of information matrices, and we will explain how both standard and robust standard errors can be computed. Finally, in the third part, we will show how the so-called robust (Satorra-Bentler, Yuan-Bentler) test statistics are computed. All of this will be demonstrated with base R. This workshop aims to deepen your understanding of how Structural Equation Modeling works.
WORKSHOP Theory and analysis of conditional and average causal effects

Instructor: Rolf Steyer

Tuesday, July 24, 14:00 - 18:00
Multimediazentrum - HS E028

This short course is an introduction to the stochastic theory of causality, which is a generalization of the theory of causal effects in the tradition of J. Neyman and D. B. Rubin. In the course I will present the stochastic theory of causal effects and show how to use EffectLiteR for the analysis of conditional and average total effects.

Contents

• Motivation: Simpson’s paradox, non-orthogonal ANOVA
• The scope of the theory: random experiments
• The mathematical structure of causal models: causality space
• True outcome variables, average and conditional causal effects
• Prima facie effects
• Sufficient conditions for unbiasedness
• The role of randomization and other design techniques and strategies of data analysis
• Estimating and testing average and conditional total effects via structural equation modeling (Applications using EffectLiteR) in some empirical examples
Poster
Poster Session 1

Wednesday, July, 25, 14:00 - 15:00
Foyer Saal Friedrich Schiller

Plan

A1 Algebraic Formulas for Data Integration in Measurement of Ambivalence
SLAWOMIR PASIKOWSKI

A2 Analyzing the Reliability and Validity of the Gender Role Conflict Scale-Short Form in German
NIKOLA KOMLENAC, Heidi Siller, Harald R. Bliem, Margarethe Hochleitner

A3 Bayesian Heterogeneous Latent Growth Curve Model with a Semi-Parametric Spline Variance Function
MARK ANDERSON, Holger Brandt

A4 Concreteness-Abstractness Issue in Measuring Levels of Epistemological Understanding
NATALIA ZYLUK, Karolina Karpe, Mikolaj Michta

A5 Discovering Structures in Longitudinal Data: An Integrated Network Analysis and SEM Approach
ALEXANDRU AGACHE, Lilly Bilhler, Birgit Leyendecker

A6 Examination of IRT Model in Detection of Response Styles
QIANQIAN FAN, Moon-Ho Ho

A7 Generalizability in Research in Psychology and Economics
LUKAS RÖSELER
A8 Latent Growth Curve Evidence for Need Supportive and Thwarting Styles As Distinguishable Constructs
CARME VILADRICH, Ariadna Angulo-Brunet, Jaume Cruz

A9 Latent Transition Analysis and Subjective Marker to Measure the Transition to Adulthood
ANGELA SORGENTE, Margherita Lanz, Semira Tagliabue, Joyce Serido, Soyeon Shim

A10 Managerial Research: Qualitative and Quantitative Approaches to the Social Reality
SLAWOMIR BANASZAK

A11 Multidimensional Cross-Recurrence Analysis – A Method for Correlating Multivariate Time-Series
SEBASTIAN WALLOT

A12 Nonresponse in Income Data Using Linked Data
NADINE BACHBAUER

A13 Performance of Techniques to Correct Publication Bias Based Only on Significant Results
DESIRÉE BLÁZQUEZ, Juan Botella, Manuel Suero

A14 Psychological Needs, Self-Determined Motivation and Engagement in Athletes: A SEM Model Approach
CONSTANTINO ARCE, Cristina de Francisco, Elisa Isabel Sánchez-Romero, María del Pilar Vílchez

A15 Review and Validation of Dyadic Adjustment Scale in a Sample of Spanish Pregnant Women
F. JAVIER DEL RÍO, Manuel Antonio García Sedeño, Laura Verdugo, F. Javier Ruiz

A16 Review and Validation of Rosenberg Self-Esteem Scale in a Sample of Spanish Pregnant Women
F. JAVIER DEL RÍO, Manuel Antonio García Sedeño, Ana María Cuevas, F. Javier Ruiz

A17 Review and Validation of the Female Sexual Function Index (FSFI) in a Sample of Spanish Pregnant Women
F. JAVIER DEL RÍO, Manuel Antonio García Sedeño, Yolanda Sánchez, F. Javier Ruiz

A18 The Effect of Incentives in Survey Experiments on Panel Stability and Sample Composition
TOBIAS RETTIG, Barbara Felderer, Ulrich Krieger

A19 Tools to Assess the Measurement Properties of Quality of Life Instruments: A Meta-Review
Sonia Lorente, Josep-Maria Losilla, CARME VILADRICH, Jaume Vives

A20 Use and Usefulness of Individual Scores in Dynamic Factor Models in Case of Misspecified Models
KATINKA HARDT, Martin Hecht, Manuel C. Voelkle
A21 Using Timing Data to Model Item Omission: Estimation and Implications for Implementation
KEVIN HOPPE, Esther Ulitzsch, Steffi Pohl

A22 Validity of Single-Item Measures to Assess Self-Determination Theory Constructs in the Elderly
ARIADNA ANGULO-BRUNET, Carme Viladrich, Susana Pallares, Yago Ramis, Marta Borrucco
The aim of the presentation is to reveal the properties of algebraic formulas used in research on ambivalence and ambiguity, as well as the properties of models which are created using these formulas. Investigation of ambivalence and ambiguity issues requires using instruments sensitive to the coexistence of opposite tendencies, first at the stage of gathering empirical data and then at the stage of modeling the tension between these tendencies. The presentation provides illustrations of the main solutions in this area. It will enable us to characterize the methodological conditions for the application of algebraic formulas and then present methods of their evaluation. The results of research on these algebraic formulas will be presented separately. The conducted simulations made it possible to fit mathematical functions, and to describe the advantages and drawbacks of each of the formulas. Based on the results an original formula was proposed. It meets the conditions required for this type of formulas, and it has distinctive properties that make it more robust to the defects and distortions observed in the case of other formulas, and the results they provide. The problem of artefactuality with respect to data yielded by the algebraic formulas, as well as the possibility of using these formulas in studies on methodological orientation will be discussed.
Even though the Gender Role Conflict Scale has been validated in numerous studies, the Gender Role Conflict Scale-Short Form (GRCS-SF) has not been subject to such rigorous testing yet. The current study was conducted in order to test whether the GRCS-SF is reliable and valid in an Austrian sample of men with diverse socio-demographic backgrounds. The GRCS-SF was translated into German and handed to a sample of men in Austria. The study sample comprised 127 in-patients of a university hospital. The sample’s mean age was 59.5 (SD = 14.6, range 23 – 84) years. A confirmatory factor analysis (CFA) was performed to test the construct validity. The convergent validity was assumed if the GRCS-SF correlated with the Sexual Performance Belief Scale (SPBS) that assesses whether people believe in masculine gender role norms concerning "sexual performance". Normal distribution could not be assumed for the GRCS-SF scales. Therefore, the maximum likelihood method (ML) and the Bollen-Stine Bootstrap method with 500 samples were used to estimate the fit indices of the CFA. Spearman correlations were calculated to test convergent validity. The CFA provided support for the four factor structure of the GRCS-SF ($\chi^2(98) = 115.9; p = .126$; RMSEA = .06; SRMR = .08; GFI = .88). The internal consistency was satisfactory (Cronbach’s $\alpha = .65 – .78$). All four scales correlated significantly with the SPBS to a small or moderate extent ($r = .18 – .42$). These findings support the claim that the GRCS-SF is reliable and valid in a diverse sample. This study adds first findings that recommend the use of the GRCS-SF to reliably and validly assess men’s gender role conflict in Austria.

KOMLENAC, NIHOLA 1
Siller, Heidi 1
Bliem, Harald R. 2
Hochleitner, Margarethe 1

A2 Analyzing the Reliability and Validity of the Gender Role Conflict Scale-Short Form in German

1 Gender Medicine Unit, Medical University of Innsbruck, Austria
2 Faculty of Psychology and Sport Science, University of Innsbruck, Austria

nikola.komlenac@i-med.ac.at
The heterogeneous growth curve model (HGM; Klein Muthén, 2006) and the robust version of the HGM (HGM-R, Brandt Klein, 2015) are statistical models to account for heterogeneity within longitudinal growth rates. Specifically, these models incorporate a heteroscedastic residual structure for the latent slope factor. These modeling approaches successfully build upon growth curve mixture modeling literature and other traditional growth curve modeling specifications. However, the approximation of heteroscedastic growth rates in the HGM so far is modeled using a quadratic function, but other functional forms may exist within the data and hence the model may be misspecified (Hedeker, Mermelstein, Demirtas, 2012). The spline heterogeneous growth model (S-HGM) is an extension of both the latent growth curve model and the HGM that can flexibly model any type of heteroscedasticity for the residual structure without a priori specification of its functional form. The current model expands on the previously available estimators by incorporating the model in a Bayesian framework. In the present simulation study, the performance of S-HGM is investigated for different types of heteroscedasticity and it is compared to results for the HGM. One of the main advantages of this model is the identification of misspecified variance functions in other models, which is difficult to detect by model fit indices, and hence this method allows the researcher to inspect the appropriateness of certain functions. To demonstrate practical applicability, the S-HGM is illustrated in a longitudinal data set concerning social media use within a representative sample of the Dutch population.
The aim of this poster is to describe the process of modification of the research tool designed for measuring the development of personal epistemology — Standardized Epistemological Understanding Assessment (SEUA; Żyluk et al., 2016). SEUA was constructed as an improved version of the instrument originally proposed by Kuhn et al. (2000). The main changes to the initial version included extending the list of test items, a new administration procedure (a structured interview) and an introduction of a quantitative scoring method. SEUA was proved to be more reliable instrument than its predecessor, however, further changes were necessary to obtain better reliability and easier to administer form. Interactive manner of research with SEUA in interview setting allowed us to observe that test items used in this tool can be divided into two groups: abstract (these that refer to general class of objects, properties or beliefs) and concrete (these that refer to known, particular objects, properties or beliefs), which were approached differently by subset of our participants. In their cases, the inability to suppress personal preferences in responding to concrete items threatened the validity of the tool, as the instrument may measure preferences rather than epistemological beliefs in this situation. SEUA was therefore modified to create a full-abstract version (SEUA-A). Both versions were administered in an online form. The performance of two versions of the tool was compared: the extended, 25-items tool presented in Żyluk et al. (2016), and a newly developed, abstract-only version. The results combined from this and previous studies inform that online SEUA-A with internal consistency above 0.7 threshold for every test domain is the most reliable, an easy to administer version of the tool.
We propose an integrated network analysis and structural equation model (SEM) method to explore multivariate network structures for the case of sparse longitudinal data. The method aims at quasi-regular spaced longitudinal data, where the number of repeated measurements available per analysis unit is small. In recent years, network theory has gained increasing popularity in psychology, especially for observations that do not fit well in the traditional conceptualization of latent variables (e.g., networks of interacting symptoms of different disorders; Borsboom, 2017). Current approaches in longitudinal network analysis handle the problem of estimating structures involving high numbers of parameters using multilevel autoregressive models and regularization techniques (Epskamp et al., 2018). These methods require intensive time series data. In contrast, the random and fixed effects SEM (Allison, 2009; Bollen Brand, 2010) can be estimated with 3 or 4 repeated measurements and are flexible enough to handle different modeling assumptions. For instance, different variants of random and fixed effects models can be compared with fit statistics, time invariant unmeasured confounders can be controlled for, and tests of lagged, reciprocal causation can be incorporated (Allison, Williams, Moral-Benito, 2017). The present paper illustrates how the existing network and SEM analysis techniques can be integrated across several analytical steps. These steps include the investigation of trajectories and stationarity assumptions, the estimation of undirected longitudinal networks based on unregularized or regularized partial correlations, and the estimation of directed networks based on dynamic fixed effects SEM. We illustrate this method using data on teacher-child interactions observed within 177 preschool classrooms across four repeated cycles. Plans for the future include simulation work and the development of an add on for existing R packages.
In psychological research, rating scales were widely used to measure individuals’ attitudes and behaviors. However, it was found ratings may be affected by Response Styles (RS) (Paulhus, 1991). Therefore, various approaches were proposed to identify response styles (RS), such as count procedure, Representative Indicators for Response Styles (RIRS), IRT models, etc. (Vaerenbergh Thomas, 2012). Among these approaches, IRT models received most research interest. Nevertheless, there were some limitations in previous research. For example, IRT models were mainly employed to detect ERS (extreme RS; the tendency to use the highest and lowest response categories); ERS was simulated from a normal distribution with mean of 0 and sd of 1, which consisted of 2 RS – ERS and MLRS (mild RS; the tendency to avoid the highest and lowest response categories); previous research did not examine parameter recovery of IRT models in detection of RS, etc. Therefore, the present study will address these limitations, and explore how to implement a widely used IRT model - Multidimensional Nominal Response Model (MNRM) with a SEM software to identify ERS, ARS (acquiescence RS; the tendency to agree with items), and MRS (mid-point RS; the tendency to use the middle response category). Data will be simulated from MNRM. Impact of four factors will be examined: 1. sample size (300 vs. 800); 2. number of items per content factor (10 vs. 20); 3. RS (NRS–no RS, ARS-the tendency to agree with items, and MRS-the tendency to use the middle response category); 4. correlations between items (items from content factor only vs. 15 items from 5 uncorrelated factors). Results of simulation studies will be focused on: 1. Model fit indices in true/incorrect models; 2. Pearson and Spearman correlation between true and estimated content factor/RS scores; 3. Parameter recovery (category intercepts) in true/incorrect models. Simulation results and application of IRT models in SEM framework will be discussed.

FAN, QIANQIAN 1
Ho, Moon-Ho 1

A6 Examination of IRT Model in Detection of Response Styles

1 Nanyang Technological University, Singapore
qfan002@e.ntu.edu.sg
Even though generalizability plays a central role in research in psychology and economics, researchers have yet to agree on what generalizability of results actually means. Moreover, researchers have yet to come to an agreement on whether generalizability (i.e., external validity) is more or less important than internal validity. On the basis of the distinction between effects application and theory application by Calder, Phillips, and Tybout (1981), I will suggest solutions to previously observed problems: These problems include the artificial nature of laboratory settings, the use of student samples, possible trade-offs between internal and external validity, theoretical and empirical views on external validity, and the question of the extent to which theories or empirical findings possess external validity. I argue that advocates from both (effects and theory application) views on external validity criticize each other’s positions but disregard relevant underlying philosophical assumptions. I draw the conclusion that a clear formulation of a researcher’s point of view regarding induction versus deduction (and effects vs. theory application) in a specific study is essential for avoiding misunderstandings because the conflict between researchers who prefer effects application and those who prefer theory application needs to be addressed on a philosophical level rather than on merely a methodological one.
According to Self-Determination Theory, the same authoritative individual can adopt a need supportive or a need thwarting interpersonal communication style depending on circumstances. To date, basic psychometric evidence for distinguishable styles such as factors showing discriminant validity and non redundant relationships with external variables has been provided. The aim of this research is to add a new piece of evidence comparing the trajectories of both styles using parallel latent growth modeling. Seven secondary school physical education teachers who volunteered to participate in the area of Barcelona (Spain) were assigned into the control arm of a larger study and their 718 students pertaining to 27 groups reported their perception of teacher’s interpersonal style of autonomy support and thwarting four times during the course 2016/17. Results showed longitudinal scalar measurement invariance except for the intercepts of two items tapping autonomy support which decreased across waves. All correlation coefficients between supportive and thwarting styles lay between -.47 and -.54 supporting the discriminant validity of both concepts. Parallel latent growth curve analysis showed an acceptable fit ($\chi^2_{SB} = 697.74$, df = 436, CFI = .94, TLI = .93, RMSEA = .03, SRMR = .05). Autonomy support and thwarting trajectories showed mixed evidence with the same latent slopes ($\beta_s = .14$, $p = .26; \beta_t = .15$, $p = .13$) but sizeable different latent intercept variances ($\sigma^2_s = 0.36 \sigma^2_t = 0.10$). Psychometrically speaking, these results constitute a fresh piece of evidence of the distinguishability of autonomy supportive and thwarting styles in a part of the intended interpretation of these constructs less explored to date and with particular relevance for follow-up assessments. Funding was provided by grant MINECO/DEP2014-52481-C3-1R from Spanish Government.
The expression “transition to adulthood” refers to the passage from adolescent dependence to adult autonomy historically marked by the sequential achievement of five observable life events: completing school, entering into the labor market, moving out of the parental home, forming a partnership, and becoming a parent. Social scientists have paid this unfolding sequence of events a considerable amount of attention over the past decade mainly because of its changing duration and increasing complexity. Before the 1970s the transition to adulthood was “early, contracted, and simple”, while today it is “late, protracted, and complex” (Billari Liefbroer, 2010). However, most of the techniques used previously to measure this transition (survivor analysis, transition rates, ‘mirrored’ survivor functions, non-recurrent sequences, sequence analysis, and entropy analysis) cannot capture the complexity of today’s transition phenomena. Specifically, these techniques can manage only one event at time, are unable to treat events as repeatable, or require multiple cohort-comparisons to offer useful information. Furthermore, the techniques suffer from sequencing fallacies and an inherent subjectivity in the way they select the trajectories to be studied. Moreover, scholars suggest that nowadays to fully transition into adulthood the individual must also begin to perceive himself as an adult (Reitzle, 2006).

The current study aims to establish a new method for measuring the transition to adulthood, one that involves (1) using 5 objective markers and 1 subjective marker and (2) applying the Latent Transition Analysis technique, which is characterized by its person-centered approach and its longitudinal design. This new method has been tested on a sample of 913 young adults in the United States aged 20-25 at time 1. The results show that there are three different configurations of adulthood marker and multiple transition sequences from one configuration to another over time.
Managers are the very important social category in modern societies. They are also an important object of social research. The author examines managers as a social class. Some remarks on the structure of the contemporary societies are revealed. The author emphasizes the need for managerial research as well. The research has to be projected as a multidisciplinary project of cognition in roles in contemporary economies and societies. The author used a mixed research strategy, i.e. he used both an in-depth interview and a collection of data collection techniques. The effects of this study do not give a full picture of managerial identities, it is rather a sketch for a managers’ portrait. First of all, managers do not appreciate the academic education. Rather, they tend to recognize that leadership qualities are natural in individuals. Also in the case of managing people, they prefer non-modern ways of motivating and evaluating employees. At the level of methodological reflection, the author looks for strengths and weaknesses of individual research techniques. And he also tries to develop his own model of managers’ research.

BANASZAK, SLAWOMIR

A10 Managerial Research: Qualitative and Quantitative Approaches to the Social Reality

1 Adam Mickiewicz University, Poland
s.ban@amu.edu.pl
Multidimensional Cross-Recurrence Quantification Analysis (MdCRQA) is an extension of Multidimensional Recurrence Quantification Analysis (MdRQA), which allows to quantify the (auto-)recurrence properties of a single multidimensional time-series. MdCRQA extends MdRQA to bi-variate cases to allow for the quantification of the co-evolution of two multidimensional time-series (coupling of arousal between two participants as measured by heart rate, breath-rate and skin conductance). Moreover, it is shown how a Diagonal Cross-Recurrence Profile (DCRP) can be computed from the MdCRQA output that allows to capture time-lagged coupling between two multidimensional time-series. The core concepts of these analyses are described, as well as practical aspects of their application.
Nonresponse is a phenomenon that primarily comes up in surveys, especially for sensitive questions. For example, a not negligible amount of respondents refuse to provide individual income. Reason for this refusal of information can be memory errors as well as social desirability. People at the margins of income distribution, in particular, tend not to indicate their income or to state it incorrectly. The used database contains the data of the National Education Panel Study of Starting Cohort 6 linked to the administrative data of the Institute for Employment Research at the Federal Employment Agency (NEPS-SC6-ADIAB). The question arises whether the missing income data is randomly distributed. Whether a Missing at random (MAR) process or a not missing at random (NMAR) process exists concerning the gaps in the survey income information. To analyse this research interest, a model is estimated that includes the administrative income data as well as characteristics typically used to impute income data as independent variables. Imputation methods are often used to close up the gaps of missing data and to increase the capabilities for analysis. Subsequently, the aim is to impute the missing incomes in the survey data and to compare the distribution with the administrative income data in order to check the imputation procedure for validity. This research project is still in progress partly caused by the upcoming data release of NEPS-SC6-ADIAB.
Introduction: A recent strategy to address publication bias in meta-analysis focuses only on the primary studies with significant results. The underlying assumption is that, within the statistically significant studies, the conditional probability of being published is independent of the p-value. The techniques based on this new strategy model the p-values associated to the significant primary studies in a meta-analytic dataset and obtain an estimate of the parameter (the value tested in the null hypothesis) for the best fit to a uniform distribution of the normalized p-values. There have been proposed at least six different techniques under this perspective and the differences between them are related to two main aspects: the p-values transformation and the estimation procedure. In the present work, we use simulated and real databases to assess which transformation and estimation process obtains better parameter estimates by testing the performance of these main procedures under certain factors that have demonstrated to influence the implementation of the underlying strategy, but have not been studied systematically.

Method: The performance of the six different procedures is measured in terms of mean error, mean square error, and confidence interval coverage given a specific alpha level. The factors that may influence the performance of the ‘only significant results’ techniques are the presence of outliers, the parametric effect size, and the sample size of the primary studies with significant results. Firstly, we carry out a simulation study assuming a fixed effect model. And secondly, we use real datasets from Registered Replication Reports initiatives to test the conclusions obtained previously in the simulation study.
The theory of self-determination is based on a conceptual model that examines human motivation, and within it, how the satisfaction of basic psychological needs influences the different degrees of self-regulation of a person (the most self-regulated degrees are achieved for intrinsic reasons of the subject). The lack of satisfaction of basic psychological needs has been pointed out as one of the main reasons for the burnout and other associated symptoms in the sport context. One aspect not studied is the engagement from a self-determination perspective. Engagement is the opposite construct to burnout within a continuum. In this study, we investigated the relationship between the satisfaction of basic psychological needs and the athlete engagement taking motivation as a mediator variable. Our main hypothesis was that the satisfaction of the basic psychological needs has two effects on the engagement, one direct positive effect and one indirect effect through motivation. The satisfaction of basic psychological needs in addition to high levels of self-determined motivation will increase the engagement of the athletes and in addition to low levels of self-determined motivation will decrease the athlete engagement. In the study participated 426 female and male athletes aged between 13 and 28 years that were practicing some of 24 different sports modalities. Three specific questionnaires developed in the sport context were administered. One to measure the satisfaction of basic psychological needs, one to measure the self-determined motivation and other to measure athlete engagement. The data analysis was performed from the perspective of the SEM models. The satisfaction of basic psychological needs was taken as an exogenous variable, the self-determined motivation as a mediator variable and the athlete engagement as an endogenous variable. The results offered evidence in favor of the hypothesis. The satisfaction of basic psychological needs has a positive direct effect of .29 (p < .001) on the athlete engagement and an indirect effect via motivation, being positive when the levels of motivation are high and negative when are low. The total effect is of .5579 (p < .001) and the R-square value equal to .41. Measures of the overall fit of the model were GFI = .91, CFI = .91, RMSEA = .11 (90
A good relationship is an important factor for the mental health of people, since their absence is related to the increased risk of depression, anxiety and general health problems. For this reason, it is essential to have good instruments that evaluate the dyadic adjustment. One of the instruments available for this is the dyadic adjustment scale. The aim of this study was to review the reliability and the factorial structure of the scale in a sample group of Spanish pregnant women. 152 of them participated, their ages span in between 18-43 years old (average, 32.15; standard deviation, 4.987). FACTOR was the programme used to carry out the statistical analysis. The correlation item-total is good (.310-.596). The reliability of the scale was high, in both the Cronbach’s alpha (.8733) and the McDonald Coefficient Omega (.8734). The KMO test (.8028), Bartlett test (.000) and the determinant of the matrix (.040) assure that the adaptation of the factorial analysis fulfilment. For the analysis we used the polychoric correlation matrix. Therefore, the original version of the factorial structure is confirmed, explaining the 59.90
Self-esteem is a much studied concept, and one that has a direct relationship with the general well-being of the person, being, therefore, a good indicator of mental health. Previous research has shown the inverse relationship between self-esteem and depression. Pregnant women, as a result of the physiological and hormonal changes of their state, usually have a mood and a concept of oscillating themselves. For that reason, it is important to have a questionnaire that adequately evaluates your self-esteem. One of the instruments available for this is the Rosenberg self-esteem scale. The aim of this study was to review the reliability and the factorial structure of the scale in a sample group of Spanish pregnant women. 152 of them participated, their ages span in between 18-43 years old (average, 32.15; standard deviation, 4.987). FACTOR was the programme used to carry out the statistical analysis. The correlation item-total is good (.303-.655). The reliability of the scale was high, in both the Cronbach’s alpha (.878) and the McDonald Coefficient Omega (.879). The KMO test (.771), Bartlett test (.000) and the determinant of the matrix (.000) assure that the adaptation of the factorial analysis fulfilment. For the analysis we used the polychoric correlation matrix. We found a factorial structure of two factors that explains 69.97

DEL RÍO, F. JAVIER ¹
García Sedeño, Manuel Antonio ¹
Cuevas, Ana María ¹
Ruiz, F. Javier ²

A16 Review and Validation of Rosenberg Self-Esteem Scale in a Sample of Spanish Pregnant Women
¹ University of Cadiz, Spain
² Centro de Salud San Benito, Spain

fjdelrio@iaosexualgia.com
DEL RÍO, F. JAVIER ¹
García Sedeño, Manuel Antonio ¹
Sánchez, Yolanda ¹
Ruiz, F. Javier ²

A17 Review and Validation of the Female Sexual Function Index (FSFI) in a Sample of Spanish Pregnant Women

¹ University of Cadiz, Spain
² Centro de Salud de San Benito, Spain

fjdelrio@iasexologia.com

Pregnancy is considered one of the most critical stages in a woman’s life. The most recent studies about women’s sexuality during the pregnancy shows that the ability to reach an orgasm decreases and the likelihood of dyspareunia increases, primarily during the second half of the pregnancy; likewise, due to the fear over the consequences in the normal progress of the pregnancy, pregnant women refrain from carrying out some coital positions. These changes start taking place from the first term of the pregnancy and last until the end of the gestation. One of the tools offered to assess women’s sexual response is the Female Sexual Function Index. The aim of this study was to review the reliability and the factorial structure of the questionnaire in a sample group of Spanish pregnant women. 152 of them participated, their ages span in between 18-43 years old (average, 32.15; standard deviation, 4.987). FACTOR was the programme used to carry out the statistical analysis. The correlation item-total is good (.437-.857). The reliability of the questionnaire was high, in both the Cronbach alpha (.960) and the McDonald Coefficient Omega (.962). The KMO test (.916), Bartlett test (.000) and the determinant of the matrix (.000) assure that the adaptation of the factorial analysis fulfilment. For the analysis we used the polychoric correlation matrix. Therefore, the original version of the factorial structure is confirmed, explaining the 89.28
We examine the effect of differential respondent incentive payouts to panelists as a result of survey experiments on panel stability and sample composition.

Incentives have been successfully used as a method to increase or stabilize participation in longitudinal studies. Incentives have been paid out to all sample members or have been targeted towards a specific subgroup such as reluctant respondents.

In this presentation however we examine the effects of incentives that are used as rewards in survey experiments conducted as part of a panel study. Survey experiments are increasingly used by social scientists to generalize laboratory findings. Little is known about the methodological effect of conducting such experiments in general population surveys. In this study we focus on the use of respondent incentives in such experiments.

In the German Internet Panel, an online panel of the general population based on a probability sample, three survey experiments have been conducted in subsequent waves offering monetary incentives to a select group of respondents. Payouts went to a random group of respondents and their value varied between €5 and €80.

This setting creates the unique opportunity to examine the effects of such differential incentives on panel attrition and sample composition controlling for incentive value. Receiving incentives could boost panel participation, not being among the winners could increase attrition. In addition, the effect of incentives could be different across specific subgroups of the sample and thus inducing a bias in sample.

Results will add to the body of knowledge on the use of differential incentives and conducting survey experiments in longitudinal studies.
A19 Tools to Assess the Measurement Properties of Quality of Life Instruments: A Meta-Review

Lorente, Sonia¹
Losilla, Josep-Maria¹
Viladrich, Carme¹
Vives, Jaume¹

¹ Universitat Autònoma de Barcelona, Spain

Carme.Viladrich@UAB.Cat

Systematic reviews of measurement properties critically appraise the properties of instruments measuring a certain construct, being useful to provide evidence for the selection of the best instruments in clinical practice and research (Mokkink et al., 2009). To improve the quality of systematic reviews it is recommended to apply specific tools to assess the measurement properties of these instruments. However, depending on the measurement standards upon which these tools are developed, the approach to appraise the measurement properties of health instruments may vary, leading to different conclusions. Because systematic reviews of health instruments are essential to select instruments for research and clinical practice (Terwee et al., 2016), the present meta-review aims to: 1) identify systematic reviews assessing the measurement properties of instruments evaluating Health Related Quality of Life (HR-QoL) (Healthy People, 2010); 2) identify the tools applied to assess the measurement properties of HR-QoL instruments; 3) describe the characteristics of the tools applied to assess the measurement properties of HR-QoL instruments (validity, reliability, feasibility, etc.); and 4) compare the measurement standards upon which these tools were developed. A descriptive analysis of the frequency of use of tools applied to evaluate the measurement properties of HR-QoL instruments is presented; the similarities and differences among the Standards for Educational and Psychological Measurement (AERA, APA NCME, 1999, 2014), Attributes and Criteria to assess Health Status and Quality of Life Instruments (Aaronson et al., 2002; Reeve et al., 2013) and Health Status Measures in Economic Evaluation (Standing Group on Health Technology) (Brazier, Deverill, Green, Harper, Booth, 1999; Brazier et al., 2007.) are described, and some recommendations about the methodological and research applications are made.
Longitudinal dynamic factor models facilitate the understanding of mechanisms underlying processes of change. Whereas most studies focus on optimal model fit and optimal estimation of model parameters, comparably less attention has been paid to the use of individual scores. Individual scores are useful if individuals and individual developmental trajectories are of interest to researchers or practitioners, for instance, if decisions about an individual need to be made. Several approaches to obtain individual scores are available and a variety of criteria has been applied to evaluate their performance. However, performance of different individual score methods in the presence of model misspecification when tracking individuals over time has not been investigated thus far. With our poster, we will contribute our share to closing this gap. In a simulation study, we illustrate differential performance of four widely used approaches to obtaining individual scores in misspecified models: the regression method, the Bartlett method, the sum score and the Kalman filter. As these approaches differ with regard to the model parameters that are entering their computation, they are differently sensitive to different locations of misspecification. In our simulation we use autoregressive latent factor models to illustrate individual score method performance if a) residual covariance among indicators is present but not modeled, b) loadings are different but modeled as being equal, and c) the dynamics as modeled by an autoregressive factor model of order one obey a different model. As a general result, the sum score shows remarkably low performance across all conditions and may therefore lead to severely misleading individual trajectories. In contrast, differences in performance between the regression method, the Bartlett method and the Kalman filter become obvious only under specific conditions. Implications of these results will be discussed.
Item omission is often nonignorable and poses a threat for the reliability of proficiency estimates in large scale assessments such as the Programme for the International Assessment of Adult Competencies (PIAAC). Computerized test administration grants researchers access to timing data containing additional information on participants’ test taking behavior that may help modeling the missing process. In PIAAC, item nonresponses occurring after an exposure time of five seconds are scored as incorrect. Ulitzsch, Pohl, and von Davier (in review) propose an alternative approach, the SA+O model, in which they jointly model item responses and nonresponses as well as associated timing data for both types of responses. The model allows for investigating the missing process underlying item omission as well as accounting for it. However, a major challenge this approach faces is the high sparseness of available timing data due to item responses and nonresponses being mutually exclusive. Even omission rates as high as 25
According to self-determination theory, supportive interpersonal styles are determinant on the satisfaction of basic psychological needs for autonomy, competence and relatedness, which in turn can impact on subjective vitality, forming a three-concept chain. In order to support the use of single-item measures for these concepts, the aim of this study is to provide validity evidence through testing the replicability of the expected chain and also to provide test-retest reliability for each measure. Embedded in a larger study, seventy-one elderly adults who were attending to group physical activities answered the relevant seven single-item questions in a five points Likert scale for two times in a lapse of five weeks. Considering small sample size and floor and ceiling effects, descriptive correlation networks based on polychoric correlations were used to test the expected three-concept chain. Data gathered in wave 1 yielded favorable evidence for the chain showing moderate correlations (.37-.68) between adjacent points and weaker correlations (.26-.41) between distant points. In contrast, mixed evidence was obtained in wave 2, mainly due to unexpectedly low correlations (.13-.43) between basic psychological needs and the other variables in the chain. Regarding test-retest reliability, supportive interpersonal style single-items showed higher values (.50 -.58) than basic psychological need satisfaction (.31 -.45) and vitality (.37). On the basis of the evidence currently available, single-item measures of supportive interpersonal styles and vitality are promising measures whereas more validity and reliability evidence should be gathered for basic psychological need satisfaction constructs. This research was supported by the Spanish Ministerio de Economía y Competitividad (DEP2014-52481-C3-1R).

ANGULO-BRUNET, ARIADNA ¹
Viladrich, Carme ¹
Pallares, Susana ¹
Ramis, Yago ¹
Borrueco, Marta ¹

A22 Validity of Single-Item Measures to Assess Self-Determination Theory Constructs in the Elderly

¹ Universitat Autònoma de Barcelona, Spain

ariadna.angulo@uab.cat
Poster Session 2

Thursday, July, 26, 14:00 - 15:00
Foyer Saal Friedrich Schiller

Plan

<table>
<thead>
<tr>
<th>B1</th>
<th>B2</th>
<th>B3</th>
<th>B4</th>
<th>B5</th>
<th>B6</th>
<th>B7</th>
<th>B8</th>
<th>B9</th>
<th>B10</th>
<th>B11</th>
<th>B12</th>
<th>B13</th>
<th>B14</th>
<th>B15</th>
<th>B16</th>
<th>B17</th>
<th>B18</th>
<th>B19</th>
</tr>
</thead>
</table>

B1 ARIMA Models in Protest Events Analysis: the Self-Generated Character of Social Protests
RODRIGO ASÚN, Claudia Zúñiga, Cristobal Villalobos, Camila Ortiz

B2 About the Simmetry/Asimmetry Between Qualitative and Quantitative Approaches in Mixed Methods
M. TERESA ANGUERA, Oleguer Camerino, Marta Castañer, Pedro Sánchez-Algarra, Anthony J. Onwuegbuzie

B3 Adaptation of the Multidimensional Scale Body Self-Relations Questionnaire for the Adolescent Population
MARÍA PAULA FERNÁNDEZ GARCÍA, Carmen González González-Mesa, Soraya Coballes del Río

B4 Assessing Consistency in Single-Case ABAB Phase Designs
RENÉ TANIOUS, Tamal Kumar De, Bart Michiels, Patrick Onghena

B5 Can We Take Advantage of Textual Material Through Indirect Observation? Proposal of ‘Quantitizing’
M. TERESA ANGUERA, Mariona Portell, Salvador Chacón-Moscoso, Susana Sanduvete-Chaves

B6 Checklist for Reporting Observational Studies in Sports Areas: Content Validity
Salvador Chacón-Moscoso, Susana Sanduvete-Chaves, MARÍA TERESA ANGUERA, José Luis Losada, Mariona Portell, José Antonio Lozano-Lozano
B7 Detection of Response Styles with Count Procedure in Simulation Studies
QIANQIAN FAN, Moon-Ho Ho

B8 Effective and Efficient Design of a Research Vulnerable to Data Loss
MARÍA PAULA FERNÁNDEZ GARCÍA, Guillermo Vallejo Seco, Pablo Livácic Rojas, Ellíán Tuero Herrero

B9 Gender-Based Measurement Invariance of the Basque Short Version of Camir (B-Camir-R)
JONE ALIRI, Arantxa Gorostiaga, Alexander Muela, Nekane Balluerka, Goretti Soroa, Aitor Aritzeta

B10 How to Enhance Students’ Interest in Fundamentals of Methodology? A Cohort Quasi-Experimental Design
MARÍA PAULA FERNÁNDEZ GARCÍA, María Paz Suárez Coalla

B11 Investigating Chain Decisions and Convergence in Monte-Carlo Markov-Chain Estimated Analysis
ZACHARY ROMAN, Holger Brandt

B12 Psychometric Properties of Child Victimization Instruments: A Systematic Review
Flor Zaldívar-Basurto, ELENA ORTEGA-CAMPOS, Leticia De la Fuente-Sánchez, Juan García-García

B13 Sample Size Planning for Testing the Equivalence of Two Independent Simple Linear Regression Slopes
WEI-MING LUH, Jiin-huarng Guo

B14 Statistical Power of Covariance Structure Selectors for Complete Data and Missing Data Mechanisms
PABLO LIVACIC-ROJAS, Guillermo Vallejo Seco, María Paula Fernández García, Ellíán Tuero Herrero

B15 Synthesizing Single-Case Research: Nonparametric Confidence Intervals for Combined Effect Sizes
BART MICHIELS, Tamal De, Rene Tanious, Patrick Onghena

B16 Testing the Effect of the Number of Response Categories on Self-Other Agreement and Consensus
KAY BRAUER, René Proyer

B17 The Use and Disuse of the Effect Size Index in JCR Spanish Journals of Psychology
ELENA ORTEGA-CAMPOS, Leticia De la Fuente-Sánchez, Juan García-García

B18 The Use of a Bi-factor Model to Check the Internal Structure of a Scale of Attitudes
Celia Serrano, Luis Manuel Lozano, JOSE-LUIS PADILLA, Inmaculada Valor

B19 Validation of the Scenario-Based German Creative Response Evaluation at Work Scales (CRE-W)
KAY BRAUER, Christiane Hahn, René Proyer
The Protest Events Analysis (PEA), which consists of the collection, codification and systematic analysis of news recorded in the press referring to collective demonstrations during more or less prolonged periods of time, it has become in the last years, a very used and productive methodological tool for the longitudinal study of social protest cycles and the variables that affect them. However, the techniques that have been used to analyze the data produced tend to be very unsophisticated (sometimes consisting of simple visual inspection of graphs) or not fully exploiting the longitudinal nature of the available data (transversal regression models are usually applied). In addition, until now there has always been an attempt to explain the evolution of protests from events outside the protest cycles, without considering the possibility that at least part of them may be explained by self-generating processes. With the aim, so far not reported in the literature, to explore the possible self-generating character of social protests and show the productivity of applying longitudinal analysis techniques to AEP, we have used ARIMA econometric models to model the series of monthly protest events made by the Chilean student movement since the recovery of democracy in 1990 until 2014. Our results show that a substantive part - little more than 40
Mixed methods studies involve the adoption of either a qualitative-dominant or a quantitative-dominant approach, but usually the studies are frequently characterized by a lack of symmetry between the two approaches. Mixed methods studies typically focus more on qualitative than quantitative data and accordingly miss the opportunity to explore the wealth of information that a quantitative analysis of qualitative data can provide. Although it is true that some researchers apply robust statistical methods and even multiple techniques to analyze quantitative data, they frequently fail to move beyond a descriptive analysis, and consequently miss out on the opportunity to explore the richness of information within the qualitative component.

To achieve the qualitative-quantitative symmetry, it is necessary to quantitize the qualitative data and qualitize the quantitative data. Our novel proposal is the way in which we integrate or mix the two types of data. We use the sequentiality method, which takes as its starting point the annotation of the order of occurrence of all the behaviors included in a given observational dataset. This sequentiality permits the transformation of initially qualitative data into a format that can be analyzed quantitatively and robustly, achieving thus successful integration.

Several data analysis techniques that are specific to the study of sequences of behavior have a particularly important role in observational methodology due to the assignment of parameters of frequency, order, and duration to the initial qualitative data and thereby providing the necessary conditions for subsequent quantitative analysis using robust, non-standard, statistical techniques that offer highly relevant structural results.

ANGUERA, M. TERESA ¹  
Camerino, Oleguer ²  
Castañer, Marta ²  
Sánchez-Algarra, Pedro ¹  
Onwuegbuzie, Anthony J. ³  

B2 About the Symmetry/Asymmetry Between Qualitative and Quantitative Approaches in Mixed Methods  

¹ University of Barcelona, Spain  
² University of Lleida, Spain  
³ Sam Houston State University / University of Johannesburg, USA  
tanguera@ub.edu
FERNÁNDEZ GARCÍA, MARÍA PAULA 1
González González-Mesa, Carmen 1
Coballes del Río, Soraya 2

B3 Adaptation of the Multidimensional Scale Body Self-Relations Questionnaire for the Adolescent Population

1 Universidad de Oviedo, Spain
2 Colegio San Fernando, Spain

paula@uniovi.es

In 1990 Brow, Cassm Mikulka built their multidimensional scale Body Self-Relations Questionnaire for the adult population. It has been applied many times and validated many times in different countries including Spain. However, it is worth noting that it has been applied in most cases rather laxly, with little common ground in results. Many studies have put into stark relief the importance of self-image based on bodily characteristics during adolescent years and its influence in terms of self-esteem. With the objective of adapting and validating the Multidimensional scale Body Self-Relations Questionnaire for adolescents we have carried out our investigation. We have accumulated a data base of 723 subjects between nine and sixteen years of age, taken from different schools in Asturias; 398 males and 325 females. We started by carefully translating the items in the original questionnaire before revising semantic comprehension with language experts, teachers and educators. After this initial preparation we had a trial run before beginning the application of the questionnaire to our total data base of adolescents. Firstly via exploration of crucial factors involved, followed by another exploration of said factors, which partially confirmed effectiveness and comprehension, we evaluated the dimensionality of the test and its equal application and effectiveness for both sexes. The said process has led us to conclude that twenty-six items constitute between them the essential structure constituted by three factors: F1: satisfaction with corporal self-image; F2: quality capacity and fitness and F3, care taken with physical appearance. We conclude that the scale which has guided this investigation is valid for the investigation of how adolescents and pre-adolescents perceive their own self-image with reference to and based upon their physical appearance.
Single-case experimental designs (SCEDs) involve the repeated measurement of a single entity that is being exposed to different manipulations of at least one independent variable (e.g., Kratochwill et al., 2010). The application of SCEDs has risen steadily over the years (Shadish Sullivan, 2011; Smith, 2012). Alongside the number of applications of SCEDs, the number of available analytical techniques has increased, leading to intensified discussions regarding the appropriateness of these techniques (Manolov Moeyaert, 2017). Traditionally, visual analysis is the primary method for analyzing data obtained from SCEDs (e.g., Lane Gast, 2014). The What Works Clearinghouse guidelines for SCEDs (Kratochwill et al., 2010) recommend inspecting six features of the data when performing a visual analysis: level, trend, variability, immediacy of the effect, overlap, and consistency of data patterns across similar phases. However, the validity of visual analysis has been seriously questioned over the years. Harrington and Velicer (2015) concluded for example that visual analysis “is prone to bias and should not be used as a stand-alone analytical method” (p. 181). Therefore, in addition to visual analysis, various effect size measures have been proposed for the data features suggested by Kratochwill et al. However, no operationalization exists yet to quantify the degree of consistency in SCED data. In order to fill this gap, we propose several operationalizations to quantify consistency in SCEDs. We first propose two ways of operationalizing consistency as the extent to which data patterns are the same within the conditions based on a Euclidean distance measure and univariate linear regression respectively. Next, we propose an operationalization that is based on consistency as a meta-measure of the other five data aspects. We illustrate the calculation of the proposed operationalizations with three published data sets.
Indirect observation involves the analysis of textual material generated either indirectly from transcriptions of audio recordings of verbal behavior in natural settings (e.g., group discussions, conversation) or directly from several types of narratives (e.g., tweets, letters of complaint). All these materials are continuously growing with the burgeoning of new technologies for data recording, dissemination, and storage. They constitute an extremely rich source of information for studying daily life.

The current perspective of mixed methods allows us to propose an appropriate way to mix or integrate qualitative and quantitative elements. In the mixed method framework, we transform non-systematic qualitative data into a format suitable for quantitative analysis.

Extracting information of human behavior from texts and transforming it into suitably systematized and organized categorical data, without loss of key information, is a major challenge. This ‘quantitizing’ proposal consists of six stages: 1) specification of study dimensions, 2) establishment of segmentation criteria to divide the text into meaningful units, 3) building of a purpose-designed observation instrument, 4) coding of information, 5) data quality control, and 6) quantitative analysis of data.

The novel nature of our proposal is that we do not study frequency counts which, despite their serious limitations, were the only measure of quantification used in observation studies for decades. We prioritize three analytical techniques that are particularly well-suited to process qualitative data in indirect observation studies: Lag sequential analysis, polar coordinate analysis, and T-pattern detection. All three techniques are based on statistical calculations and therefore provide the necessary guarantees of replicability and robustness.
Background: Observational studies are in systematic observation, understood as a register and the quantification of behavior in its natural context. These studies applied to the specific area of sport present advantages comparing them with studies based on other designs, such as the flexibility to adapt to different contexts and the possibility of using non-standardized instruments, as well as a high degree of development in specific software and data analysis. Despite the importance and usefulness of these studies in the sports field, there is no checklist to inform these studies. Consequently, the authors do not have a guide to follow to include all the essential elements in an observational study in the sports area, and the reviewers do not have a reference tool when evaluating this type of work. To solve these problems, this work aims to develop a checklist to measure the quality of observational studies related to sports based on a content validity study. Methods: The participants were 22 judges with at least 3 years of experience in observational studies, sports areas and methodology. They evaluated a list of 60 items systematically selected and classified in 12 dimensions. They were asked to rate four aspects of each item on 5-point Likert scales to measure the following dimensions: representativeness, relevance, usefulness, and feasibility. The judges also had an open format section for comments. The Osterlind index was calculated for each item and each of the four aspects. The items were considered appropriate when obtaining a score of at least .5 in the four aspects evaluated. Results: After considering this inclusion criteria and all open format comments, the resulting checklist consisted of 54 items grouped in the same 12 initial dimensions. Finally, we highlight the strengths and limitations of this work and indicate areas for further development.

Chacón-Moscoso, Salvador 1
Sanduvete-Chaves, Susana 2
ANGUERA, MARÍA TERESA 3
Losada, José Luis 3
Portell, Mariona 4
Lozano-Lozano, José Antonio 5

B6 Checklist for Reporting Observational Studies in Sports Areas: Content Validity

1 Universidad de Sevilla/ Universidad Autónoma de Chile, Chile
2 Universidad de Sevilla, Spain
3 Universidad de Barcelona, Spain
4 Universidad Autónoma de Barcelona, Spain
5 Universidad Autónoma de Chile, Chile

tanguera@ub.edu
Response styles (RS) received increasing interest, as RS exerted negative impact on assessment of true scores by inflating or deflating observed values (e.g., Moors, 2012; Reynolds Smith, 2010). Among various approaches to detect RS, count procedure was commonly used because it is easy to use, and it does not require additional items (Bachman O’Malley, 1984). It was suggested count procedure works better when items are heterogeneous, that is, items are not significantly correlated to each other (e.g., Baumgartner Steenkamp, 2001). However, no prior studies were conducted to examine the effects of correlation between items on performance of count procedure systematically. Therefore, the present study explored how correlation between items affects count procedure to detect RS through simulation. Data will be simulated from Multidimensional Nominal Response Model (MNRM), which was widely used for survey data in previous research. Impact of four factors will be examined: 1. sample size (300 vs. 800); 2. number of items per content factor (10 vs. 20); 3. RS (NRS – no RS, acquiescence RS - the tendency to agree with items of a rating scale, extreme RS - the tendency to use the highest and lowest response categories, and mid-point RS - the tendency to use the middle response category); 4. correlations between items (items from content factor only vs. 15 items from 5 uncorrelated factors). Results of simulations will be focused on: 1. RS scores calculated by counting; 2. Pearson and Spearman correlation between true RS and estimated RS scores. Results of simulation studies will be discussed and recommendations for application of count procedure will be provided.
Without a doubt, the least of the problems caused by data loss is losing power of the test. A wise researcher will know how to prevent this situation by planning the research by calculating a priori the sample size in anticipation of data loss. If you have the possibility to select the number of subjects according to this forecast, there is no problem provided that the loss is MCAR or MAR. However, if the loss of data is non-random (MNAR), it causes a selection bias that will not be solved by increasing the sample size. It is true that if the researcher is aware of this problem, and has sufficient methodological preparation, he will plan the research based on the calculation of the sample size, and also taking care of and trying to record the data and covariates related to the loss to convert into MAR what would inevitably have been a non-random loss (MNAR). By carefully analyzing the data and verifying the convergence of results by performing sensitivity analyzes, the researcher will have managed to resolve with dignity the problem caused by the loss of data. However, the researcher is not always able to know what is causing the loss of data and therefore can not record variables related to the loss. On these occasions, the researcher must be prepared to optimize the design. We always propose to implement the SOS code but in these situations, with greater interest. The SOS code implies putting into practice three tasks: implementing solutions to avoid losing data and at the same time being prepared to test derived hypotheses, optimize observation opportunities, and implement a monitoring protocol that allows evaluating different hypotheses conditioned to the process.
Adolescence is a period of development that is marked by profound psychological changes that have a considerable impact on attachment relationships (Allen, 2008). Bowlby (1969), the founder of attachment theory, described attachment as a biologically-based adaptive system that drives the child to seek proximity to the primary caregiver during times of stress. From the cognitive point of view, attachment is associated with the construction of a set of representations related to the attachment figure, to oneself, and to the relationship between the two (López, 2006). To have self-administered, short and easy to interpret instruments that can measure the quality of the attachment relationship in adolescence is very important. Thus, the present study aimed to contribute to achieve this goal by examining the gender-based measurement invariance of the Basque adaptation of the short version of the Camir scale (B-Camir-R. Aliri et al., in press). The sample comprised 402 adolescents (53
In the 2016-2017 academic year, a cohort quasi-experimental research was initiated with the students in the first-year of the speech-therapy, and the first part of research will end in the 2019-2020 academic year when the first students finish their university studies. It aims to achieve three objectives. First. The objective in each academic year is to get the Logopedia students in their first university course to enjoy learning the different topics that the subject of Fundamentals of Methodology includes in the teaching guide, and to awaken in them enthusiasm for scientific research. This knowledge is reinforced in the subject of Introduction to Psycholinguistics. Second. In the subsequent courses, and until the end of the degree studies, it is intended that students continue to remember fundamental concepts of methodology putting them into practice in different subjects. With this it is intended that students are able to recognize that the methodological knowledge they acquire in the subject of Fundamentals of Methodology are useful, necessary and fundamental to approve the subject, of course, but above all have a major importance for growth and advancement of the science that supports all the professional activities that a speech therapist performs, and therefore, are fundamental for their integral formation as speech therapist. Third. To evaluate the experienced change of the opinions expressed about the importance and usefulness of this subject in the General Survey of the Teaching and to conclude that the effort is worthwhile. We present the quot;dynamiquequot; design of the research procedure to achieve the objectives. The first results are very encouraging. The students have substantially improved their academic grades and their opinion with respect to the three previous academic years in the General Teaching Survey is more positive.

FERNÁNDEZ GARCÍA, MARÍA PAULA 1
Suárez Coalla, María Paz 1

B10 How to Enhance Students’ Interest in Fundamentals of Methodology? A Cohort Quasi-Experimental Design

1 Universidad de Oviedo, Spain

paula@uniovi.es
In Monte-Carlo Markov-Chain (MCMC) estimated analyses, users need to choose a number of stochastic chains. The first step in determining model convergence is to investigate chain mixing, which is commonly assessed using R-hat statistics. Convergence occurs when the Markov Chains reach a similar within and between chain variance (e.g., Gelman, Shirley, 2011). The number of chains used, however, may have an effect on estimates of convergence. Literature on MCMC estimation provides few substantiated recommendations other than to use more than a single chain in order to avoid local optima that depend on starting values (Gelman, Carlin, Stern, Dunson, Vehtari, Rubin, 2014). Besides the recommendation to choose a chain number greater than one, no further evidence is supplied to suggest a specific range. Although, divergent decisions have been noted when using different numbers of chains, the impact of chain number specification on convergence statistics has not been systematically investigated.

To explore the effects of chain number, prior decision, and chain length on convergence, a MCMC simulation was conducted. As a prototype model, we generated data for a random intercept random slope (multilevel) model. The model was then analyzed under differing prior conditions, which included diffuse, strong and correct, or strong and incorrect priors using the Stan software. Both chain length (2,000 vs. 4,000), and number of chains specified (2 through 19) were varied. Convergence (R-hat) and bias of parameter estimates were investigated. Results showed that R-hat estimates were more likely to fall below 1.1 when higher numbers of chains were used (gt;8). For diffuse and incorrect prior specifications the distribution of R-hat values for large chain numbers had a mean closer to 1.0 and a lower variance (more stable), as compared to lower chain numbers (lt; 5). Results and consequences for the application of MCMC analysis are discussed.
Violence against children is a public health problem. In 2016, one billion children were victims of violence. A problem of such magnitude must have measurement instruments that present adequate psychometric properties. The aim of this research is to study the psychometric properties of child victimization instruments. A systematic review was conducted. The electronic databases consulted were: Proquest, PsycINFO, PsycTest, PsycArticles, SCOPUS, Pubmed and Web of Knowledge. Direct searches have been conducted in: (a) journals specialized in child victimization; (b) abstracts books of Congress specialized in child victimization; (c) experts on child victimization; and (d) the references of the studies included in the systematic review have been reviewed. The inclusion criteria were: (a) include one instrument of child victimization; (b) published in English or Spanish; (c) the psychometric properties of the child victimization instrument should be indicated; (d) victimization information should refer to children (under 18 years). The variables included in this research were: year of publication, sample size, age of children, nationality of the study, number of items, original instrument or adaptation, psychometric properties (reliability and validity). The results indicate that there is a wide variety of instruments for child victimization. The study of the psychometric properties needs greater involvement by the researchers to improve the methodological quality of the studies of child victimization instruments.

Zaldivar-Basurto, Flor
Ortega-Campos, Elena
De la Fuente-Sánchez, Leticia
García-García, Juan

B12 Psychometric Properties of Child Victimization Instruments: A Systematic Review

1 University of Almería, Spain

elenaoortega@ual.es
In many studies such as educational testing and drug testing, there is no knowledge regarding whether the treatment slope $\beta_1$ is superior to or inferior to the control having slope $\beta_2$. A frequent task is to determine whether an empirical relationship estimated within two independent samples is equivalent. Thus, the test of equivalence of two regression slopes is appropriate. By specifying in advance two equivalence boundaries, the largest difference that is clinically acceptable, the difference in slopes between two groups lies within a specified small interval can be tested. When the null composite hypothesis is simultaneously rejected (so-called intersection-union test), the two slopes are considered practically equal or equivalent. The conventional formulas for calculating the required sample sizes adopt z tests. In the present study, we considered the case of variances being unknown and possibly unequal and proposed new formulas by adopting Welch’s (1938) two-sample t statistic. The present study also considered different unit cost of sampling and developed optimal allocation ratio for two groups. That is, for a given statistical power, we choose a sample size allocation ratio to have a minimal sampling total cost. Or given a fixed sampling total cost, the sample size allocation ratio can be chosen to have a maximal statistical power. We provide an example to illustrate the needed sample sizes for these two purposes. The simulation results show that the proposed method can achieve the nominal statistical power or/and has minimal sampling cost. The implication of the proposed method was discussed in terms of statistical power and the sampling cost for planning sample sizes in experiments. An R program is also provided for applied researchers to calculate appropriate sample sizes as needed.
The objective of this work is to evaluate the statistical power of five covariance structure selectors to detect the missing data mechanisms in repeated measures designs with complete data vectors, monotone loss patterns and covariance structure underlying the most frequent data in psychological research. For this purpose, the research was made with Monte Carlo simulation method with SAS 9.4 statistical software, through 1000 simulations. The manipulated variables were: a) complexity of the model used to generate data; b) total sample size (50-100-120); c) pairing of the covariance matrices and size (null, positive, negative); d) number of measurements occasions; e) equality and inequality of dispersion matrices; f) normal and non-normal distribution of data. Results show that the highest power levels are for the selection criteria AIC, AICCR and HQICR (recommended for scenarios 2 and 3 under UN matrix and n=50), while BICR, CAICR show lower average power levels (recommended for ARH matrix with n=50).
Single-case experiments (SCEs) are increasingly being used to assess the efficacy of interventions for individual participants in various fields of behavioral science (Horner et al., 2005). As a result it is becoming more important to develop widely accepted meta-analytical techniques for SCEs in order to include this type of experimental evidence in verifying evidence-based practices in these substantive domains of science. Over the years, multiple proposals for the meta-analysis of SCEs have been made (e.g., Burns, 2012; Hedges, Pustejovsky, Shadish, 2012, 2013; Heyvaert et al., 2016; Parker, Vannest, Davis, 2011; Rindskopf, 2014; Van den Noortgate Onghena, 2003). In this poster we present a nonparametric technique for meta-analyzing randomized SCEs by inverting randomization tests to construct nonparametric confidence intervals for combined effect sizes (CICES). CICES is based on the combination of two separate operations. The first operation entails combining the randomization schemes of multiple randomized SCEs in order to create meta-randomization schemes. These meta-randomization schemes can be used to perform a randomization test at the meta-level using a multivariate test statistic combining the data of multiple SCEs into a combined effect size estimate. The second operation entails the inversion of this meta-randomization test to a two-sided nonparametric confidence interval for the combined effect size estimate using the principle of hypothesis test inversion (Garthwaite, 2005). We demonstrate how CICES is appropriate for meta-analyzing any combination of randomized SCEs using various types of phase designs and alternation designs. In addition, we illustrate how CICES can be used with various standardized or unstandardized effect size measures including Hedges et al.'s (2012) $d$ and Hedges’ (1981) $g$ statistic. We have implemented the CICES technique for single-case meta-analysis in a freely available R-function.
Research on the interpersonal perception of personality examines how accurately others (e.g., peers) judge a target person. Typically, observers provide their judgments by completing personality questionnaires using items formulated in the third-person form (e.g., "He/She is outgoing."). To quantify the judgment accuracy, two indices are frequently used: (1) Self-other agreement (SOA; i.e., correlation between targets’ self-ratings and observer-ratings) and (2) Consensus/Other-other Agreement (i.e., convergence between judgments of two or more observers). It has been found that the reliability and validity of self-rating scales increase with response options (Lozano et al., 2008), but it is unclear whether the number of response options influence indices of accuracy and its variance components (e.g., stereotype bias). To address this gap in the literature, we will analyze the differential effect of response categories on the two accuracy indices. Therefore, data collection of acquainted peer-dyads is underway (currently N = 110 dyads; determined = 150). Participants provide ratings on measures of broad (Big Five; life satisfaction) and narrow (adult playfulness, three dispositions towards ridicule and being laughed at) personality traits in four within-subject conditions (2/4/6/8 response categories). We will examine whether coefficients of SOA and consensus differ depending on response conditions and whether profile accuracy (i.e., overall, normative, and distinctive) is affected. Findings will provide insight regarding the selection of response options for the study of accuracy in interpersonal perception.
The movement led by Wilkinson’s The Task Force on Statistical Inference, promotes the use of the effect size index accompanying the statistical significance test in research articles. Additionally, this movement emphasizes the importance of effect size use with the publication “Statistical Methods in Psychology Journals: Guidelines and Explanations”, which indicates a series of recommendations to researchers with the aim of improving the methodological quality of research papers in Psychology. One of the recommendations was the use of effect size in the results section. In 2001, the fifth edition of the Publication Manual of the American Psychological Association included this recommendation. In Spain, Anales de Psicología, International Journal of Clinical and Health Psychology and Psicothema, includes in the JCR ranking, expressed the need to use effect size in research papers. The aim of this research is to determine whether the papers published in the Spanish Psychology Journals indexed in JCR since 2008 include an effect size index. The results indicate that the inclusion level of effect size is still far from desirable. A longitudinal study allows to check whether the effect size recommendations has been followed throughout the study period.
The use of adapted questionnaires is essential for the development of the Psychology as science in its different fields. Study of attitudes of the police officers towards Intimate Partner Violence (IPV) is an example of this. To evaluate these attitudes there are several scales, being the proposed by Chu Sun (2014) the most popular one. This scale is not adapted to the Spanish culture and there is a lack in the study of validity evidences on a Spanish sample. The aim of this research is to study the validity evidences of this scale, focusing on the evidence based on internal structure. A total of 296 Spanish police officers composed the sample. Using Mplus five nested factorial structures were analyzed (unidimensional, bi-dimensional, bi-dimensional-bi-factor, three-dimensional and three-dimensional-bi-factor) checking the model fit. The study of the bi-factor models is justified by the high correlations obtained between the dimensions. Several statistical indices derived from the bi-factor solutions were computed (hierarchical omega, relative omega, factor determinacy, percentage of uncontaminated correlations...). All these indexes will allow to interpret the internal structure, and its evidences based on internal structure, in a better fashion. The model that has a better fit is the three dimensions bi-factor one. In this model, the former indices show that although the three dimensions bi-factor model obtains the better fit, the model can be considered essentially unidimensional. The results show that the internal structure of the questionnaire that has a better fit is a multidimensional one. Although these results are important in order to study the construct validity, data also support that psychologist can consider the questionnaire as a unidimensional measurement of the construct.
Of critical importance to education policy is the monitoring of trends in education over time. Developing optimal predictive models allows researchers and policy makers to assess cross-country progress and forecasts toward that goal. The purpose of this project is to apply Bayesian model averaging to cross-country growth regressions in education achievement using data from TIMSS (Trends in International Mathematics and Science Study). Whereas it is common practice to select one particular model from a set of models based on the data fit, model averaging utilizes relevant information from a set of models. This technique allows considering the model uncertainty. Bayesian model averaging has been applied to a wide variety of content domains in economics, bioinformatics, weather forecasting, causal inference within propensity score analysis, and structural equation modeling. However, applications to education data were not common, because trend data with sufficient number of data points was not available. This study utilizes data from TIMSS, an international assessment of mathematics and science at the fourth and eighth grades. 20 years of data collection allows for analyzing changes over time on country level to predict future developments of student achievement as well as gender gaps. We expect the Bayesian model averaging for growth curve models to produce more precise forecasts of student achievement and gender gaps than conventional approaches to forecasting. The study supports basic theoretical research on the problem of prediction in international large-scale assessment data and contributes to it by adding results from one of the largest international large-scale assessments in education. The unique contribution pertains in particular to the length of the survey and a fine grain of curriculum data crucial for policy analysis. It presents a novel approach to utilize large-scale assessment data for forecasting and policy analysis.
Wednesday, July 25
Replication is a fundamental aspect of the scientific method and is central to the rhetoric of science. Yet recent empirical research has called into question the replicability of experimental research in fields as diverse as economics, medicine, and psychology. This work undermines the credibility of science and the evidence science provides. Surprisingly, there has been little research on the methodology of replication itself, including the design of replication studies and appropriate statistical analyses to determine whether a set of studies replicate one another. Perhaps as a result, some recent programs of research on replication have used multiple, and sometimes mutually contradictory, methods to study replication. This talk will draw on a meta-analytic perspective to formalize ideas about the definition of replication and the analysis of replication studies. I will focus on three problems that seem straightforward, but will argue that each of them is more complex than it first appears. One is the precise definition of replication: What exactly does it mean to say that the results of a set of studies replicate one another? The second is the statistical analysis of replications: Given a definition of replication, what statistical analysis is appropriate? The third is the design of replication studies: What kind of ensemble of two or more studies should we assemble to evaluate whether results replicate?
SESSION Causal Inference in Dynamic Models

Chair: Christian Gische

Wednesday, July 25, 11:30 - 12:30
Saal Friedrich Schiller

Issues in Causality in Discrete-Time and Continuous-Time Stochastic Process Models
JULIA GANTNER, Rolf Steyer

An Interventionist Approach to Causal Inference Based on Panel Data
CHRISTIAN GISCHE, Manuel Voelkle

A General Nonlinear Model for the Identification of Mediators Without the No Confounder Assumption
HOLGER BRANDT
Issues in Causality in Discrete-Time and Continuous-Time Stochastic Process Models

Friedrich-Schiller-Universität Jena, Germany

julia.elisa.gantner@uni-jena.de

Keywords:
Causality, Stochastic Process Models, Structural Equation Modeling

It is a classical experimental paradigm to manipulate manifest treatment conditions ($X$) assumed to have an effect on a latent theoretical variable ($\eta_1$) which in turn affects a manifest or latent outcome variable ($\eta_2$). With this paradigm researchers intend to test the theory that $\eta_1$ has a causal effect on $\eta_2$. Therefore they estimate the direct effect of $X$ on $\eta_1$ (manipulation check) and make the erroneous conclusion that a successful manipulation check means a significant total effect of $X$ on $\eta_2$ can be interpreted as evidence for their originally stated hypothesis that $\eta_1$ affects $\eta_2$. Even if the theoretical constructs considered in a case like this can be represented in a model with only the three random variables $X$, $\eta_1$ and $\eta_2$, a causal interpretation of the putative effects is only possible if the direct effect of $X$ on $\eta_2$ is zero and all relevant pre-treatment variables are taken into account. Controlling for all pre-treatment variables when estimating direct and indirect effects is even necessary in a randomized experiment as Mayer et al. (2016) have shown.

Looking at substantive psychological theories in more detail, however, it seems that a theoretical construct can often not be defined as a single random variable (such as $\eta_1$ or $\eta_2$), but rather as a stochastic process with latent variables $\eta_{it}$, $t \in T$. These stochastic processes might not even be time-discrete, as assumed in cross-lagged panel models and multivariate autoregressive processes. Instead, for many psychological constructs they have to be conceptualised as continuous-time processes. This raises new questions concerning causality in continuous-time processes, some of which are addressed in this talk.
During the last two decades a comprehensive theory of causal inference based on directed acyclic graphs (DAGs) has been developed (Pearl, 2000). It is well known that non-recursive structural equation models can be represented as DAGs and thus can be analyzed using Pearl’s framework of causal inference. Despite the crucial role of time for the study of causal effects, however, surprisingly little attention has been put on integrating longitudinal models of change into Pearl's approach to causal inference. In this presentation, we apply Pearl's general causal effect formula to the bivariate autoregressive cross-lagged panel model that incorporates variation across individuals (i=1,…,N) and over time (t=1,…,T) and allows for unobserved heterogeneity across individuals.

We apply existing sufficient criteria for causal identification (e.g. the Back-Door criterion) to the above model and show that certain causal effects of interest can be identified through adjustment with respect to unobserved heterogeneity. The suggested non-parametric procedures for estimating the identified causal effects rely on the assumption that all adjustment variables are observable. These procedures are therefore not feasible in the presence of unobserved heterogeneity.

In a next step we turn to existing parametric estimation techniques based on panel data (e.g. fixed-effects, random-effects or likelihood based estimators) and analyze their feasibility for the estimation of causal effects. We explicitly state assumptions under which these parametric techniques yield consistent estimates of the parameters of the interventional distribution. Based on the interventional distribution causal effect quantities can be computed and interval forecasts of effects of interventions can be calculated. We conclude by comparing the expected consequences of actively manipulating a putative cause variable with the consequences of passively observing changes in a putative cause variable.
The reliable identification of mediator variables that transmit an effect of an intervention to an outcome has been an ongoing problem in social and behavioral sciences. Most methods developed so far rely on the assumption of no unobserved confounders, meaning that the model under investigation included all relevant covariates. This assumption has been criticized repeatedly and it has been shown that mediators identified under this assumption often provide spurious results because even minimal misspecification may already lead to an identification of artificial mediator variables. One class of alternative models are structural mean models (SMM; Ten Have et al., 2007; Zhang & Zou, 2015) that do not rely on the no confounder assumption. These models provide unbiased results even if relevant covariates are not included in the model. The models have been extended to include interactions between intervention, mediators and covariates. However, these models suffer from two major problems. First, these models are very inefficient compared to traditional mediator models and provide a very lower power in typical applied settings. Second, the estimation of interaction effects is often impossible due to multicollinearity problems in the estimation routine. In this talk, I will present a nonlinear semiparametric extension of the SMM that substantively increases its efficiency and overcomes the multicollinearity problem. Simulation studies are presented that show that the extended model has two to three times more accurate (and efficient) estimates than the original SMM for direct and indirect effects as well as for interaction effects included in the model. Further, the results reveal that the model is robust against violations of its model assumption. In an empirical data set, its application is illustrated. Further extensions and its applicability are discussed.
SESSION Item Response Theory

Chair: Frans Kamphuis

Wednesday, July 25, 11:30 - 12:50
Salon Schlegel

A State-Space Approach for Student Growth Percentile Estimation
FRANS KAMPHUIS, Ron Engelen

Psychometric Evaluation of the d2 Test of Sustained Attention With the Rasch Poisson Counts Model
PURYA BAGHAEI, Mahsa Nadri

Forecasting Clinical Outcomes by Combining Measurement and Prediction Models for Health Evaluations
NIELS SMITS

Exchanging Selection Rules from Cognitive Diagnosis Modeling to Traditional Item Response Theory
MIGUEL A. SORREL, Juan Ramón Barrada, Jimmy de la Torre
Betebenner (2009) introduced student growth percentiles (SGP) for norm- and criterion-referenced student growth. Each student’s current test score is expressed as a percentile rank in the distribution of current test scores among students who had the same past test scores. In this approach a vertical scale is not necessary. Recently Lockwood and Castellano (2015) suggested two alternative approaches for estimating SGP. One based on modelling the conditional cumulative distribution functions the other based on multidimensional item response theory. The last approach can overcome problems with measurement error in both the past and current test score. Kamphuis and Moelands (2000) formulated a states-space approach: the measurement problem is separated from the structural model. In this framework we can handle a lot of measurement models like IRT, CTT etc. The structural part can include many domains, occasions and background information. In the talk the methodological aspects and the validity issues related to monitoring growth are discussed. Emphasized is the need for a framework of reference for the interpretation of growth. A student monitoring system for secondary schools in Kazakhstan will be presented as an illustrative Example.
The d2 Test is a cancellation test of attention and concentration in which respondents have to cross out target variables among similar nontarget stimuli (Brickenkamp & Zillmer, 1998). The target variable is a "d" with two dots above or below it. The targets are randomly interspersed among nontarget characters. The nontarget characters are d’s with one, three or four dots above or below them and p’s with one, two, three, or four dots above or below them. The target and nontarget characters are presented in 14 consecutive lines. Separate time limits of 20 seconds are allotted for each line. The construct validity of the test has been investigated with classical methods of factor analysis and criterion measures. To the best of our knowledge no study so far has examined the fit of the d2 Test to IRT models. In this study the fit of the test to the Rasch Poisson Counts Model (RPCM) is examined. The structure of the test (a combination of 14 lines of stimuli each with a separate time limit) makes it an ideal candidate for RPCM scaling. In this study the overall fit of the d2 Test to RPCM, the fit of the individual items (lines), and the reliability of the test are investigated.

BAGHAEI, PURYA 1
Nadri, Mahsa 1

Psychometric Evaluation of the d2 Test of Sustained Attention With the Rasch Poisson Counts Model

1 Islamic Azad University, Mashhad Branch, Iran

puryabaghaei@gmail.com

Keywords:
Rasch Poisson Counts Model, Sustained Attention, d2 Test
Both in medical research and clinical practice, questionnaire-based assessments are used to obtain information about the physical, mental, and social well-being as experienced by patients. Commonly, measurement models, such as those based on Item Response Theory (IRT), are employed to meaningfully reduce a patient's item scores to a single test score (typically denoted $\theta$) or a set of test scores. Such scores are mostly used to evaluate and monitor the patient, and to provide feedback on his or her status. For such purposes, measurement precision is key because the scores should accurately represent the patient's attributes. However, tests are also used for predictive purposes, such as forecasting a future health state, or a diagnosis based on the gold standard. It has been shown that in test design a trade-off exists between measurement and prediction, and thus that accurate measurement is no prerequisite for good prediction (e.g., Smits et al., 2017). This raises several questions concerning the use of test data for prediction purposes.

In the present study, three questions are answered: (1) Is a single test score sufficient when the purpose of the test is prediction, or should the individual item scores be used? (2) How should multiple test scores be combined for optimal prediction? (3) How should measurement errors of test scores under an IRT model be incorporated into a prediction model to obtain sound estimates of predictive power?

To answer these questions an illustrative data file is used consisting of 735 patients with scores on 42 PROMIS Pain Quality items, and 26 PROMIS Affective Pain items. The patients either did or did not have a clinical condition associated with neuropathic pain; this outcome is considered the gold standard and used as target variable for prediction.
Nowadays, there are two predominant approaches in adaptive testing. One is based on cognitive diagnosis models and is referred to as cognitive diagnosis computerized adaptive testing (CD-CAT), and the other one is the traditional CAT based on item response theory. The present study evaluates the performance of two item selection rules (ISRs) originally developed in the CD-CAT framework, the double Kullback-Leibler (DKL) and the global discrimination index (GDI), in the context of traditional CAT. The accuracy and test security associated with these two ISRs are compared to those of the point Fisher information and likelihood-weighted KL using a simulation study. Five dependent variables were evaluated: bias and root mean square error relative to the measurement accuracy; overlap rate as indicator of test security; mean values of the a and c parameters administered, with the aim of analyzing the kind of item that tends to be selected by each ISR; and the correlation between the item exposure rates for each pair of ISRs, as indicative of the convergence between them. The impact of the trait level estimation method was also investigated. Maximum likelihood (ML) and expected a posteriori (EAP) estimation methods were compared. The results of the study show that the new ISRs can be used to improve the accuracy of CAT with fewer items administered, particularly in the case of DKL. This is of major importance in contexts (e.g., educational, medical) where testing time is always an issue. In addition, both rules selected a different set of items: with DKL, items with the highest a parameter were administered, whereas items with the lowest c parameter were administered with GDI. Regarding the trait level estimation methods, we found that EAP was generally better in the first stages of the CAT, and converged with ML when a medium to a large number of items was administered. Several implications and possible future directions are provided.

SORREL, MIGUEL A. 1
Barrada, Juan Ramón 2
de la Torre, Jimmy 3

Exchanging Selection Rules from Cognitive Diagnosis Modeling to Traditional Item Response Theory

1 Universidad Autónoma de Madrid, Spain
2 Universidad de Zaragoza, Spain
3 The University of Hong Kong, Hong Kong

miguel.sorrel@uam.es

Keywords: Cognitive Diagnosis Modeling, Computerized Adaptive Testing, Gender Development Index, Item Response Theory, Item Selection Rules, Kullback–Leibler Divergence
SESSION Applied Statistics

Chair: Patricia Martinkova

Wednesday, July 25, 11:30 - 12:10
Salon Hölderlin

Model-Based Reliability to Check for Disparities in Ratings of Internal and External Applicants
PATRICIA MARTINKOVA, Dan Goldhaber, Elena Erosheva

Impact of Formal Educational Upgrading on the Likelihood Leaving Unemployment
Sushant Pandkar, JOHANNES JAENICKE
In this work we address disparities in ratings of internal and external applicants. We develop model-based inter-rater reliability (IRR) estimate to account for various sources of measurement error, their hierarchical structure and the presence of covariates, such as assessed status, that have the potential to moderate IRR. Using dataset of ratings of applicants to teaching positions in Spokane district in Washington, USA, we first test for bias in ratings of applicants external to the district, which is shown to be significant even after including various measures of teacher quality in the model. Moreover, with model-based IRR, we show that consistency between raters is significantly lower when rating external applicants. We further address how IRR affects the predictive power of measurement in different scenarios and conclude the work by discussing policy implications and applications of our model-based IRR estimate for teacher hiring practices.
Using the adult cohort from the National Educational Panel Study (NEPS), we analyze the probability leaving short- or long-term unemployment taking into account structural educational background, complete professional history and demographic characteristics. Our discrete choice model is based on human capital theory including life-long learning and professional experiences. The results show that professional history, the age of an individual at the time of unemployment, and educational background are significant predictors for getting out of unemployment. The job finding probabilities are driven by general and model specific explanatory factors in case of short- or long-term unemployment.

Pandkar, Sushant 1
JAENICKE, JOHANNES 2

Impact of Formal Educational Upgrading on the Likelihood Leaving Unemployment

1 Alumnus of the Willy Brandt School of Public Policy, University of Erfurt, Germany
2 Independent Scientist, Germany

Jaenicke@Johannes-Jaenicke.de

Keywords:
Discrete Choice, Hartz IV, Marginal Effects, Panel, Structured Education, Unemployment
SESSION Replication Crisis

Chair: Jörg Blasius

Wednesday, July 25, 11:30 - 12:30
Salon Novalis

Fabrication of Interview Data in PISA and PIACC
JÖRG BLASIUS

Transparency and Replicability in Cross-National Survey Research
ELENA DAMIAN, Bart Meuleman, Wim Van Oorschot

Questionable Research Practices in Student Theses - Prevalence, Antecedents and Implications
ANAND KRISHNA, Sebastian Maximilian Peter
The quality of survey data is a function of the three levels of actors involved in survey research projects: the respondents, the interviewers, and the employees of the survey research organizations. I argue that task simplification dynamics can occur at each of these levels resulting in reduced data quality. The precise form of task simplification differs at each of the three levels. For respondents, it might take the form of utilizing only specific parts of the available response options; interviewers might ask the demographic questions and some basic information only and then fabricate plausible responses for the remainder; the third type applied by employees of research institutes is the near-duplication of entire questionnaires. I will use data from the Programme for International Student Assessment (PISA) 2012 and the Programme for the International Assessment of Adult Competencies (PIAAC) to document various task simplification techniques performed at each of these levels and I propose a new statistical method to discover interviewer falsifications through the field work period.
Transparency and replicability in cross-national survey research: Demarcation of problems and possible solutions

Abstract This paper offers insights into the level of transparency and replicability of cross-national survey research. The first contribution is theoretical as we provide an overview of the current measures taken to achieve research transparency in cross-national survey studies. We do so by developing a heuristic theoretical model regarding the actors, factors and processes that influence the level of transparency of an academic article. The second contribution is empirical and regards our dependent variable - article transparency. Specifically, using a random sample of 305 comparative studies published in one of 29 peer-reviewed Social Sciences journals (1986-2016), we show that most articles do not provide crucial empirical information for independent researchers to evaluate the validity and reliability of study's findings or perform a direct replication. Additionally, we develop and propose a set of transparency guidelines tailored for reporting cross-national survey research.

DAMIAN, ELENA
Meuleman, Bart
Van Oorschot, Wim

Transparency and Replicability in Cross-National Survey Research

1 KU Leuven - University of Leuven, Belgium

elena.damian@kuleuven.be

Keywords: Crossnational Survey Research, Replication, Secondary Analysis, Transparency
Questionable Research Practices in Student Theses - Prevalence, Antecedents and Implications

1 University of Würzburg, Germany

krishna@psychologie.uni-wuerzburg.de

Keywords: Questionable Research Practices

Although questionable research practices (QRPs, e.g. Fiedler & Schwarz, 2016) have received more attention in recent years, little research has focused on their prevalence and acceptance in student samples. Students are the researchers of the future and will also represent the field outside of academia. Therefore, it is vital to establish that they are not learning to use and accept QRPs, which would negatively impact their ability to produce and evaluate meaningful research. 207 psychology students and fresh graduates provided data on the prevalence and predictors of QRPs. We focused on attitudes towards QRPs, beliefs about whether significant results constitute better science or lead to better grades and motivation and stress levels as predictors. Furthermore, we assessed perceived supervisor attitudes towards QRPs as an index of perceived normativeness and an important influencing factor on student’s QRP attitudes and use. The results were generally in line with Fiedler & Schwarz’ (2016) estimates of QRP prevalence. The best predictor of QRP use was students’ QRP attitudes, although perceived supervisor attitudes exerted both a direct and indirect effect via student attitudes. Motivation to write a good thesis was a protective factor, whereas stress had no effect. Students did not subscribe to the belief that significant results were better for science or their grades, which may explain why these beliefs did not impact their QRP attitudes or use. Finally, students engaged more in QRPs pertaining to reporting and analysis than those pertaining to study design. These results imply that supervisors have an important function in shaping students’ attitudes towards QRPs and the opportunity to further improve their research practices by motivating them well. Furthermore, this research provides some impetus towards identifying predictors of QRP use in academia.
Incidental data are data that people produce incidentally, as a byproduct of the normal course of operations of a platform, business, or government. Well-known examples include using Twitter, Facebook, Google search, smartphones, badges, etc. to study social phenomena, such as election behavior, attitudes, employment, or consumer confidence. It has been almost ten years since various high-impact papers and books have proclaimed the end of traditional social research and the beginning of a new era of exciting new possibilities for social science.
Tests represent the evaluation technology most widely used by psychologists in their professional practice. In recent years there have been great advances in psychological evaluation, which have affected both the tests themselves and their use. In this presentation, recent advances in the construction and use of tests are reviewed, and some future challenges discussed. This review is structured around six dimensions of change: the evolution of psychometric models, changes in the technology used, developments in the construction of items, estimation of reliability, conceptualization of validity, and use of tests in professional practice. Finally, some future perspectives are discussed, taking into account the great impact of new information technologies on the evaluation methods, tests included.
SESSION Latent Variable Analysis

Chair: Tobias Koch

Wednesday, July 25, 16:30 - 17:50
Saal Friedrich Schiller

Analyzing Different Types of Moderated Method Effects in Confirmatory Factor Models for Structurally
TOBIAS KOCH, Augustin Kelava, Michael Eid

Repeated Measures ANOVA with Latent Variables Using the Latent Growth Component Approach
BENEDIKT LANGENBERG, Axel Mayer

DIF of Self-Assessment Items Across Different Levels of a Latent Variable: Positive Affect
ANA HERNÁNDEZ, Vicente González-Romá, Inés Tomás

Parameter Associations in Bivariate Dual Change Score Models: Implications for Simulation Studies
HOLLY O’ROURKE, Kevin Grimm
Koch, Tobias ¹
Kelava, Augustin ²
Eid, Michael ³

Analyzing Different Types of Moderated Method Effects in Confirmatory Factor Models for Structurally

¹ Leuphana University Lüneburg, Germany
² Eberhard Karls University of Tübingen, Germany
³ Freie Universität Berlin, Germany

tobias.koch@leuphana.de

Keywords:
Latent Variable Analysis, Structural Equation Modeling

In this talk, we present two confirmatory factor models for multimethod designs with structurally different methods for the analysis of latent moderations: the nonlinear latent difference (NL-LD) model and the nonlinear correlated trait–correlated method minus-one (NL-CTC[M-1]) model. We explain how different moderated method effects can be examined in the NL-CTC(M-1) model and why the classical NL-LD model does not permit this fine-grained analysis of method effects. To fully recover the results of the NL-CTC(M-1) model, we propose an extended version of the NL-LD model. The different versions of the nonlinear multimethod models are compared with regard to the psychometric definition and meaning of the latent moderated method effects and are illustrated using real data from a multirater study. Finally, the advantages and challenges of incorporating latent interaction effects in multimethod confirmatory factor models are discussed.
In this presentation, we introduce a way of testing hypotheses of interest in repeated measures designs with latent variables using structural equation modeling (SEM). Traditionally, such designs are analyzed using repeated measures analysis of variance (repeated measures ANOVA). A limitation of repeated measures ANOVA is that only manifest variables can be used. To overcome this limitation, we extend the traditional method to incorporate latent variables. We build on the latent growth components approach (Mayer, Steyer & Mueller, 2012) for this purpose. The latent growth approach is a flexible method that can be used to define latent effect variables in a structural equation modeling framework. We propose to use a comprehensive structural equation model that is specified in three steps. In the first step, a measurement model is formulated for the latent dependent variables. In the second step, a contrast matrix is used to decompose the latent dependent variables into several latent effect variables (e.g., latent difference score variables) that represent main effects and interaction terms. The matrix used to decompose the latent variables corresponds to the transformation matrices that appear in multivariate general linear hypotheses. In the third step, the structural coefficients for the SEM are derived by inverting the contrast matrix. All parameters of the resulting structural equation model are estimated simultaneously. This approach allows us to directly examine main effects and interaction terms by testing the means of the latent effect variables using Wald tests.
Differential Item Functioning (DIF) is said to occur when the item performs differently for one group of a population compared to another group, after controlling for the construct being measured and the possible existing construct differences between the groups (Holland, 1993). In other words, the expected item responses after controlling for the latent trait, depend on group membership. However, apart from group membership (for example sex, culture, etc.), DIF may be attributed to a “quantitative” variable (for example age or impulsiveness). Typically, in these cases variables have been categorized in order to test for DIF (e.g. Fleishman, Spector & Altman, 2002) with the corresponding loss of information.

MIMIC models have been used to test for uniform DIF by regressing item responses on the grouping variable, after allowing for mean differences across groups (Woods, 2009). In the present study we propose that existing DIF across different levels of a latent construct can also be assessed by means of MIMIC models. Specifically, considering that, there is some evidence that positive affect increases the perceptions of self-evaluative constructs such as self-efficacy (e.g. Medrano, Flores-Kanter, Moretti & Pereno, 2016) we assess whether item responses to a Core-Self Evaluation questionnaire (CSE; Judge, Erez, Bono, & Thoresen, 2005) show DIF across different levels of positive affect, when controlling for the relation between the two latent variables (CSE and positive affect). We fitted a MIMIC model in a sample of 503 participants by means of SEM, and regressed CSE item responses on the two latent variables and the interaction between them. Results show that, after controlling for the latent CSE construct, positive affect and the interaction significantly predicted CSE item responses. As expected, positive affect positively biased the CSE assessments. Interestingly this bias was not constant along all levels of CSE.
Latent change score models are longitudinal structural equation models that can be used to concurrently investigate growth over time and dynamic relations between two variables (known as bivariate dual change score [BDCS] models). These models are useful instruments for researchers studying development, and as such their use in the social sciences has increased in recent years (Ferrer & McArdle, 2010), particularly the use of the BDCS model. Methodological researchers have used simulation studies to examine the complexities of these models using a variety of procedures for parameter selection and data generation (Grimm, 2006; Hamagami & McArdle, 2001; O’Rourke, 2016; Prindle & McArdle, 2014; Usami, 2014; Usami, Hayes, & McArdle, 2016; Voelkle & Oud, 2015). With simulation work for the BDCS model, there is currently no standardized procedure to select parameters and produce data trajectories that appropriately simulate trajectories seen in real-world data. In this study, we first review the current simulation work on BDCS models and describe the parameter selection procedures that have been used. Many studies either use parameters based on published studies with BDCS models or use arbitrary criteria for parameter selection. We then describe the mean and covariance expectations for BDCS models (Grimm & McArdle, 2005) in the context of their usefulness for parameter selection and data generation. Finally, we propose a parameter selection procedure that retains the unique associations among parameters of the BDCS model, which produces trajectories that mimic trajectories found in empirical work.
SESSION Multilevel Analysis

Chair: Johannes Hartig

Wednesday, July 25, 16:30 - 17:50
Salon Schlegel

Estimation of Random Group DIF Using Two- and Three-Level GLMMs
JOHANNES HARTIG, Carmen Köhler, Alexander Naumann

Multilevel Models for Evaluating the Effectiveness of Instruction: ANCOVA vs. Change-Score Approach
CARMEN KÖHLER, Johannes Hartig

The Optimal Design of Cluster Randomized Trials With Outcomes at Individual and Cluster Level
MIRJAM MOERBEEK

Comparative Performance of Single Trial Multilevel Analyses of Event-Related Brain Potentials
JUAN CARLOS OLIVER-RODRÍGUEZ, Mirjam Moerbeek
HARTIG, JOHANNES ¹
Köhler, Carmen ¹
Naumann, Alexander ¹

Estimation of Random Group DIF Using Two- and Three-Level GLMMs

¹ German Institute for International Educational Research (DIPF), Germany

hartig@dipf.de

Keywords:
Differential Item Functioning, Invariance, Item Response Theory, Multilevel Analysis

Of critical importance to education policy is the monitoring of trends in education over time. Developing optimal predictive models allows researchers and policy makers to assess cross-country progress and forecasts toward that goal. The purpose of this project is to apply Bayesian model averaging to cross-country growth regressions in education achievement using data from TIMSS (Trends in International Mathematics and Science Study). Whereas it is common practice to select one particular model from a set of models based on the data fit, model averaging utilizes relevant information from a set of models. This technique allows considering the model uncertainty. Bayesian model averaging has been applied to a wide variety of content domains in economics, bioinformatics, weather forecasting, causal inference within propensity score analysis, and structural equation modeling. However, applications to education data were not common, because trend data with sufficient number of data points was not available. This study utilizes data from TIMSS, an international assessment of mathematics and science at the fourth and eighth grades. 20 years of data collection allows for analyzing changes over time on country level to predict future developments of student achievement as well as gender gaps. We expect the Bayesian model averaging for growth curve models to produce more precise forecasts of student achievement and gender gaps than conventional approaches to forecasting. The study supports basic theoretical research on the problem of prediction in international large-scale assessment data and contributes to it by adding results from one of the largest international large-scale assessments in education. The unique contribution pertains in particular to the length of the survey and a fine grain of curriculum data crucial for policy analysis. It presents a novel approach to utilize large-scale assessment data for forecasting and policy analysis.
In instructional research, the primary focus lies on identifying teacher behavior that positively influences relevant student outcomes. Analytical models in this field are typically complex, since (a) the involved constructs are measured using multiple indicators, (b) research questions on the classroom level require the application of multilevel models, and (c) the outcome variable should be observed at two time points at least, in order to infer that classes develop differently under different forms of instruction. Various multilevel models can be applied for analyzing such research questions. Two prominent approaches are (1) covariance analytical approaches in which the outcome at the second measurement occasion is controlled for by regressing it on the outcome at the first measurement occasion and (2) latent change-score models, where the change in the outcome between the two measurement occasions is modeled as an additional latent variable. Both approaches have been widely discussed regarding their differences and respective assumptions in models without a multilevel structure (Allison, 1990; Holland & Rubin, 1982; McArdle, 2009). The aim of this contribution is to apply them to the field of instructional research and to outline under which specific circumstances which model is more appropriate with respect to the underlying assumptions, and which inferences they each allow. We use empirical examples to illustrate these differences. We further simulate various data sets in order to examine influences of time specific, stable and random error components of the measured variables on both levels of measurement.

KÖHLER, CARMEN ¹
Hartig, Johannes ¹

Multilevel Models for Evaluating the Effectiveness of Instruction: ANCOVA vs. Change-Score Approach

¹ German Institute for International Educational Research (DIPF), Germany
carmen.koehler@dipf.de

Keywords:
Conditional Model, Latent Change, Multilevel Modeling
With cluster randomized trials complete groups of subjects are randomized to treatment conditions. An example is a study on the effectiveness of neighbourhood-level interventions to improve quality of life in impoverished neighbourhoods. Here residents are nested within neighbourhoods and outcomes may be measured on residents (e.g. perceived safety) and the level of the neighbourhood (e.g. crime index).

The optimal design determines the number of neighbourhoods and the number of residents per neighbourhood in the intervention and control conditions. It is found by taking a cost constraint into account: costs are associated with implementing the intervention or control in a neighbourhood, and with taking measurements at the resident and neighbourhood level. The optimal design is found such that the effect of intervention is estimated with highest efficiency, and the total costs do not exceed the budget that is available.

The design that is optimal for the outcome at the resident level is not necessarily optimal for the outcome at the neighbourhood level. Multiple objective optimal designs are used to take both outcomes into account. The aim is to find a design that has high efficiencies for both outcome measures. A Shiny App that can be used to find the optimal design is demonstrated.

The optimal design ensures financial resources are used in the most efficient way and the power for finding an effect on intervention is maximized.
Repeated measures Anova or Manova are frequently used for analyzing event-related brain potentials. They are typically performed on averaged trials as a way of increasing the reliability of the electroencephalogram signal. Averaging, however, leads to information loss concerning the covariance matrix of random individual differences of participant treatment and time effects, which could be of substantive interest. Lack of adequate specification of the covariance matrix has also been shown to lead to inferential biases. The objective of the present study is to compare performance of this traditional approach with that of multilevel models which allow for an explicit modeling of these random effects using single trial measures. The numbers of stimuli and participants as well as the magnitude of variance and covariance components between participants, treatment by participant and time by participant interactions will be manipulated in a simulation of a facial perception experiment. Empirical power, Type I Error, and effect sizes will be obtained as performance measures. Results will be presented and discussed.

OLIVER-RODRÍGUEZ, JUAN CARLOS ¹
Moerbeek, Mirjam ²

Comparative Performance of Single Trial Multilevel Analyses of Event-Related Brain Potentials

¹ Universitat Jaume I, Spain
² Utrecht University, The Netherlands
oliverr@uji.es

Keywords:
Event-Related Potentials, Multilevel Analysis
SESSION Structural Equation Modeling

Chair: Keith Widaman

Wednesday, July 25, 16:30 - 17:50
Salon Hölderlin

Unreliability Has Important Negative Effects: Correcting May Be Easier Than You Think
KEITH WIDAMAN

Multilevel SEM for Discrete Data With the Pairwise Likelihood Estimation Method
MARISKA BARENDSE, Yves Rosseel

An Alternative Estimation Method for Multilevel SEM Based on Factor Scores
INES DEVLIEGER, Yves Rosseel

Omitted Cross-Loadings in Nonlinear SEM: A Monte Carlo Study
KARINA RDZ-NAVARRO, Karin Schermelleh-Engel
Unreliability can have important negative effects on parameter estimates in multiple regression analysis, a problem that has long been understood, is often largely ignored, but recently has returned to prominence. In this presentation, the population basis of the effect will be demonstrated, and simulations will demonstrate the magnitude of effects on Type I error rates, Type II error rates, and parameter coverage rates. A more nuanced evaluation of the problem will be presented, with comparisons of effects on a target predictor based on perfect or imperfect reliability of the other predictor. Correcting for unreliability may be easier than you think, but is not without caveats. One way to correct for unreliability is to disattenuate correlations for unreliability and perform multiple regression analyses on these corrected estimates. Examples will demonstrate that this can resolve the pernicious effects of unreliability. The presentation then will be extended to path analysis with the analysis of an educational status attainment data set that has been used as a classic example since the early days of structural equation modeling. In the past, the standard approach to path analysis has been taken with this data set, an approach that fails to correct for unreliability. By correcting for unreliability, model fit can be affected in substantively important ways, parameter estimates are altered, and the ultimate form of the acceptable path model is altered in fundamental ways. The major caveat: all of these effects of correcting for unreliability require optimal estimates of reliability. This implies that some commonly used reliability indices, such as coefficient alpha, may not be recommended; model-based estimates, such as coefficient omega, might be preferred. Sensitivity analyses can inform about whether choice of reliability estimate has a notable effect on model fit and parameter estimates.
Social and behavioural research frequently involves multilevel data with a large number of latent variables (i.e., random slopes, random intercepts, and hypothetical constructs). Current full-information approaches for discrete data typically involve computationally intensive numerical methods (e.g., adaptive Gauss-Hermite quadrature). Alternatively, in the Pairwise Likelihood (PL) estimation method (Jöreskog, & Moustaki, 2001), the full likelihood is replaced by a sum of (bivariate) pairwise likelihoods, which are easier to handle. In this presentation, we will examine the ‘wide format’ or ‘multivariate’ approach to multilevel data. In this approach, all the individuals in a cluster are row-wise displayed and analysed (see Bauer, 2003, and Mehta & Neale, 2005 for continuous data). In a simulation study using discrete data, we will compare the ‘wide format’ PL estimation method with the multilevel weighted least squares (WLS) approach (Asparouhov & Muthen, 2007) and the multilevel marginal maximum likelihood approach (Hedeker & Gibbons, 1994) under different conditions (i.e., sample size, model misspecification, and balanced/unbalanced data). Overall, results show that PL estimation in the ‘wide format’ approach is quite close to the multilevel marginal maximum likelihood estimates, which is often considered to be the golden standard. During this presentation we will illustrate the use of the ‘wide format’ approach for discrete data and discuss the advantages and disadvantages of the different estimation methods with multilevel data.

BARENDSE, MARISKA ¹
Rosseel, Yves ¹

**Multilevel SEM for Discrete Data With the Pairwise Likelihood Estimation Method**

¹ Ghent University, Belgium

Mariska.Barendse@ugent.be

Keywords:
Discrete Data, Mixed Effects, Multilevel Analysis, Pairwise Likelihood, Structural Equation Modeling
An Alternative Estimation Method for Multilevel SEM Based on Factor Scores

UGent, Belgium
ines.devlieger@ugent.be

Multilevel SEM is an increasingly popular technique, used to analyze data that is both hierarchical and contains latent variables. When using the within-between framework, the parameters are usually jointly estimated using a maximum likelihood estimator (MLE). This has some drawbacks. First, a large number of clusters is needed to obtain unbiased estimates. Second, misspecifications in one part of the model, may influence the whole model. To overcome these issues, we propose a stepwise estimation method, which is an extension of the Croon method for factor score regression (Croon2002). A factor analysis is performed for every latent variable, resulting in factor scores. Next, the between- and within-cluster covariance matrices of these factor scores are calculated and corrected using the formulas of Croon. New data is simulated using these corrected covariance matrices, which can be used in subsequent analyses, such as multilevel regression or multilevel path analysis. A simulation study was set up to compare this new estimation method to the standard MLE. The results of uncorrected multilevel factor score path analysis were also considered. Two software packages were used to perform MLE, namely lavaan and Mplus. Five criteria were considered, namely bias of the regression parameters at the within level and between level, bias of the within and between variance and the proportion of successful replications. No major differences were found between lavaan and Mplus. The uncorrected path analysis resulted in biased estimates of all regression parameters and variance components. On the within level, the Croon method and MLE resulted in very similar, unbiased estimates for the regression parameters and variance components. The proportion of successful replications was also very similar for both methods. On the between level, the Croon method outperformed MLE when the number of clusters was low. In conclusion, the Croon method seems to be a promising alternative to MLE.
Researchers are often interested in detecting nonlinear effects using nonlinear structural equation modeling (SEM). Generally, it is assumed that the measurement models of the latent predictor variables are unidimensional, i.e. that all indicator variables load only onto a single latent variable, although this may not always be true. There already exists strong evidence for linear SEM that omitted cross-loadings might lead to biased covariances between latent predictor variables as well as biased structural parameter estimates (Bandalos, 2014; Hsu, Troncoso-Skidmore, Li & Thompson, 2014; Li, 2016). However, until now effects of omitted cross-loadings have not been investigated in the context of nonlinear SEM. Given that omitted cross-loadings affect the estimates of factor variances and covariances, and that the estimation of nonlinear effects depends on these estimates, it might be expected that omitted cross-loadings will also affect estimates of latent interaction and quadratic effects. In a Monte Carlo study using the LMS method (Klein & Moosbrugger, 2000) of the Mplus program, we investigated the effects of omitting cross-loadings on estimates of nonlinear effects by varying the number, size, and sign of the secondary loadings, the number of nonlinear effects, and the size of the latent covariance. Our results indicate that cross-loadings may lead to severely biased parameter estimates. We will show under which conditions spurious nonlinear effects occur (type I error rates) or existing nonlinear effects vanish (power rates). In order to detect omitted cross-loadings, an empirical researcher might think of testing the fit of the measurement models using the likelihood ratio test for linear SEM, because a global test for nonlinear SEM does not yet exist. However, we will also demonstrate that despite a good model fit undetected omitted cross-loadings may bias the nonlinear effects.

Omitted Cross-Loadings in Nonlinear SEM: A Monte Carlo Study

1 University of Chile, Chile
2 Goethe University Frankfurt, Germany

rdznavarro@uchile.cl

Keywords:
Cross Loadings, Interaction Effects, Nonlinear Structural Equation Modeling, Quadratic Effects
SYMPOSIUM New Developments in Mokken Scale Analysis

Chair: Andries van der Ark

Wednesday, July 25, 16:30 - 17:50
Salon Novalis

Introduction to Mokken Scale Analysis
KLAAS SIJTSMA

Checking Assumptions in Two-Level Mokken Scale Analysis
LETTY KOOPMAN

Two-Level Mokken Scale Analysis: The State of the Art
ANDRIES VAN DER ARK

Using Mokken Scaling Techniques to Evaluate Educational Assessments
STEFANIE WIND
Over the past decade, Mokken scale analysis (MSA) has shown a quickly growing popularity among researchers from many different research areas. This introduction discusses a set of techniques and a procedure for their application, such that the construction of scales that have superior measurement properties is further optimized, taking full advantage of the properties of MSA. First, I define the conceptual context of MSA, discuss the two item response theory models that constitute the basis of MSA, and discuss how these models differ from other IRT models. Second, I discuss dos and don'ts for MSA; the don'ts include misunderstandings frequently encountered in applications of MSA. Third, I discuss a methodology for MSA on real data consisting of a sample of persons who have provided scores on a set of items that, depending on the composition of the item set, constitute the basis for one or more scales, and I use the methodology to analyze an example real-data set.
The nonparametric IRT models that underlie Mokken scale analysis consist of four main assumptions: unidimensionality, local independence, monotonicity, and invariant item ordering. These assumptions imply certain observable properties of the data. For example, local independence and monotonicity imply conditional association; for dichotomous items scores, monotonicity implies manifest monotonicity; and invariant item ordering implies manifest invariant item ordering. Mokken scale analysis provides methods to investigate the assumptions of the nonparametric IRT models by investigating the observable properties. When dealing with multi-rater data, some adjustments of the assumptions are necessary. For example, for multi-rater data, the monotonicity assumption concerns the latent trait of the subject combined with the rater effect. In addition, multi-rater data require a different way to estimate the item probabilities. As a result the methods that are used to investigate observable properties must be adapted for multi-rater data. I will discuss the necessary adaptations to make the methods from Mokken scale analysis useful in a multi-level context, and I will discuss how these adaptations may be implemented.

KOOPMAN, LETTY

Checking Assumptions in Two-Level Mokken Scale Analysis

1 University of Amsterdam, The Netherlands

V.E.C.Koopman@UvA.nl

Keywords:
Item Response Theory, Item Response Theory, Mokken Scaling, Multilevel Modeling, Nonparametric Statistics
Currently, Mokken scale analysis for two-level data is being developed. It is a scaling procedure that allows test constructors to investigate the scalability, reliability, and validity of measurement instruments producing two-level data. This talk provides an overview what has been achieved so far. I will first discuss the types of data for which two-level Mokken scale analysis is a useful procedure, and I will briefly discuss underlying nonparametric IRT models. Then, I will discuss the estimation of scalability coefficients and their standard errors, and finally I will demonstrate two-level Mokken scale analysis using a multi-rater dataset.
The purpose of this study is to illustrate and consider the use of Mokken Scale Analysis (MSA) a method for evaluating the quality of educational assessments, including multiple-choice (MC) assessments and rating quality in rater-mediated educational performance assessments. I focus on the following questions:

1. What information does MSA provide about the psychometric quality of MC items? 2. What information do traditional applications of MSA provide about rating quality in rater-mediated performance assessments? 3. What information does an adjacent-categories scaling procedure adapted from MSA provide about rating quality in rater-mediated performance assessments? 4. What are the practical implications for researchers and practitioners of using MSA as an evaluative framework for educational assessments?

To address these questions, I used two data sources: (1) middle school students’ responses to an MC-format engineering design process assessment, and (2) ratings of middle school students’ written compositions composed during an administration of a rater-mediated writing performance assessment. I analyzed the first dataset using a traditional application of dichotomous MSA to the MC items. For the performance assessment, I calculated indicators of rating quality adapted from Molenaar’s original polytomous MSA models and an adjacent-categories approach to MSA (ac-MSA). Specifically, I examined indicators of rater monotonicity, rater scalability, and invariant rater ordering using the original models and the ac-MSA adaptations. I discuss the implications of considering rating quality from the perspective of MSA and ac-MSA.

Together, the results indicated that MSA provides diagnostic information about the psychometric quality of individual MC items and individual raters that can provide valuable insight during assessment development and revision. Implications for research and practice are discussed.
Thursday, July 26
KEYNOTE Towards a deeper understanding of the effectiveness of interventions: New methods based on structural equation models and causal inference (Axel Mayer)

Chair: Rolf Steyer

Thursday, July 26, 08:30 - 09:15
Saal Friedrich Schiller

Traditionally, the majority of studies investigating the effectiveness of interventions focused on the average effect, but there is much more to learn about the effects of a treatment or an intervention. Researchers are for example interested in heterogeneity of effects, in subgroup effects, or in conditional effects given values of one or multiple covariates. In addition, there is interest and need for visualizing the effects of interventions and in making the key findings from intervention studies usable for selecting the best treatment for a concrete person. In personalized medicine and related fields, attention is shifting towards estimation of interindividual differences in effects and there is a variety of statistical approaches that can be used for this purpose. However, the investigation of conditional effects and interindividual differences in the effects is particularly challenging in the social and behavioral sciences, because many constructs-of-interest such as depression or anxiety are latent variables. In addition, the selection of variables for the analysis is crucial, and modeling conditional effects oftentimes requires interactions and potentially non-linear relationships. In this keynote, I will bring together concepts from the causal inference literature and from structural equation modeling to allow researchers to gain a deeper understanding of the effectiveness of interventions based on latent variable models. I will use causality conditions and effect definitions from the causal inference literature and show how multigroup structural equation models with stochastic group sizes can be used to estimate the effects of interest in experimental and quasi-experimental studies. The new approach (and the accompanying open source R package) is termed EffectLiteR approach Several empirical examples from psychology and educational science are used to illustrate the proposed approach. Furthermore, I will show that many popular methods like ANOVA or moderation analysis are special cases of the general multigroup SEM approach for analyzing treatment effects. A SEM-based approach also has the advantage that many recent advancements that have been made in this area, like robust estimators and standard errors, modern fit statistics, and measurement models can be used for estimating causal effects. Finally, I will show some extensions of the proposed model, namely how to include propensity scores in the analysis, how it can be extended to multilevel SEM approaches and how Bayesian non-linear structural equation models can be used to deal with latent interactions and non-normally distributed latent variables.
SESSION Large Scale Data

Chair: Steffi Pohl

Thursday, July 26, 09:30 - 10:30
Saal Friedrich Schiller

**Item-Person Mismatch and Parameter Recovery Accuracy in Sparse Multi-Matrix Booklet Designs**
ANTA AKURO, Martin Brunner, Steffi Pohl

**Measurement Invariance of the Academic Performance for Fifteen Countries With the Alignment Method**
Pamela Woitschach, Bruno Zumbo, Ruben Fernandez-Alonso, JOSÉ MUÑIZ-FERNÁNDEZ

**Treatment of Measurement Error and Missing Data Using Nested and Non-Nested Multiple Imputation**
SIMON GRUND, Oliver Lüdtke, Alexander Robitzsch
Multi-matrix booklet designs are important in educational large-scale assessments as they help reduce respondents’ burden and enable test administrators to save time and money. However, accurate or efficient parameter recovery from response data is a central problem when analyzing data from multiple matrix booklet designs in conjunction with IRT. Important factors such as sample size and the match between person and item location parameter distributions could influence parameter recovery accuracy. This study investigates the degree to which person and item parameters are recovered as a function of matrix sparseness and sample size when using balanced incomplete block multi-matrix designs with varying degrees of alignment between the person and item location parameter distributions. To achieve this, data was simulated where person and item location parameters are generated from a population assuming a normal distribution with unit variance. The mean for the item location parameter is fixed at 0 in all conditions while that for the person parameters vary to give differing levels of mismatch ($\mu = 0, 0.2, 0.4, 0.8, 1.2, 1.6, 2.0$). The R package irttools (Partchev, 2016) was used in simulating the data under a Rasch model; as well as, for estimating item and person parameters using MML. The sparse multi-matrix designs used were like those used by Gonzalez & Ruthowski (2010) but with a test length of 42 items and varying sample sizes of 300, 500, 1000, 3000, 4500 and 6000 examinees. The root mean squared error and bias between true and estimated parameter values were used to assess parameter recovery accuracy. The study design was fully crossed with 1000 replications used in each condition to ensure stable results. The results showed that parameter recovery accuracy was affected by the match between person and item location parameter distributions. However, the size of the effect was negligible as the difference between the RMSEs in the perfectly matched cases and the most mismatched cases was always in all conditions less than 0.02. As expected, parameter recovery accuracy decreased as sample size reduced or when the multi-matrix designs became sparser.
In the context of international assessments, the comparability of scores between countries is based on the assumption that the measures are equivalent. UNESCO’s Third Regional Comparative and Explanatory Study (TERCE) program reports on the results for mathematics, science and reading for 15 Latin American countries and the State of Nuevo León in Mexico. A standard reporting practice is to rank order the countries according to their performance levels in each of these three subjects. An implicit assumption in this ranking is that the measures are sufficiently invariant to allow an unconfounded interpretation. The objectives of our research were to investigate the use of a relatively newly developed psychometric method – the alignment method (Asparouhov & Muthén, 2014) – for the analysis of the measurement invariance and determine the comparability of the scores obtained in the assessment. The analysis was carried out with 82 items of the Science test applied to 61,921 students. The alignment method was used for the item pool of the test, under the MLR estimation strategy to test the approximate measurement invariance. The data analyses were performed with the Mplus 8 program. The preliminary results indicate that the alignment method based on a configural model automates the process of invariance measurement. In summary, the research shows the effectiveness of the use of the technique for the detection of invariance in complex samples, providing evidence of non-invariant items that may affect the validity of interpretations in cross-cultural comparisons.
In educational research, plausible values (PVs) are frequently used to correct measurement error and represent students’ latent achievement scores while taking background information, such as students’ interests or attitudes, into account (Mislevy, 1991). This method follows the multiple imputation (MI) approach of Rubin (1987) by considering latent variables as missing data, thus predicting achievement scores on the basis of a measurement model and background information.

However, this procedure requires that the background data are completely observed, raising the question of how best to treat measurement error and missing data. In the present talk, we consider different strategies for dealing with measurement error and missing data using MI. This includes the procedures currently employed in educational large-scale assessments such as PISA but also strategies involving nested and non-nested MI (Rubin, 2003) which are commonly used in secondary analyses in educational research. We present the results from a simulation study that compared these methods under different conditions (e.g., sample size, reliability, amount of missing data). We show that the procedures currently employed in large-scale assessments could lead to biased parameter estimates (e.g., correlations, regression coefficients) when the data are not missing completely at random. By contrast, nested and non-nested MI are shown to provide unbiased estimates even with systematically missing data (missing at random). In addition, we show that simplified procedures can often be used, which are based on nested MI but do not require the use of specialized pooling methods normally required to analyze the data obtained from nested MI. In this context, we provide recommendations for practice, emphasizing the different roles of researchers who are involved in the scaling of student achievement data and those who conduct secondary analyses.
SYMPOSIUM Differential Item Functioning (DIF) in Educational Settings: Methods, Simulations and Applications

Chair: Rudolf Debelak, Martin Tomasik

Thursday, July 26, 09:30 - 10:50
Salon Schlegel

A Regularized Moderated Item Response Model for Assessing Differential Item Functioning
ALEXANDER ROBITZSCH, Oliver Lüdtke

Calibration of a Criterion-Referenced Computerized Adaptive Test in Higher Education
HANNA KÖHLER, Andreas Frey, Sebastian Born, Aron Fink, Christian Spoden, Johannes Bauer

Differential Item Functioning in the Context of Multistage Testing
SEBASTIAN APPELBAUM, Thomas Ostermann, Martin J. Tomasik

A Flexible Method for the Detection of Differential Item Functioning
RUDOLF DEBELAK, Lennart Schneider, Achim Zeileis, Carolin Strobl
A Regularized Moderated Item Response Model for Assessing Differential Item Functioning

1 Leibniz Institute for Science and Mathematics Education / Centre for Internat, Germany
robitzsch@ipn.uni-kiel.de

Keywords:
Differential Item Functioning, Item Response Theory, Moderated Factor Analysis, Regularization

The evaluation of differential item functioning (DIF) is an important aspect in the evaluation of measurements. Very often, DIF is investigated for several categorical (e.g., gender or region) and continuous (e.g., social status) covariates and a statistical model including all covariates is warranted. In this paper, a moderated item response model (also labelled as moderated nonlinear factor analysis, MNLFA) for polytomous responses is investigated which allows the dependence of item parameters on both types of covariates (Bauer, 2017). The model could include all possible DIF effects which leads to a highly parametrized model. Alternatively, statistical tests can be employed in a modelling phase to select DIF parameters which should be estimated in the MNLFA model which is essentially a multi-step approach. To circumvent both unfavourable strategies, we propose to include regularization methods into the MNLFA model using penalized marginal maximum likelihood estimation to assess DIF items (see Tutz & Schaubberger, 2015, for a similar approach). Simulation studies and an application demonstrate the usefulness of the proposed method.
The use of digital technologies opens up new opportunities in the field of higher education, for both teaching and testing. Regarding testing (e.g. written exams), besides other advantages, digital technology makes it possible to introduce state of the art methods from Psychometrics and Educational Measurement into the daily practice in higher education. In particular, criterion-referenced adaptive testing (CRT-CAT) has the potential to make exams more individualized, more accurate and fairer. However, from a practical point of view, the calibration of the item pool needed for CRT-CAT poses a critical challenge since a separate calibration study is typically not feasible and/or sample sizes of one written exam are too low to allow for a stable estimation of item parameters. To overcome this problem of small sample sizes in one written exam, within an actual construction of a CRT-CAT for student competences in statistics, the same test items were presented at different university locations. Thereby, the number of responses per item is increased making scaling with item response theory models possible. The test takers at all locations attended lectures with comparable content. However, as students in one university share exactly the same learning opportunities, they might be more similar than students between universities. A possible result could be differential item functioning (DIF) due to testing location. To examine for which items DIF due to testing location does not occur (so that they can be used in a joint calibration), several analyses with the multi-facet Rasch model are conducted. The data collection is carried out at the end of the winter term 17/18. The analyses will be finished before the conference. On the basis of the results, it will be discussed if the calibration across several university locations of an item pool for such a specific construct, that is only recently generated in students through lectures, can be recommended for future studies.
Multi-stage tests (MST) based on item-response theory (IRT) are becoming more and more common in educational research. MSTs have properties of both linear and of computer adaptive tests (CAT). In a CAT scenario, it is evaluated after each question whether the test taker should be presented an item that is more or less difficult. In a MST scenario, the test taker is presented whole sets of questions. After one set is completed, an algorithm decides the difficulty of the next set. The aim of MST is to evaluate the test taker’s ability with fewer questions and/or with a higher precision as compared to linear testing.

A routine procedure in test fairness evaluation is the identification of differential item functioning (DIF). DIF is observed when individuals from different subgroups but with identical abilities have different probabilities of solving an item. Using the 2PL-IRT model, DIF can occur in two different forms, namely as uniform DIF and non-uniform DIF (also called crossing DIF). In uniform DIF, the item characteristic curves do not intersect while in crossing DIF they do.

For linear tests various statistical tests for DIF detection are readily available but most of these methods are not suitable for DIF detection in MST. One of the methods, namely the SIBTEST, was adapted to the CAT scenario (CATSIB) that also seems to work well in MST scenarios. However, CATSIB cannot detect crossing DIF that occurs quite often in 2-PL models. Our aim, therefore, was the implementation of CATSIB in an R package and to develop an R function for the detection and evaluation of CDIF in MST. We compared these R functions in a simulation study to check their statistical properties. We also conducted a DIF analysis study with real data from a MST of competencies in the entire population of students from grade 8 and 9 (N > 3000). Of special interest was the detection of DIF between students from the two grades based on the null hypothesis that there would be none.
Educational assessments are required to be fair. This includes that persons of comparable ability obtain comparable results in educational tests and that no bias against specific groups of respondents is present. Tests which show such a bias are said to exhibit differential item functioning (DIF). In practical measurements, the statistical framework of item response theory (IRT) provides tools for DIF detection.

We present a class of score-based tests, which are named M-fluctuation tests, for the detection of DIF in the IRT framework. These tests can be applied with a wide range of IRT models to test the hypothesis that the characteristics of test items, for instance their difficulty or item discrimination, are invariant with regard to the characteristics of test takers, for instance their age and gender. In contrast to alternative tests for DIF detection in IRT, M-fluctuation tests do not require the definition of groups of respondents (e.g., age groups) between which the parameters are compared, but are directly applied to person characteristics like age and gender.

Using simulated data, which are generated based on the two- and three-parametric logistic models, we first demonstrate that M-fluctuation tests are able to detect changes in the different model parameters (like the difficulty or the guessing parameter) between groups of respondents, but are conservative if these parameters are stable. We then present how software packages in the free statistical software environment R can be used to calculate M-fluctuation tests.
SESSION Factor Analysis

Chair: Florian Scharf

Thursday, July 26, 09:30 - 10:50
Salon Hölderlin

Orthogonal Versus Oblique Rotation in Temporal EFA for Event-Related Potentials
FLORIAN SCHWARF, Steffen Nestler

Common Factor Analysis and Principal Component Analysis: Competing Indeterminacies
KEITH WIDAMAN

The Number of Factors in Exploratory Factor Analysis
MAX AUERSWALD, Morten Moshagen

On the Influence of Processing Speed on Investigations of Structural Validity: A Simulation Study
KARL SCHWEIZER
SCHARF, FLORIAN 1
Nestler, Steffen 1

Orthogonal Versus Oblique Rotation in Temporal EFA for Event-Related Potentials

1 Leipzig University, Germany
florian.scharf@uni-leipzig.de

Keywords:

Temporal exploratory factor analysis (EFA) is widely used to reduce the dimensionality of event-related potential (ERP) data sets and to reduce the ambiguity with respect to the underlying components. Typically, EFA is conducted on a data matrix in which the columns are time points and data from all participants, electrodes and conditions are commingled in the rows.

The central goal of this procedure is to test whether there are differences in the factor scores (i.e., amplitudes) between the conditions. In the past, the risk of incorrect allocation of condition effects between factors has raised concerns. Simulation studies have shown that orthogonal rotation methods are more prone to this variance misallocation than oblique rotation methods. However, orthogonal rotations such as Varimax are still applied.

Here, we outline the reasons for the superior performance of oblique rotation from the perspective of EFA as a statistical model. Specifically, we show that factors in temporal EFA for ERP data are inevitably correlated due to the condition effects and the scalp topography. We also demonstrate these principles in a Monte Carlo simulation comparing orthogonal Varimax rotation with the oblique Promax and Geomin rotations.

In line with previous research and our mathematical derivations, Varimax rotation was prone to spurious cross-loadings between the factors. This pattern occurred even when the factors were uncorrelated across participants. Oblique rotation methods showed much weaker biases that increased as a function of the temporal overlap between the factors.

In order to circumvent correlated factors as a major cause of variance misallocation, oblique rotation methods should generally be the method of choice when applying temporal EFA to ERP data.
To date, comparisons between exploratory versions of common factor analysis (CFA) and principal component analysis (PCA) have tended to focus on differential parameter estimates across methods. The purpose of this presentation is to extend prior work to additional issues. When using any method of analysis, the analyst should understand: (a) how to perform analyses in reasonable, step-like fashion; (b) how to interpret parameter estimates; (c) how to simulate data from the mathematical model or representation; and (d) the relations between parameter estimates and key data from which they were derived. In general, these four aspects of analyses are well understood when using CFA. However, I submit that only the first – how to perform analyses in step-like fashion – is well understood in PCA, and the remaining issues are not well known in relation to PCA. The presentation will stress the nature and range of parameter estimates under CFA and PCA, discuss issues related to simulating data from each procedure, and outline relations between parameter values and the correlation matrices from which they were derived. One key distinction historically stressed involves indeterminacy: that CFA has several crucial indeterminacies (e.g., rotational indeterminacy, factor score estimation), whereas PCA provides a determinate solution. Investigation of points (c) and (d) uncovered heretofore unaddressed, pernicious problems and/or indeterminacies associated with PCA. The nature of these problems with PCA will be illustrated. A major conclusion is that PCA offers a weak basis for scientific generalization in precisely those situations in which it is typically used, situations in which between 3 and 6 variables load highly on each component. If exploratory analyses should be used to hone measurement models for subsequent confirmatory analyses in new samples, CFA is the appropriate exploratory method to use, and the use of PCA should be strongly discouraged.
Exploratory factor analysis (EFA) is a widely used statistical method to study the underlying latent structure of a large number of observed variables, especially if there is no strong a priori justification for a particular theoretical model. Many criteria have been suggested to determine the correct number of factors. In this study, we present an extensive Monte Carlo simulation comparing traditional parallel analysis (PA), the Kaiser-Guttman criterion, and sequential $\chi^2$ model tests (SMT) to four recently suggested methods: revised PA, comparison data (CD), the Hull method, and the Empirical Kaiser Criterion (EKC). We manipulated the number of latent factors, the correlation among the factors, the number of items per factor, the magnitude of loadings, the underlying distribution, the presence of cross-loadings, minor factors, and the number of observations. No single extraction criterion performed best for every factor model. In unidimensional and orthogonal models, traditional PA, EKC, and Hull consistently identified the correct number of factors, even in small samples. Models with correlated factors were more challenging, where CD and SMT outperformed other methods especially for scales with fewer items. Given that the correct number of factors was reliably retrieved when SMT and either Hull, EKC, or traditional PA indicate the same number of factors to retain, we suggest that investigators first apply these methods to determine the number of factors. When the results of this combination rule are inconclusive, CD performed best. However, disagreement also suggests that factors will be harder to detect, in which case we recommend a sample size of $N \geq 500$. 
A time limit in testing prevents slow participants from reaching their highest possible scores, (besser mit Komma) whereas fast participants can reach their highest possible scores within the available time span. The consequences of such time limits for the investigation of the structural validity by means of confirmatory factor analysis are investigated in a simulation study. The following questions are addressed: Does the time limit in testing impair model fit in investigating structural validity? Does the representation of the effect prevent the impairment of model fit? Is it possible to identify and discriminate this method effect from another method effect? An important characteristic of the study is the assumption that omissions due to the time limit in testing reflect the participants’ processing speed. Four sets of 500 matrices showing a strong effect, a medium effect, a weak effect and no effect due to a time limit were generated and investigated. Impairment of model fit resulting from the effect due to the time limit was signified by some fit indices only. The inclusion of a factor for representing processing speed into the model of measurement in confirmatory factor analysis improved model fit. But there was no full compensation for the impairment. The precise representation of the effect enabled the discrimination of the effect due to the time limit from a uniform effect.

Schweizer, Karl ¹

On the Influence of Processing Speed on Investigations of Structural Validity: A Simulation Study

¹ Goethe University Frankfurt, Germany

K.Schweizer@psych.uni-frankfurt.de

Keywords:
Structural Equation Modeling
SESSION Applied Statistics

Chair: Thomas Schäfer

Thursday, July 26, 09:30 - 10:50
Salon Novalis

Cohen Revised: Empirical Redefinition of the Conventions for Interpreting Effect Sizes in Psychology
THOMAS SCHÄFER

Introducing Indigenous Methodology Into the Practice of the European Social Research
JUSTYNA PILARSKA

The Impact of Test-Review Models on Improving Tests and Testing: The Case of Spain
Hidalgo M. Dolores, HERNÁNDEZ ANA

Neets in French Labour Market: A Multidimensional and Fuzzy Approach
CLAIRE BONNARD, Jean-François Giret, Yann Kossi
Effect sizes are the currency of psychological research. They quantify the results of a study to answer the research question and are used to calculate statistical power. The interpretation of effect sizes—when is an effect small, medium, or large?—has been guided by the conventional definitions Jacob Cohen suggested in his pioneering writings starting in 1962. But do Cohen’s suggestions stand up to comparison with the empirical distributions of effect sizes as they are really found in psychology? For the present analysis, 900 effect sizes were randomly drawn from the whole history of psychological research. The distributions of effect sizes revealed that effects are much larger than suggested by Cohen, warranting a redefinition of the standards of their interpretation. New benchmarks are provided for the effect sizes $r$, $d$, and eta-squared. In addition, large differences were found between psychological subdisciplines, calling for the careful use of general guidelines.
The talk will evolve around the issue of indigenous methodology and its application in the social as well as cultural studies within the practices of the European qualitative research. It puts emphasis on innovative trends in qualitative methodologies that focus on culturally diverse environments as exemplified by the cultural borderland of Bosnia-Herzegovina’s society, yet stemming from emancipatory paradigm and post-colonial critique. The contents of the presentation will entail some key points to the indigenous methodology, studied by the author in the selected aboriginal academic and educational centres in Australia this year during a research visit. Indigenous approach can be referred to as “an ethically correct and culturally appropriate, indigenous manner of taking steps towards the acquisition and dissemination of knowledge about indigenous peoples” (Porsanger, 2004). Yet, despite the fact of being originally designed to emancipate and empower the indigenous peoples (e.g. in Canada, Australia, New Zealand or North America), it bears paramount significance for the many culturally fragile (i.e. subject to marginalization or discrimination) groups in Europe (such as Bosnian Muslims, Roma communities in Slovakia, Arabic refugees, etc.). It can, therefore, serve as a great tool of empowering and subjectifying the researched individuals, whose cultural identities are manifested in at times full of tensions cultural borderlands, exposed to ethnic, religious and national diversity. Introducing indigenous methods and paradigms can significantly enrich culturally sensitive practices of the field researchers, who respect the research communities and acknowledge them as the active co-producers of knowledge. Taking the above into consideration, the author discusses applied, and empirically developed theoretical models concerning indigenous methodology that can be successfully and mutually beneficially used in European social sciences and humanities.
Tests are essential tools in educational and psychological assessment. In order to improve test quality, different associations across different countries have proposed to use Test-Review Models (see for example Prieto & Muñiz, 2000; Bartram, 2002; Evers, Braak, Frima, & Van Vliet-Mulder, 2009; Lindley, 2009; Nielsen, 2009, Evers et al., 2013; Hernández, Ponsoda, Muñiz, Prieto & Elosua, 2016). When applying these models, qualified professionals make a rigorous assessment of a number of tests, and the results of the assessments are made available to test users and professionals in order to improve testing and help them to make the right assessment decisions. In Spain, Prieto and Muñiz (2000) proposed the Spanish Test-Review model, which was revised in 2016 based on the updated EFPA model (see Hernández et al., 2016). Since the model was applied for the first time, in 2010, the National Tests Commission of the Spanish Psychological Association (COP) has periodically reviewed some of the tests most commonly used by Spanish psychologists. To date, six rounds of test reviews have been completed and a total of 65 tests have been assessed. In this study, we present the results of the sixth test-review edition, where 10 new tests have been evaluated. In addition, we assess the impact that after these years, the model and the review process has had on different stakeholders: a) professional psychologists, b) psychometricians and psychological-assessment scholars, and c) test builders and test publishing houses. We conclude with some actions that could increase the impact of the test-review models.
The NEET measure concerns young people that are neither in employment, nor in education or training. Since 2010, the European Commission has introduced this NEET measure to monitor the labour market and social situation of youth. This measure is supposed to represent the problematic labour market transitions among early school-leavers and more specifically to identify the more vulnerable one (Furlong, 2006). However, as revealed in a number of studies conducted, NEETs are a very heterogeneous population that cover a very wide range of situations, of which certain accumulate vulnerability factors: demotivated youths, unoccupied, living with parents, disadvantaged, looking for a career path, having families responsibilities, youths taking a year out (Eurofound, 2011). If France has a NEET rate close to the European and OECD average (16.6...
SESSION Bayesian Statistics

Chair: Herbert Hoijtink

Thursday, July 26, 11:30 - 12:50
Saal Friedrich Schiller

Computing Bayes Factors From Data With Missing Values
HERBERT HOIJTINK, Xin Gu, Joris Mulder, Yves Rosseel

Handling Ordinal Predictors in Regression Models via Monotonic Effects
PAUL BÜRKEN, Emanuel Charpentier

Bayesian Estimation for Cases of Empirical Underidentification
JONATHAN HELM

Operationalizations of Inaccuracy of Prior Distributions in Simulation Studies
MILICA MIOCEVIC
The Bayes factor is increasingly used for the evaluation of hypotheses. These may be traditional hypotheses specified using equality constraints among the parameters of the statistical model of interest or informative hypotheses specified using equality and inequality constraints. So far no attention has been given to the computation of Bayes factors in the presence of missing data.

This paper will show how observed data Bayes factors, that is, Bayes factors computed using only the observed data, can be approximated using multiple imputations of the missing values. After introduction of the general framework elaborations for Bayes factors based on priors specified using training data and Bayes factors based on default or subjective prior distributions will be given. It will be illustrated that the approach proposed can be applied using the packages MICE, Bain, and BayesFactor.
Ordinal predictors are commonly used in regression models. Yet, they are often incorrectly treated as either nominal or metric thus under- or overestimating the contained information. This is understandable insofar as general applicable solutions or corresponding statistical software is still missing. We propose a new way of handling ordinal predictors, which we call monotonic effects. Here, the regression coefficients are reparameterized in terms of a scale parameter $b$ taking care of the direction and size of the effect and a simplex parameter $\zeta$ modeling the normalized differences between categories. This ensures that predictions are monotonically increasing or decreasing, while changes between adjacent categories may vary across categories. This formulation nicely generalizes to both interaction terms and multilevel structure. Fitting monotonic effects in a fully Bayesian framework is straightforward with the R package brms, which also allows to incorporate prior information and to test whether the assumption of monotonicity is justified.

BÜRKNER, PAUL $^1$
Charpentier, Emanuel $^2$

Handling Ordinal Predictors in Regression Models via Monotonic Effects

$^1$ University of Münster, Germany
$^2$ Assistance publique – Hôpitaux de Paris, France

paul.buerkner@gmail.com

Keywords: Bayesian Statistics, Multilevel Modeling, Ordinal Data, Regression
HELM, JONATHAN

Bayesian Estimation for Cases of Empirical Underidentification

1 San Diego State University, USA

jhelm@sdsu.edu

Keywords:
Bayesian Statistics,
Multitrait-Multimethod, Structural Equation Modeling

The correlated trait-correlated method model (CT-CM) and the true-score multitrait-multimethod model (TS-MTMM) are two structural equation models that can summarize multitrait-multimethod data. However, these two models often produce inadmissible solutions (e.g., failed convergence, out-of-bounds parameter estimates), which may be due to empirical underidentification (i.e., the data do not contain enough information to produce a unique set of parameter estimates, even though the model may accurately represent the data generation process). This presentation describes how Bayesian estimation can alleviate these estimation problems. A large-scale simulation showed that Bayesian estimation produced admissible solutions for 99.99%.
The most controversial aspect of a Bayesian analysis is the selection of prior distributions. There have been several simulation studies evaluating the impact of inaccurate prior distributions on statistical properties of posterior summaries of model parameters (Depaoli, 2013; 2014; Miočević, Levy, & MacKinnon, under review). However, the findings of such simulation studies depend heavily on the operationalization of inaccuracy and informativeness of priors. This talk presents several possibilities for designing inaccurate priors for a simulation study, and communicates the results of a simulation study used to compare the consequences of two different conceptualizations of inaccurate priors on recommendations made for applied researchers. The talk offers several options for constructing inaccurate priors for simulation studies that mimic real-life scenarios for how applied researchers might inadvertently specify inaccurate priors. This project aims to demonstrate that as methodologists, we should include multiple types of inaccurate priors in simulation studies used to examine statistical properties of Bayesian methods with inaccurate priors.
SESSION Causal Inference

Chair: Rolf Steyer

Thursday, July 26, 11:30 - 12:50
Salon Schlegel

Average Effects Based on Regressions with Log Link: A New Approach with Stochastic Covariates
CHRISTOPH KIEFER, Axel Mayer

Ignoring Ignorability: Towards a Realistic Public Policy Evaluation
TRINIDAD GONZALEZ, Ernesto San Martin

How to Model Production in Psychology? A Bayesian Stochastic Frontier Structural Equation Model
RÜDIGER MUTZ

Bias in Estimating Treatment Effects of Latent Non-normal Distributed Outcome Variables With Binary Indicators
JAN PLOETNER, Rolf Steyer
Researchers oftentimes use a regression with logarithmic link function (e.g., Poisson regression) to evaluate the effects of a treatment or an intervention on a count variable. In order to judge the average effectiveness of the treatment on the original count scale, they compute so-called marginal or average treatment effects, which are defined as the average difference between the expected outcomes under treatment and under control. Current practice is to evaluate the expected differences at every observation and use the sample mean of these differences as point estimate of the average effect. Standard errors for average effects are then obtained using the delta method. This approach makes the implicit assumption that covariate values are manifest and fixed, i.e., do not vary across different samples. We present a new way to analytically compute average effects based on regressions with log link with stochastic and/or latent covariates and develop new formulas to obtain standard errors for the average effect. In a simulation study, we evaluate the statistical performance of our new estimator and compare it to the traditional approach. Our findings suggest that the new approach gives unbiased effect estimates and standard errors and outperforms the traditional approach in some conditions.
Evidence-based public policies is a widely accepted requirement (OECD, 2007). Although the search of objectivity is a reason (King, 2010), to ensure the public trust constitutes the main motivation (Holt, 2008). This leads to evaluate a public policy by maximizing the social welfare (Manski, 1996). Thus, the evaluation itself reduces to compare the probability of the outcome when all the units are under the treatment, \( P(Y_1|X) \); and the probability of the outcome when all the units are not exposed to the treatment, \( P(Y_0|X) \); here \( X \) characterizes contexts in which the public policy is intended to be applied.

These conditional probabilities are non-identified due to the fundamental problem of causal inference (Holland, 1986): \( P(Y_1|X, Z=0) \) and \( P(Y_0|X, Z=0) \) are non-identified, where \( Z=1 \) means that a unit is under the treatment, \( Z=0 \) otherwise. An identification restriction widely used is the strong ignorability condition (Rosenbaum & Rubin, 1983) that establishes that \( (Y_0,Y_1) \) is independent of \( Z \) given \( X \). This condition is not empirically refutable, but only justified by substantive considerations. However, there are two problems associated to this condition: at the logical level, the question is to know if the recommendations obtained under the strong ignorability condition are still valid under weaker assumptions. At the policy level, it would be relevant to realize to what extent the ignorability condition determines the policy maker behavior with respect to the implementation of the treatment. In this talk, we intend to enhance the scope of public policy evaluations by introducing partial identification restrictions leading to solve the fundamental problem of causal inference. More specifically, we show how a public policy can be evaluated from different scenarios. We show not only how each of these scenarios induce specific policy maker behaviors, but also to what extent one scenario allows to falsify the policy recommendations reached under a complementary scenario.
How to Model Production in Psychology? A Bayesian Stochastic Frontier Structural Equation Model

1 ETH Zurich, Switzerland

mutz@gess.ethz.ch

Keywords:
Bayesian Statistics, Causal Inference, Stochastic Frontier Analysis, Structural Equation Modeling

In the realm of Social Psychology or Work and Organizational Psychology it is often the case that groups, for example research teams, are observed which produce an output for a given input. Modeling this relationship as a simple prediction task miss the idea of production. It is not of interest to predict the average outcome of a unit, but to estimate for a unit the maximal possible outcome (frontier of production) for a given input and the deviation of the observed outcome from the maximal possible outcome (inefficiency). Stochastic Frontier Analysis (SFA), adopted from Econometrics, should be favored in this case toward ordinary regression. Considering psychological applications, however, four problems arise: First, the outcomes are often count variables (e.g., number of publications). Second, the outcomes are prone to measurement errors (e.g., number of publications for a unit vary among different bibliographic databases). Third, not only one, but multiple outputs are often reported (e.g., number of publications, number of Phds). Fourth, the input (e.g., funding) as a treatment might be confounded (e.g., scientific fields). A Bayesian Stochastic Frontier Structural Equation Modeling approach (BSFSEM) addresses all of these problems: The outcome variables are assumed to be Poisson distributed. A measurement model is assumed to assess several latent output and input factors. The structural part of the BSFSEM is defined by the stochastic frontier component as a skewed normal distribution model. The main input factor (e.g., funding) provides for a continuous treatment by adopting concepts of the theory of causal effects of Steyer and the General Propensity Score Matching. Besides the methodological concepts, a simulation study will show how the model behave under different sampling conditions. The Austrian Science Fund (FWF) provided for research reports of N = 1,046 funded projects in order to illustrate the proposal.
To evaluate the effects of a treatment, it is useful to estimate the average treatment effect (ATE). Often, the outcome does not only depend on the treatment, but on one or several (qualitative or quantitative) covariates and the interactions between treatment and covariates as well. To estimate the ATE, one can use multi-group structural equation modeling (SEM), which also allows to analyze latent outcome variables and latent covariates. We focus on the case of a single latent outcome variable, which is measured by several binary indicators (items) and we use the Rasch model as the measurement model. 

Previous simulation studies have shown that the ATE can be estimated unbiasedly if all covariates which the latent outcome variable depends on are included in the model. However, if some covariates are ignored the ATE estimate can be biased. This can even lead to a bias in a random experiment where the treatment and the ignored covariates are stochastically independent and consequently the true ATE is unbiased. One explanation for this bias is that ignoring relevant covariates can lead to a skewed distribution of the latent outcome in one or more treatment groups. Subsequently, using conventional estimation methods, like the maximum-likelihood-estimator, results in a biased estimate of the ATE, because they assume a normal distribution of the latent outcome variable. In order to investigate to which extent a non-normal distribution affects the ATE estimate, we conducted a simulation study in which we systematically varied the conditional distribution of the latent outcome in control and treatment. Results showed that skewed distributions only led to a bias if the distributions of the latent outcome variable differed in control and treatment. We will further discuss these results and their implications for estimating treatment effects.
SYMPOSIUM Response Time Modeling in Psychometrics

Chair: Steffi Pohl

Thursday, July 26, 11:30 - 12:50
Salon Hölderlin

Disentangling Missingness Due to a Lack of Speed From Missingness Due To Quitting
ESTHER ULITZSCH, Steffi Pohl, Matthias von Davier

Response Time Models for Automated Test Assembly
BENJAMIN BECKER, Sebastian Weirich, Dries Debeer

Response Times and Latent Response Style Classes in Noncognitive Measures
ARTUR POKROPEK, Lale Khorramdel, Matthias von Davier

A Finite-State Machine Approach to Extract Item Response Times From Questionnaire Item Batteries
ULF KROEHNE, Janine Buchholz, Frank Goldhammer
Missing values at the end of a test can occur for a variety of reasons: On the one hand, examinees may not reach the end of a test due to time limits and a lack of speed. On the other hand, examinees may not attempt all items and end the test early due to, e.g., fatigue or a lack of motivation. We use response times retrieved from computerized testing to distinguish missing data due to a lack of speed from missingness due to quitting. On the basis of this information, we present an approach that allows to disentangle and simultaneously model and account for different missing data mechanisms underlying not reached items. The proposed model combines research on missing data and research on response times. In doing so, the model a) supports a more fine-grained understanding of the processes underlying not reached items and b) allows to obtain less biased and more efficient ability estimates. In a simulation study we evaluate the proposed model and compare its performance to current state of the art models for not reached items. In an empirical study we show which insights can be gained regarding test taking behavior using this model.
In the recent past, Automated Test Assembly (ATA) methods have been developed to enable the automatic generation of comparable test forms. In high-stakes assessments, comparable test forms are required for the comparison of persons across test forms. As it is common practice in high-stakes tests to use a fixed time limit across test forms and to score not reached items as incorrect responses, test length is a crucial property of test forms. Research has shown that items differ not only regarding their average response times but regarding how discriminating they are for differences in participants’ speed. Van der Linden (2011) showed that the balancing of parameters retrieved via the Hierarchical Response Time Model is useful to achieve the same degree of speededness across test forms, even for different levels of speed. However, the proposed model is rarely used in practice, partly as the consequences of differential speed sensitive test forms in high-stakes settings have not yet been investigated.

In a simulation study we show that test forms with equal average difficulty and length but with different speed discrimination can yield different person parameter estimations for certain speed levels. This demonstrates that it is necessary to include speed discrimination parameters into the assembly of test forms to prevent the allocation of test forms to individuals from having an impact on the ability estimation. In a second simulation study we found that the approach of van der Linden (2011) can prevent bias in person parameter estimation. The approach was compared to ATA methods using average response times and no response time information. Both of the latter approaches resulted in bias in person parameter estimation. We conclude that using the Hierarchical Response Time model in Automated Test Assembly is crucial for the assembly of fair test forms.


**Response Time Models for Automated Test Assembly**

1 Institute for Educational Quality Improvement, Germany
2 University of Zurich, Switzerland

b.becker@iqb.hu-berlin.de

**Keywords:** Item Response Theory, Response Times, Test Assembly
The purpose of the study is to examine the relationship between response time and response styles in the assessment of noncognitive constructs. Personality constructs, attitudes, and other noncognitive variables are often measured using rating scales. These scales can be biased by respondents giving invalid responses in form of response styles (RS) due to low motivation, fatigue effects, or problems understanding the questions. RS are defined as respondents’ tendencies to respond in a systematic way independent of the item content (Paulhus, 1991). They can affect the dimensionality of the measurement (Chun, Campbell, & Yoo, 1974), the validity of survey data (cf. Baumgartner & Steenkamp, 2001; Dolnicar & Grun, 2009; Morren, Gelissen, & Vermunt, 2012) and the comparability of test scores. The current study is based on a multi-process IRTree approach (Böckenholt, 2012) and multidimensional extensions (Khorramdel & von Davier, 2014; von Davier & Khorramdel, 2013) to detect and correct for response styles in rating data. A multidimensional IRTree approach combined with mixture IRT models is applied to rating data from PISA 2015. This approach allows modeling respondents’ behavior closely by decomposing rating data into multiply nested response sub-processes (binary pseudo items) separating different types of RS from trait-related responses. Because the measurement of RS is not straightforward – not all respondents show RS and the ones who do may not show it to the same extent or in the same direction – mixture IRT models are applied using the pseudo items to differentiate between groups of respondents with different response behavior. The resulting latent classes are then used to identify respondents who show RS and are related to response times as covariates in the mixture distribution modeling approach.
A potential benefit from modeling response times is to increase measurement precision of constructs defined by responses if response times are systematically related. Response time models for test items (e.g., van der Linden, 2007) or Likert type questions (e.g., Ferrando & Lorenzo-Seva, 2007) require time measures at the item level, defined as time differences between the onset of item presentation and the answer. Response times can easily be recorded for each item under very general conditions (labeled as “one-item-one-screen” design by Reips, 2002) in computer-based assessments. However, until now there is no well-established method that allows deriving response times when multiple questions are presented on one screen simultaneously, a design often used for item batteries.

Based on a general framework for the analysis of log data using finite state machines (Kroehne et al., 2016), a method to extract response times for item batteries is presented. Instead of using the total time on screen, the proposed approach uses log-events like answer-change events to aggregate time differences between subsequent answers as response times at the item-level, while accounting for the self-selected order of responses within the battery. The proposed method is illustrated using data from the ICT familiarity questionnaire of the PISA 2015 context assessment. Item response times are used to increase measurement efficiency by using them as predictors of the measured constructs in latent regression models (Ramalingams, 2017). It shows that adding item-level response times increases the reliability of the generalized partial credit models, for instance, for the subscale “Use of ICT outside of school” from 0.781 to 0.794, while adding only the total time on screen does not (0.782). The implied relationship between the measured construct and time of this method as well as alternative psychometric approaches will be discussed with respect to response times for item batteries.
SESSION Experimental Design

Chair: Volker Kraft

Thursday, July 26, 11:30 - 12:40
Salon Novalis

Statistical Power in Pooled Time Series
MARCEL ELIPE MIRAVET, Patricia Flor Arasil, Francisco Herrero Machancoses, Jesús F. Rosel Remírez, Pilar Jara Jiménez

Understanding (the) Power in Designed Experiments
VOLKER KRAFT

Degrees of Freedom Approximations in Multilevel Meta-Analysis of Standardized Single-Case Experiment
LALEH JAMSHIDI, John M. Ferron, Mariola Moeyaert, S. Natasha Beretvas, Wim Van den Noortgate
With the aim to find the total number of weeks that the smoking behaviour is rooted in the organism, a daily register was taken during 12 weeks (84 days) counting the amount of cigarettes smoked by a sample of 62 Spanish university students, to subsequently carry out a study using pooled time series. The results show that the smoking behaviour has an AR model (2)(7)8, that is, the smoking behaviour has a 56 days memory. We want to check if our study has enough statistical power to confirm that our results obtained here are sensitive to the real values of our estimators. Given the absence of previous studies, 10 subjects of the initial sample were taken a posteriori, and when we analysed their data, an AR model (1)(7)1 was obtained (R2 = .50), having the lag 56 a statistical power of .249. Using GPower statistical software, the sample size needed was estimated to obtain a statistical power of .90 for the lag 56, resulting in a minimum of 1428 useful data, that is, 51 subjects, because when the dependent variable is lagged, each subject would only have 28 useful data (84-56). To sum up, we conclude with the affirmation that time series analysis has a poor statistical power, so samples for this type of analysis should be quite large. Furthermore, the ideal number of subjects to obtain an adequate statistical power and effect size should be checked by a previous study, or if that is not possible, a posteriori analysis.
Computer-generated or optimal designs provide an accessible approach to efficiently learn from data in many situations. A live demo of a designed experiment will focus on data visualization at several steps of the integrated workflow. Visualization does not only help to better understand interactions and variability in the factor space, but also to communicate findings derived from the analyzed experiment. Based on a real-world scenario, JMP Pro will be used for designing a repeated measures (split-plot) experiment and for the mixed-model analysis. Integrated simulators support a-priori power calculations during the design phase and empirical statistics based on modeling results during the analysis phase. The demo case and other resources will be shared to get everybody started with state-of-the-art tools for statistically designed experiments.

**KRAFT, VOLKER**

**Understanding (the) Power in Designed Experiments**

1 SAS Institute GmbH – JMP Division, Germany

Volker.Kraft@jmp.com

**Keywords:**
Multilevel modeling can be used to synthesize the results of multiple single-case experimental designs. We are interested to discover to what degree using different methods of estimating the degrees of freedom, including the default ‘containment’, the Satterthwaite, and the Kenward-Roger methods, in analysis of standardized raw data and standardized effect sizes can improve statistical inferences on the fixed effects. We also used the ‘sandwich estimator’ to evaluate the improvement in the standard errors of the fixed-effects’ estimates. The raw data were simulated using a three-level model that models a possible effect of the intervention on the level and on a time trend. The standardized raw data were synthesized in a three-level meta-analysis. We also calculated the ordinary least square estimations of the case-specific intervention’s effects on the level and on the slope, and combined these standardized effect sizes over cases and studies in two separate univariate three-level meta-analyses. Results indicate as expected that the quality of the fixed effect estimates is unaffected by the method of estimating the degrees of freedom and for all methods the estimations are less biased as the number of measurement occasions increase. The number of measurement occasions and number of studies have a significant impact on the standard error estimates’ bias, operationalized as the gap between the standard error estimates and the standard deviation of the fixed effect estimates: as the number of occasions and/or studies increase, the negative bias decreases significantly. Surprisingly, the standard error estimates of the mean effect sizes are more underestimated when sandwich estimation was applied, and as a result, the sandwich estimation procedure produced a too small coverage proportion of the confidence intervals for effect estimates. The approach for approximating the degrees of freedom did not considerably affect the coverage proportion of the confidence intervals.
STATE-OF-THE-ART The role of time in dynamic models of change (Manuel Völkle)

Chair: Noémi Schuurman

Thursday, July 26, 15:30 - 16:00
Saal Friedrich Schiller

To better understand psychological mechanisms, researchers are increasingly capitalizing on longitudinal studies. The lead-lag structure and the possibility to disentangle between-person and within-person sources of variance, are two major assets of longitudinal (panel) data. However, how to best exploit this information in data analysis and interpretation? In this presentation, I want to identify several common problems and show how to avoid them by paying closer attention to the role of time. I will begin with a short example that illustrates some of the typical problems and questions faced by applied researcher and practitioners. Second, I will distinguish between static versus dynamic and discrete versus continuous time modeling approaches and discuss their advantages and disadvantages in the study of psychological mechanisms. Third, I will review different approaches to dealing with between-person differences, highlighting their dual role as a potential source of confounding as well as a source of information to improve estimation and causal inference. I will outline a possible way to better integrate information on between person differences and within person changes in the search for causal mechanisms in future research and end with a discussion of current problems and limitations.
Optimal design allows for estimating parameters of statistical models according to important optimality criteria, e.g., minimizing standard errors of estimators. Thus, optimal designs may considerably reduce the number of experimental units, such as respondents or items in empirical studies. For a long time, optimal design has not received much attention within psychology, but meanwhile interest for this subject is rapidly increasing as such designs are needed in large scale assessment, e.g. PISA, or for adaptive testing.

In this presentation, first, fundamental principles of optimal design are introduced using well-known linear models, e.g. analysis of variance or multiple regression. The rationale of adaptive, Bayesian, and minimax designs needed for nonlinear models will then be outlined. Such designs are presented for fixed and random effects models, e.g. IRT models or growth curve models. Finally, two R packages for deriving Bayesian and minimax designs based on recently developed algorithms will briefly be demonstrated.
SESSION Structural Equation Modeling

Chair: Jonathan Helm

Thursday, July 26, 16:30 - 17:50
Saal Friedrich Schiller

Dust Yourself Off and Try Anew: Reproducing ANOVA Using SEM
JONATHAN HELM

Dealing With Hypotheses That Depend on the Scaling of Latent Variables
JULIANE WILCKE

STEFAN KLOSNER, Eric Klopp

The Interpretation of Parameter Estimates in Structural Equation Models
Stefan Klößner, ERIC KLOPP
This presentation demonstrates how to reproduce the results (e.g., F-values, p-values) from different kinds of analysis of variance (between subjects, repeated measures, and multivariate ANOVA) using structural equation modeling (SEM). The presented approach differs from prior approaches, which incorporated indicator variables (e.g., dummy variables, effects codes) into a single SEM (analogous to regression). The approach presented here translates the main effects, interaction effects, and distributional assumptions of ANOVA into a set of SEMs with specific equality constraints, and then reproduces the ANOVA by statistically comparing the SEMs (i.e., difference testing). The results are virtually identical (i.e., sample statistics and p-values are equivalent to the third or fourth decimal) across the two approaches for a range of empirical examples, and the models can be extended to relax distributional assumptions (e.g., homogeneity of variance and sphericity) underlying ANOVA. Therefore, this presentation provides researchers with a series of stepping stones for using SEM in place of ANOVA, may facilitate analyses that extend beyond mean differences.
Latent variables in a structural equation model must be scaled in order for the model to become identified. Scaling methods include setting one loading per latent variable to unity (fixed marker method), setting the variances of all latent variables to unity (fixed factor method), or setting the average of all loadings per latent variable to unity (effects coding method). Unfortunately, model parameters estimate different quantities under these scaling methods and thus have different interpretations (Klößner & Klopp, 2017). This talk explores the types of hypotheses about model parameters that are affected by the choice of scaling method, as well as options for testing one’s original hypotheses of interest in these cases.

WILCKE, JULIANE ¹

Dealing With Hypotheses That Depend on the Scaling of Latent Variables

¹ University of Kassel, Germany

juliane.wilcke@uni-kassel.de

Keywords:
Identification Constraints, Structural Equation Modeling
In multi-group and longitudinal studies, it is important to test for metric measurement invariance (MI). Recently, it has been pointed out that currently used test procedures for MI are not complete in the sense that additional assumptions about the referent indicator’s invariance are needed in order to conclude that actual data satisfy MI (Raykov, Marcoulides & Li, 2012, Educational and Psychological Measurement, 72, 954-974).

Introducing the new concept of proportional factor loadings (PFL), we show that tests for metric MI actually only test for PFL, because PFL is empirically indistinguishable from metric MI. More precisely, if the loadings in the population are only proportional over groups or over time, the manifest variables’ implied distribution is identical to one that stems from invariant factor loadings in the population. Thus, it is impossible to differentiate between metric MI and PFL based on empirical data only.

Using Monte Carlo studies, we demonstrate that the power to detect violations of metric MI drastically deteriorates the closer the factor loadings’ pattern of non-invariance comes to a PFL pattern, explaining results in the literature that uniform non-invariance is very hard to detect (Yoon & Millsap, 2007, Structural Equation Modeling, 14, 435-463). Furthermore, our results on PFL explain why the choice of a referent indicator does not affect the results of testing for MI (Johnson, Meade & DuVernet, 2009, Structural Equation Modeling, 16, 642-657), while tests about the equality of latent variables’ variances may lead to potentially wrong conclusions when the data only satisfy PFL, but not MI.

With respect to partial metric MI, we find that empirically, it is impossible to differentiate invariant indicators from non-invariant ones: one can only detect which indicators form subsets whose loadings are proportional over groups or time. For detecting these subsets of indicators, we develop the partition method, and show that it works well.
Estimated parameters in statistical models usually represent population parameters. A classical example is a regression model with observed variables. The estimated slopes and intercept represent the respective population parameters. But this is no longer the case when latent variables are invoked in the model. This results from the necessity to scale the latent variable or, stated otherwise, to impose identification constraints on the latent variable. Thus, the estimated parameters do not represent population values, but an algebraic expression combining the population values of the parameter of interest and the parameter(s) used to impose the identification constraint. In practice, things become even more complicated because there are three well-known methods for achieving identification: the fixed marker, the fixed factor and the effects coding method.

In the literature, the correct interpretation of estimated parameters has not received wide attention. An exemption is Raykov, Marcoulides and Li (2011, Educational and Psychological Measurement, 72, 954–974), who demonstrated that when using the fixed marker method in factor models, an indicator’s estimated loading represents the ratio of the indicator’s population loading to the referent indicator’s population loading.

The aim of this contribution is to derive the correct interpretation of estimated parameters in structural equation models for all three widely used scaling methods. First, we deduce the interpretation of all estimated parameters in confirmatory factor analysis models, extending the results of Raykov et al. to the other scaling methods and model parameters. Second, we deduce the interpretation of parameters in recursive structural equation models. Finally, we show how our findings help to explain some recent findings regarding the effects of ignoring measurement invariance for path coefficients in structural equation models (Guenole & Brown, 2014, Frontiers in Psychology, 5, 980).
SESSION Latent Variable Analysis

Chair: Heidelinde Dehaene

Thursday, July 26, 16:30 - 17:30
Salon Schlegel

Semiparametric Regression Models for Indirectly Observed Outcomes
HEIDELINDE DEHAENE, Jan De Neve, Yves Rosseel

Impact and Dimensionality: The Performance of Logistic Regression in Differential Item Functioning
HUI-FANG CHEN, Kuan-Yu Jin

Assessing Individual Change Without Knowing the Test Properties: Item Bootstrapping
Juan Botella, DESIRÉE BLÁZQUEZ, Manuel Suero, James F. Juola
Although it is not always obvious at first glance, research studies across different fields often care about latent variables. These are variables that are not directly observed, but rather theoretically postulated or empirically inferred from observed variables (also known as proxies). The reason for using proxies can be motivated by theoretical or practical considerations (e.g. availability of sophisticated devices). Examples include the body mass index as a proxy for body fat percentage and the Beck Depression Inventory-II as a proxy for depression. In the first part of this presentation we illustrate by examples that the relationship between the outcome of interest and the proxy can be non-linear. The majority of available methods (e.g. standard structural equation models), however, typically assume that this relationship is linear. We illustrate how slight deviations from linearity can have a substantial impact on the validity of these inferential procedures.

In the second part of this presentation we present a new methodology that no longer imposes this linearity restriction, but only relies on monotonicity. Our methodology originates from the combination of three major statistical concepts: measurement error, a semiparametric linear transformation model and binary regression. The result is a model that enables us to quantify the effect of observed covariates on a summary measure of the unobserved outcome. We propose to quantify the effect of a covariate on the outcome in terms of the probabilistic index, i.e. the probability that the outcome of one subject exceeds the outcome of another subject, conditional on covariates (Thas et al., 2012, De Neve and Thas, 2015). We evaluate the proposed estimators empirically in a simulation study.
Conventional differential item functioning (DIF) approaches such as logistic regression (LR) often assume unidimensionality of scales and match participants in the reference and the focal groups based on total scores. However, many educational and psychological assessments are multidimensional by design, and a matching variable using total scores that does not reflect the test structure may not be able to eliminate the impact of multidimensional items on DIF detection. Thus, we proposed using all sub-scores of a scale in LR for DIF detection and compared the performance with three alternative matching methods in LR, including using total score and individual subdomain score. The present study manipulated three factors (the test structure, group impact and the number of cross-loading items) to compare false positive (FP) and true positive (TP) rates. We assumed that 500 or 1000 participants in each group answered 20 items, reflecting two dimensions; the tested item (the 21st item) measured a single or two domains; group impact included no impact, one group with a higher average ability on the first domain and with no impact on the other domain, and one group with a better ability on the first domain, whereas the other group was more capable on the second domain; 0, 20

CHEN, HUI-FANG 1
Jin, Kuan-Yu 2

**Impact and Dimensionality: The Performance of Logistic Regression in Differential Item Functioning**

1 City University of Hong Kong, Hong Kong
2 University of Hong Kong, Hong Kong

hfchen@cityu.edu.hk

**Keywords:**
Differential Item Functioning, Logistic Regression, Measurement Invariance, Multidimensionality
Introduction: A change between an individual’s responses in two administrations of a test reaches statistical significance when the two confidence intervals for the true scores do not overlap. When elaborating the confidence interval for the true score of a person in a test, the classical procedures involves knowing several properties of the test given a fixed sample, such as the population's variance and the reliability or internal consistency. Sometimes, those procedures cannot be employed because these properties are unknown or are not trustworthy. We propose to use the 'non parametric bootstrap’ method (Efron & Tibshirani, 1994) to the responses given by an individual to the items of the test in order to create confidence intervals for an individual’s true test score for situations in which classical procedures cannot be used.

Method and Results: Six databases containing the responses of several groups to one or more subscales have been analyzed: In two of them, there was not an expected change, while in the other four, a change in the criterion of interest was expected after an intervention. In each database two procedures have been applied to create the confidence intervals; a classical one, Estimating the True Score (ETS; Gulliksen, 1950), and the Bootstrap of items (BSI; Botella, Blázquez, Suero, & Juola, in press). The rates of significant change obtained with both procedures were very similar, suggesting that BSI is a promising solution when other methods cannot be applied.

Discussion: The BSI procedure has some advantages over ETS given that BSI requires no distributional assumptions, never needs to be adjusted because of inadequate confidence interval limit values, and the amplitude of the interval generated varies from one individual to another depending on the variability of their responses to the items. But BSI is still a very new procedure and, for assessing its performance, we still need evidence from different research contexts.
SESSION Item Response Theory

Chair: Sebastian Born

Thursday, July 26, 16:30 - 17:30
Salon Hölderlin

Comparing Fixed-Precision Multidimensional Computerized Adaptive Tests for Various Assessment Areas
SEBASTIAN BORN, Muirne Paap, Johan Braeken

Concurrent Adaptive Tests for Formative Assessments in School Classes
DANIEL BENG, Ulf Brefeld, Ulf Kröhne

The Dual Side of Classical Test Theory: The Geometry of the Axiom of Common Cause
ERNESTO SAN MARTÍN
Comparing Fixed-Precision Multidimensional Computerized Adaptive Tests for Various Assessment Areas

1 Friedrich Schiller University Jena, Germany
2 University of Groningen, The Netherlands
3 University of Oslo, Norway

sebastian.born@uni-jena.de

Keywords:
Computerized Adaptive Testing, Item Banks, Multidimensional Item Response Theory

Computerized adaptive testing (CAT) is an efficient approach to assemble tests: by selecting the most informative items for a given person, relatively short but precise tests can be constructed. In the last years, several studies have shown that taking into account the correlation structure when measuring multiple dimensions simultaneously can improve test efficiency even further. Up to now it is not entirely clear to what extent the research results on multidimensional computerized adaptive testing obtained in the field of educational testing can be generalized to fields such as health assessment where design factors differ considerably from those typically used in educational testing. In a simulation study, we examine the impact of different item bank properties which are thought to typify two assessment scenarios: health assessment (polytomous items, small to medium item bank sizes, high discrimination parameters) and educational testing (dichotomous items, large item banks, small to medium-sized discrimination parameters) on the performance of fixed-precision (variable-length) CATs. The average absolute bias across dimensions and the total test length were used to evaluate the quality of trait recovery and the measurement efficiency, respectively, for between-item multidimensional CATs and multiple unidimensional CATs. The study shows that the benefits associated with fixed-precision multidimensional CAT hold under a wide variety of circumstances.
Tests developed for summative assessment are administered to describe the performance of groups such as classes, for instance, relative to a comparable group. A formative use of such tests (e.g., the interpretation of student’s results in feedbacks to parents) necessitates a higher frequency of assessments and thus requires the increased measurement efficiency afforded by Computerized Adaptive Testing (CAT). CAT allows implementing tests with higher reliability which can be used to identify student needs and to make according adjustments to teaching and learning strategies in the course of ongoing learning events. However, the formative use requires that the test results are not only interpreted at the construct level, but rather at the level of individual items. The sizable item pools required for efficient CAT and the personalization of resulting tests mean that a huge number of different items are administered to students of one classroom, turning the item-level review into a daunting task for teachers. The talk shows how effective control of test overlap can be implemented by constraining adaptive item selection appropriately. Different from e.g. content constraints, the focus of overlap constraints extends across a group of students concurrently taking part in a testing session. The talk shows how test assembly with overlap can be formulated as an extension of the Shadow Test Approach, in which the overlap requirement can be understood as a constraint on a “Shadow Item Pool” from which individual students’ tests are assembled. The assumption of concurrent test-taking poses a number of challenges arising e.g. from individual differences in working speed and their correlation with proficiencies. Simulation studies are presented which explore the measurement properties of the proposed concurrent CAT under different assumptions on student proficiencies and working speed as well as group size, test length and target test overlap.
The Dual Side of Classical Test Theory: The Geometry of the Axiom of Common Cause

Faculty of Mathematics, Pontificia Universidad Católica de Chile, Chile
esanmart@mat.uc.cl

Keywords:
Classical Test Theory

In Classical Test Theory (CTT), the latent variable \( \theta \) is characterized by the Axiom of Local Independence (Lazarsfeld, 1950) or the Axiom of Common Cause (Laplace 1774; Reichenbach, 1956). According to Lazarsfeld (1950, p. 367), this axiom means that “all the within relationships between the [observed scores] should be accounted for by the way in which each [observed score] alone is related to the latent variable”. In Psychometrics, the interest focuses on the latent variable, which is “estimated” by the empirical Bayes estimator (EBE) \( \hat{\theta}|Y \). Its theoretical status, meaning and existence are of interest (Suppes & Zanotti, 1981; Borsboom et al., 2003). This paper makes a contribution to these aspects.

By endowing the CTT with a Hilbert space structure, we make a review of the principal axioms and results of CTT as established by Novick (1968) and Zimmerman (1975). Thereafter, we extend those results by exploiting the duality of \( \hat{\theta}|Y \) wrt \( E(\theta|Y) \). The following results will be discussed: 1) We define a dual reliability and we prove that it is equal to the standard reliability. As a consequence, we prove a more general Spearman-Browne result and we state that the quality of the EBE is directly measured by the reliability. 2) If \( \theta \) is the minimal subspace such that the weak version of the Axiom of Local Independence (Bollen, 2002), then the null space of the dual operator \( E(\theta|Y) \) is reduced to 0 and consequently \( E(Y|\theta)=\theta \); and reciprocally. This last result opens the door to identify the distribution generating \( \theta \). 3) Under the minimality of \( \theta \), the Axiom of Local Independence has a geometrical formulation, namely the orthogonal to \( Y \) have any element common with \( \theta \) except 0. 4) Finally, we make explicit linear subspaces of observables in between the subspace generated by the latent variable is contained. The main consequence is that the latent variable is an unobservable rather than an unobserved.
SESSION Applied Statistics

Chair: Patricia Flor Arasil

Thursday, July 26, 16:30 - 17:30
Salon Novalis

Measuring Intercultural Competence: Methodological Issues and Challenges
MAGGIE YUE ZHAO

Irritability and Its Memory: A Time Series Model
PATRICIA FLOR ARASIL, Marcel Elipe Miravet, Pilar Jara Jiménez, Francisco Herrero Machancoses, Jesús F. Rosel Remírez

Use of Low Cost Tools in Ergonomic Research of Mobile Restaurants in Western India
PRABIR MUKHOPADHYAY, Vipul Vinzuda, Thridev Suvarnan, Nakul Lakhtar
Increased globalization in education demands that students develop global perspective and intercultural competence in order to be competitive contributors in the global economy. Correspondingly, there has been a growing interest among educators, policy makers, and researchers in understanding and measuring students’ intercultural competence, also known as cultural intelligence.

When it comes to assessing intercultural competence, methodological issues and challenges remain. Conceptually, intercultural competence can be considered as a noncognitive attribute/trait and it contains attitudes/beliefs, knowledge strategies and behaviors. These conceptual dimensions often relate to motivation, self-efficacy and personality, which are interconnected. This creates complexity and unique challenges for assessing noncognitive attributes/trait, such as intercultural competence. Furthermore, with regards to scoring and reporting, most of the existing assessment instruments were developed and validated based on classical measurement theory. Alternatively, the modern measurement theory (i.e., item response theory) offers promising solutions to address issues that have been difficult to solve through classical methods. It offers a greater psychometric metric that brings assessment development to a new level of precision, efficiency and standardization, and provides greater flexibility and unique benefits for value-added measurement and group comparisons.

This presentation seeks to discuss methodological issues, challenges and possible solutions in relation to assessment development, data analysis and score reports of intercultural competence assessments, drawing on preliminary findings of a study on assessing intercultural competence for college students.
Diary registration used as a tool to collect data to explore the change of a person over time within natural settings has increased during the past years (Cranford et al., 2016). Daily data can be very useful to gather cyclical patterns and the possibility of studying this effect (Liu & West, 2016). Furthermore, emotions are subjective phenomena that initially were understood as internal states that could not be observed or measured. Emotions can be activated because of different reasons (Izard, 1977) and their duration is shorter than mood states. In fact, it is said that mood can last hours or days (Ekman, 1992; Frijda, 2009), but it’s not always easy to know why you are in a certain mood or in another because the event or object that has made you get in that mood may not be present at that moment, as Lazarus (1991) mentioned. Negative mood reflects irritability. We also wanted to prove the influence of days of the week, holidays, gender and age could have.

A total of 74 college students voluntarily participated. They had to register their mean mood irritability at the end of the day during a minimum period of 60 consecutive days. To do this we used a longitudinal design to collect the data, and a complex method to obtain the results: intensive longitudinal method with temporal series.

Results show that mood irritability has a memory of up to seven weeks. We didn’t find significant differences in days of the week, holidays, age, gender and the interaction between gender and sexual desire.
MUKHOPADHYAY, PRABIR¹
Vinzuda, Vipul²
Suvarnan, Thridev²
Lakhtar, Nakul²

Use of Low Cost Tools in Ergonomic Research of Mobile Restaurants in Western India

¹ Indian Institute of Information Technology Design and Manufacturing Jabalpur, India
² National Institute of Design, India

prabir@iiitdmj.ac.in

Keywords:
Anthropometrics, Direct Observation, Ergonomics, Photography

Mobile restaurants are very popular food joints in Gandhinagar city in Western Indian state of Gujarat. It's imperative that such mobile restaurants need to be designed with different human factors issues else drawing customers or increasing efficiency in the kitchen area could lead to decline in productivity and business for the people associated in this business. One such restaurant selling a very special snacks called "vada pau" and "dabelie" which is essentially a modified burger like food was analysed on the request of the restaurant association. The methodology applied was direct observation and activity analysis of both the customers and the cooks in the kitchen with an eye to locate the nature, quantum and exact location of problems related to ergonomic issues in the space. This methodology also gave insights into customer behaviour in terms of group formation, standing in the queue, postural analysis and reachability issues. This methodology was supplemented with still and videography of the entire site to record the postural issues and the time taken for performing different activities. To get further insights questionnaires both open and closed ended were also used. After collecting the data from the field the data was analysed in the laboratory to come up with concept of ergonomically designed spaces which were again tested on the grid board with mannequins and also on real users. The output of this research showed that it's possible to conduct reliable ergonomic research even by using low cost tools and this is pertinent in a country like India where usage of expensive tools for data collection are always not feasible.
Friday, July 27
In longitudinal modeling, maximum likelihood estimators of model parameters are consistent if missingness depends only on the covariates and/or observed outcomes. Such missingness processes are hence ignorable. When missingness depends on unobserved outcomes or on the random effects in mixed effects/multilevel models, it is said to be not missing at random (NMAR) and is no longer ignorable. For such NMAR missingness, joint modeling of the outcomes and missingness has been advocated, but these approaches rely on strong, unverifiable assumptions, such as parameteric specification of the missingness process. In this talk, I will show that minimal assumptions about missingness, such as whether it depends on random effects or contemporaneous observed/unobserved outcomes, often allows us to make the missingness ignorable. In other words, we can obtain consistent estimates of (some) model parameters using standard estimators, a concept also referred to as “protective” estimation. Perhaps surprisingly, one approach is to simply discard more data. Another approach is to change the estimator, for example, by switching from a random-effects model to a fixed-effects model (Skrondal & Rabe-Hesketh, 2014, Biometrika 101, 175-188). This is joint work with Anders Skrondal.
SESSION Item Response Theory

Chair: Christoph König

Friday, July 27, 10:00 - 11:00
Saal Friedrich Schiller

Reducing Sample Size Requirements of the 2PL With a Bayesian Hierarchical Approach
CHRISTOPH KÖNIG, Christian Spoden, Andreas Frey

Instructional Sensitivity of Polytomous Test and Questionnaire Items
ALEXANDER NAUMANN, Johannes Hartig, Stephanie Musow, Jan Hochweber

Estimation of a Multidimensional Item Response Model Using Bayesian Nonparametrics
FELIX NAUMANN
Given their complexity, item response theory (IRT) models require large samples for accurate item calibration and are considered primarily large-scale application methods. Therefore, approaches are needed which allow for a resource-efficient way to estimate complex IRT models, which yield accurate parameter estimates even in smallest sample sizes. Bayesian hierarchical IRT modeling with a non-centered parameterization of the item parameters, combined with a Cholesky decomposition of their covariance matrix, is a promising approach to increasing the accuracy of parameter estimation in small sample situations, when appropriate prior information is not available. Careful consideration is required especially regarding the choice of hyperprior parameters for the variance components of the item parameters $\tau_\alpha$ and $\tau_\beta$. In the context of the hierarchical 2PL model and small sample sizes ($N < 500$), this Monte Carlo simulation study investigates differences in sensitivity of the Inverse Gamma, the Cauchy, and the Exponential distribution regarding their impact on parameter accuracy across different specifications. Results show that estimation accuracy of the discrimination parameter is sensitive to different specifications of the Inverse Gamma, and robust against different specifications of the Cauchy and Exponential distributions. Differences in sensitivity between the three hyperprior distributions are most distinct for short test lengths ($k = 25$) and very small sample sizes ($N < 100$). Thus, the use of either the Cauchy or the Exponential distribution as hyperprior distributions for variance components, considerably reduces the sample size requirements of the 2PL model. This is advantageous when there are few items and respondents (e.g., in university exams), where the estimation of item parameter variance is often problematic because of sparse data, and presents resource-efficient possibilities for accurate item calibration in small sample situations.
Student achievement is widely used in educational research for drawing inferences on teaching. Valid inferences on teaching require that tests are instructionally sensitive, that is, capable of capturing effects of classroom instruction (Polikoff, 2010). Yet, despite numerous instructional sensitivity (InSe) measures to evaluate dichotomous items, none are available for polytomous items. Polytomous items may not only be sensitive as a whole, but also response categories may vary in their degree of InSe. Thus, the aim of our study is to provide measures for the InSe of polytomous items. We advance a longitudinal multilevel IRT model (LMLIRT; Naumann, Hochweber, & Hartig, 2014) to fit polytomous items. The LMLIRT model provides measures for two facets of InSe: a) global sensitivity (average change in classroom-specific item difficulty), and b) differential sensitivity (variation of change across classes). We combine the LMLIRT model with the partial credit model (Masters, 1982), using a) the standard parametrization for achievement items and b) the expanded parametrization for Likert items. So far, Likert items have been neglected when evaluating InSe. We exemplarily apply our model to achievement test and questionnaire data from the DESI (3613 students, 135 classes) and the InSe (815 students, 46 classes) studies. Results indicate that the model works well in empirical application. Polytomous items can be considered a) insensitive, if average change in location and thresholds is zero and there is no variation across classes, b) globally sensitive, if there is nonzero change in location or at least one threshold, c) differentially sensitive, if there is variation of change across classes in location or at least one threshold, and d) globally and differentially sensitive, if both b and c apply. We are confident that such information fosters valid use and interpretation when individual level student responses are used for drawing inferences on teaching.

**Instructional Sensitivity of Polytomous Test and Questionnaire Items**

1. German Institute for International Educational Research (DIPF), Germany
2. University of Teacher Education St. Gallen (PHSG), Switzerland

naumann@dipf.de

Keywords:
Educational Assessment, Item Response Theory, Multilevel Modeling
Estimation of a Multidimensional Item Response Model Using Bayesian Nonparametrics

1 LMU München, Lehrstuhl für Psychologische Methodenlehre und Diagnostik, Germany

felix.naumann@psy.lmu.de

Keywords:
Bayesian Statistics, Multidimensional Item Response Theory, Nonparametric Statistics

Parametric Item Response models do not always show acceptable fit to the data obtained from psychological tests. In these cases, one option is to resort to more flexible nonparametric models. Peress (2012) provides an identification proof for a very general Item Response model, which can be viewed as a multidimensional compensatory model with nonparametric Item Characteristic Curve (ICC). The subject of this talk is the application of Bayesian nonparametrics to the estimation of the ICC and the item- and person parameters of this model. A reparameterisation is proposed, which allows for a Bayesian formulation of the model. As the parameter space of the ICC is a function space, a Dirichlet Process Mixture of normal cumulative distribution functions can be chosen as a prior for the ICC. This allows for the derivation of all full conditionals of the joint posterior distribution of the parameters and thus for an implementation of a Gibbs Sampler.
SESSION Missing Data

Chair: Mario Lawes

Friday, July 27, 10:00 - 11:00
Salon Schlegel

Planned Missing Data Designs: Investigating the Efficiency of a Three-Method Measurement Model
MARIO LAWES, Martin Schultze, Michael Eid

Handling Missing Data in Single-Case Experiments
TAMAL KUMAR DE, René Tanious, Bart Michiels, Patrick Onghena

Testing Missingness for Continuous and Categorical Data
SERGUEI ROZINOV, André Berchtold
Planned missing data designs are an elegant way to incorporate expensive gold standard methods (e.g., biomarker, ambulatory assessment) and cheaper but systematically biased methods (e.g., self- and informant ratings) in research designs to ensure high statistical power while keeping the costs low. This talk outlines a planned missing data design with one expensive and two cheap methods (three-method measurement [3-MM] design). The statistical efficiency of this 3-MM design is investigated in a simulation study and compared to the efficiency of corresponding two-method measurement (2-MM) designs. In most conditions, planned missing data designs yielded higher statistical efficiency compared to complete-cases design. Within the planned missing data designs, 3-MM designs can increase statistical efficiency compared to 2-MM, when the additional cheap measure of the 3-MM design is inexpensive and shares only small amounts of method variance with the initial cheap measure as well as when the gold standard measure is highly expensive compared to the cheap measure. Implications for the needed sample sizes are discussed.
Single-case experiments have become increasingly popular in educational and behavioral research. However, the analysis and meta-analysis of single-case data is often complicated by the frequent occurrence of missing or incomplete observations in the data series. One reason for this large frequency of missing or incomplete observations in educational and behavioral single-case research might be that practitioners or participants are often required to record data themselves at regular intervals, sometimes even retrospectively, and that the human element in this process leads to missingness and incompleteness. If missingness or incompleteness cannot be avoided it becomes important to know which strategies are optimal if missing or incomplete data do occur because presence of missing or incomplete data or inadequate data handling strategies may lead to the experiments no longer "meeting standards" set by, for example, the What Works Clearinghouse. For the examination and comparison of several strategies to handle missing data, we simulated complete data sets for phase designs, alternating treatments designs, and multiple baseline designs. We then introduced missingness in the simulated datasets using multiple probabilities of data being "Missing Completely At Random". We evaluated the Type I error rate and the statistical power of a randomization test for the null hypothesis that there is no treatment effect, using the different strategies of handling missing data. We compared the operating characteristics for the original dataset (before the introduction of missing data points) with the operating characteristics for three strategies: (1) randomizing a missing data marker and calculating all reference statistics only for the available data points; (2) estimating the missing data points by using minimum mean square error linear interpolation; and (3) multiple imputation methods based on resampling the available data points.
The probability of having missing data in a survey is very close to 1. Their handling should be correct in order to have unbiased and consistent results, and to do so it is essential to know their type. Missing data are generally divided into three single types (Rubin, 1976): missing completely at random, missing at random, and missing not at random. Additionally, they can be a mixture of these mechanisms as well. The first step to understanding the type of non-observed information generally consists in testing whether or not the missing data are missing completely at random. Several tests have been developed for that purpose, but they have complications when dealing with non-continuous variables.

Our approach tests whether or not the missing data are missing completely at random using a regression model and a distribution test specific to the type of the incomplete variable. Formally, for a variable with missing data, we compare the predictions of the regression model given for observed data with those given for unobserved data. Consequently, our test can be applied both for continuous and categorical variables, which is not the case of the usual procedures. Simulations with six types of single and mixed missing data mechanisms were performed, linear and multinomial regression models with full and restricted information were used, and the quantity of missing data varied between 50%.
SYMPOSIUM Applying Subjective Bayes to Real Life Data

Chair: Fayette Klaassen

Friday, July 27, 10:00 - 11:00
Salon Hölderlin

Including and Assessing Expertise via Prior Probabilities.
FAYETTE KLAASSEN, DUCO VEEN

Bayesian Approximate Measurement Invariance: When You Have Too Little or Too Much Data.
KIMBERLEY LEK, SONJA WINTER

Expert-Weighted Prior Information: Applications in Psychology and Veterinary Medicine
HAIFANG NI, Mariëlle Zondervan-Zwijnenburg, DUCO VEEN
Including and Assessing Expertise via Prior Probabilities.

1 Utrecht University, The Netherlands

f.klaassen@uu.nl
d.veen@uu.nl

Keywords:
Bayesian Statistics, Elicitation, Prior Knowledge

The relative evidence for a set of hypotheses can be updated continuously with new data into posterior probabilities. In order to do so, (subjective) prior probabilities need to be specified for the hypotheses considered. Prior probabilities are often considered equal in applied research, while this might not be an accurate representation of researcher’s beliefs. This talk discusses the need for subjective prior probabilities in Bayesian hypothesis testing, and proposes an application to facilitate applied researchers (say, psychologist) to better understand and formulate prior probabilities based on subjective ideas.

In practice companies have many employees that are on the same level and are each considered experts in their field. Yet evidence for such assumed expertise is hardly ever obtained. We proposed to let experts predict new data in the form of a probability distribution, thereby explicitly incorporating both tacit knowledge and uncertainty. These predictions are then compared to actual new data and we adapted a prior-data conflict measure to assess appropriateness of the experts’ beliefs. We used this measure to rank regional directors in a large financial institution based on their capabilities to predict future turnover.
In Bayesian structural equation modeling (BSEM), informative, small-variance priors with mean 0 are used to replace exact zero constraints. One application of BSEM is in the testing of measurement invariance (MI). Traditionally, MI is accepted when exact zero constraints on the difference in intercepts and factor loadings over different groups (e.g. countries) or time hold. This exact approach works well with a reasonable number of countries or time points, but in large-scale applications it is cumbersome and leads to a frequent rejection. We show how BSEM can overcome the difficulties of exact MI in large-scale applications.

In addition to being beneficial in large-scale applications, BSEM estimation of MI can also be advantageous when samples are small. In such scenarios, traditional methods of testing MI can run into computational problems and fail to converge to a reliable solution. BSEM overcomes these computational issues and also allows for the introduction of priors on other parameters in the model. In this talk, we discuss an empirical application of Approximate MI using longitudinal survey data.

LEK, KIMBERLEY ¹
WINTER, SONJA ²

Bayesian Approximate Measurement Invariance: When You Have Too Little or Too Much Data.

¹ Utrecht University, The Netherlands
² University of California Merced, USA

k.m.lek@uu.nl
swinter@ucmerced.edu

Keywords:
Bayesian Statistics, Measurement Invariance
We systematically acquired prior knowledge for a structural equation model on the development of working memory in young adolescents with severe behavioural problems for which a subgroup frequently used cannabis. To collect prior information, a systematic search for meta-analyses, reviews, and empirical papers was conducted. A clinical and scientific expert weighted the information. Combined with general knowledge, the final prior distributions were constructed. Based on our experience, we present a set of general guidelines for collecting prior knowledge and formalizing it in prior distributions. Furthermore, we present two examples from veterinary medicine. First, the treatment effect of oral Glucosamine/Chondroitin on gait characteristics of aged horses was examined. Historical studies from across species resulting from a systematic review were weighted by the expert for their clinical relevance and incorporated into power priors. In another example, a random effects logistic regression diagnostic model was developed for detecting subclinical ketosis in dairy cows. Cluster level information about feed content and milk production was used by the expert to estimate the relative position of each herd regarding other herds in the population. The elicited judgement was incorporated into the priors for the random effects.
SESSION Longitudinal Data

Chair: Manuel Arnold

Friday, July 27, 11:30 - 12:50
Saal Friedrich Schiller

Individual Parameter Contribution Regression for Longitudinal Data
MANUEL ARNOLD

Generalized Continuous Time Models and the Continuous Time Rasch Model as an Example
MARTIN HECHT, Katinka Hardt, Charles C. Driver, Manuel C. Voelkle

The Association Between Depression and Education in UK Adolescents: A Cross-Lagged Panel Analysis
JOSÉ LÓPEZ-LÓPEZ, Rebecca Pearson, Liz Washbrook, Mina Fazel, Kate Tilling

LEONIE V.D.E. VOGELSMEIER, Jeroen K. Vermunt, Kim De Roover
Structural equation models (SEM) are widely applied in the behavioural and social sciences to analyse the relationship between observed and latent variables. A standard assumption underlying many SEMs is that parameter values are equal for all observations in the sample. Thus, researchers need to determine whether there exists relevant heterogeneity in their samples or they run the risk of reporting meaningless parameter estimates and inaccurate standard errors. During the last decades, several SEM extensions have been developed to identify and account for heterogeneity. One of those approaches is “individual parameter contribution” (IPC) regression, proposed by Oberski (2013). IPC regression is conducted in three steps. First, a theory-guided SEM is fitted. Second, the contributions of every individual to the model parameters are calculated based on the first-step model. Third, heterogeneity in the model parameters is explained by regressing the contributions on grouping variables or individual characteristics. This talk aims to illustrate how IPC regression can be used as a data-driven procedure to detect and provide estimates of individual or group differences in contemporary longitudinal structural equation models, focusing on autoregressive panel models in discrete and continuous time. It will be shown that equality and nonlinear parameter constraints, often encountered in longitudinal models, may bias the IPC regression estimates. A novel and robust estimation procedure for IPC regression and its software implementation will be presented.
Autoregressive models can be used to analyze longitudinal data. However, depending on the spacing of the discrete measurement occasions, autoregressive models will come to different results. In consequence, results from studies with different time intervals will differ. To overcome this and other shortcomings of discrete time autoregressive models, continuous time autoregressive models have been formulated. The basic idea of continuous time modeling is the assumption of an underlying continuous time process that can describe all associations of variables between any discrete points in time. This allows for using unequally spaced assessment designs to study psychological processes, and simplifies comparisons of results from studies with differing time intervals. Besides the repeated assessment over time, latent variable measurement models can be used at each measurement occasion to quantify (and increase) the reliability and precision of the measurement. Previous research has already exploited the idea of combining autoregressive models and latent variable models to a certain extent. However, continuous time models have rarely been combined with a general modeling framework that is suitable for incorporating a relatively large number of measurement models.

The main goal of this work is to combine continuous time and Generalized Linear (Mixed) Models to a new class of models termed "Generalized Continuous Time Models". We provide and describe the general equations needed and give some concrete examples. We chose one popular measurement model, the Rasch model, as an illustration of Generalized Continuous Time Models. For this model we conducted a "proof-of-concept" simulation study and give an illustrative real data example. A publicly accessible R package is available (ctsem) that can be used to model some Generalized Continuous Time Models.

HECHT, MARTIN ¹
Hardt, Katinka ¹
Driver, Charles C. ²
Voelkle, Manuel C. ¹

Generalized Continuous Time Models and the Continuous Time Rasch Model as an Example

¹ Humboldt-Universität zu Berlin, Germany
² Max Planck Institute for Human Development, Germany

martin.hecht@hu-berlin.de

Keywords:
Autoregressive Models, Classical Test Theory, Continuous Time Models, Generalized Linear Mixed Models, Generalized Linear Models, Item Response Theory, Longitudinal Data, Measurement
LÓPEZ-LÓPEZ, JOSÉ ¹
Pearson, Rebecca ²
Washbrook, Liz ³
Fazel, Mina ⁴
Tilling, Kate ¹

The Association Between Depression and Education in UK Adolescents: A Cross-Lagged Panel Analysis

¹ Population Health Sciences, Bristol Medical School, University of Bristol, United Kingdom
² Centre for Academic Mental Health, Population Health Sciences, Bristol Medical School, University of, United Kingdom
³ Centre for Multilevel Modelling, School of Education, University of Bristol, United Kingdom
⁴ Department of Psychiatry, University of Oxford, United Kingdom

ja.lopez-lopez@bristol.ac.uk

Keywords:

AIM: Mental health and educational achievement are both important in child development. There is previous evidence for links between them, however, the nature and direction of this relationship is unclear. In this paper, we explore the strength and directionality of the associations between both areas throughout adolescence. DATA: We used data from the Avon Longitudinal Study of Parents and Children, a birth cohort study including 15,390 children from the Bristol area who were born in the early 1990s. This resource includes four repeated measures of depressive symptoms (Short Mood and Feelings Questionnaire) and educational records (key stages 2-5), at mean ages 11, 14, 16 and 16 years. ANALYSIS: we fitted cross-lagged panel models within a structural equation modelling framework, using Mplus v8. We used full information maximum likelihood for missing data imputation in our main analyses. FINDINGS: results from the preliminary main analyses suggest possible age specific directionality of associations. Between ages 11 and 16, early measures of depression are negatively associated with later measures of education (standardised coefficient -0.034, 95
New technology facilitates the collection of time-intensive longitudinal data to study daily-life dynamics of psychological constructs (such as well-being) within persons over time (e.g., by means of Experience Sampling Methodology; ESM). However, the measurement quality can be affected by time- or situation-specific artefacts such as response styles or substantive changes in item interpretation. These distortions can be traced as changes in the measurement model (MM), which evaluates the constructs that are underlying a participant’s answers. If not captured, these changes might lead to invalid inferences about the targeted psychological constructs. Existing methodology can only test for a priori hypothesized changes in the MM. However, typically we have no prior information on (changes in) the MMs. Thus, an exploratory method that detects and models MM changes is needed before we can benefit from the full potential of ESM data. To this end, we present a method called latent Markov factor analysis (LMFA). In LMFA, a latent Markov chain captures the changes in MMs over time by clustering observations per subject into a few states and the data are factor analysed per state. The states indicate for which time points the construct measurements may be validly compared and within-subject MM differences can be explored by comparing the state-specific MMs. A simulation study shows a good performance of LMFA in recovering parameters under a wide range of conditions. The practical value of LMFA is illustrated with an empirical example.
SESSION Latent Variable Analysis

Chair: Jose-Luis Padilla

Friday, July 27, 11:30 - 12:50
Salon Schlegel

On the Effect of Observations and Parameters on the fit of SEM Models With Large Sample-Sizes
PIOTR TARKA

Evaluating Model Quality in Exploratory Bi-factor Modelling
EDUARDO GARCIA-GARZON, Francisco J. Abad, Luis E. Garrido

Analyzing Approximate Invariance From a Mixed-Method Ecological Approach to Validation
JOSE-LUIS PADILLA, Isabel Benitez, Luis Manuel Lozano, Pablo Doncel, Catalina Estrada, Carolina Barrios, Alejandra del Carmen Domínguez

A Correlated Covariate Amplifies the Bias of a Fallible Covariate in Causal Effect Estimates
MARIE-ANN SENGEWALD, Steffi Pohl
According to Tanaka (1987) there appears a general agreement that sample-size appropriateness should be tied to the ratio of the number of objects to the number of parameters estimated. In the works of Bentler (1980, 1983), Jöreskog (1978), Everitt (1984), James et al. (1982), and McDonald (1985) issue of sample size was invariably raised but typically not treated in sufficient detail. In fact, issue of how many observations and the number of parameters are needed before estimating and testing SEM has bothered researchers for years. Some of them have learnt lessons about the necessity of applying large samples. Marsh et al. (1998) explained that more observations (N) in sample size mean always better results for SEM models. Also Boomsma (1982, 1985) found that the percentage of proper solutions for SEM, including the accuracy of parameters, sampling variability in parameter estimates and the appropriateness of chi-square test statistic were all favorably influenced by having larger N's, recommending N ≥ 100, but also noting the desirability of N ≥ 200. Given a widely acknowledged consensus that ‘more means better’, in this study we ask the question: what precisely means a desirable level of sample size (in particular, exceeding the level of 200 observations, and reaching 300, 900, 1500, 2000, 3000, 5000, 10000, 19000, 28000, 60000 ob.) if we assume more or fewer parameters in SEM model? Besides, can we argue that larger or extreme samples will appear better than small and medium samples, as far as the quality of SEM models is concerned? Given this, we focus on the effects of observations and parameters on the fit of structural equation models (constructed on the basis of large sample-sizes). With this idea in mind, a series of calculations were conducted on one example of SEM specified model, while the data were derived from national survey taking place in Poland. The model included three varying levels of the number of parameters and ten levels of sample size.
Bi-factor models are usually applied to separate general and specific sources of variance. Bi-factor model quality measures include omega hierarchical, omega hierarchical subscale and the Explained Common Variance (ECV) index. The former represents superior choices to classical reliability statistics, while the latter evaluates the relative strength of a latent factor. They can be conjointly used to assess the extent that a general factor accounts for the common variance or if the unidimensional is supported. However, under realistic settings (e.g., cross-loadings are present) these indexes could be biased if obtained by either CFA or Schmid-Leiman solutions, as both methods lead to incorrect estimates.

Two promising algorithms for exploratory bi-factor modelling are: a) the iterative target rotation based on an initial Schmid-Leiman solution (SLi); b) the bi-factor target rotation based on obliquely rotated solution (biFAD; Waller, 2017). As they differ on how target matrix is defined (empirically vs fixed cut-off point; totally vs partially specified) and applied (iteratively vs non-iteratively), they are expected to have a dissimilar impact on quality indices.

A Monte Carlo simulation was conducted by manipulating sample size, number of specific factors, number of indicators, factor loading range, factor loading average and cross-loading size. BiFAD fixed cut-off points ranged from .05 to .20. Both methods showed good recovery, but antagonistic behaviors regarding quality indexes: SLiD overestimated omega hierarchical and ECV for the general factor (due to overestimation of the mentioned factor), while underestimating omega hierarchical subscale and ECV for the specific factors (due to underestimation of such factors). BiFAD results were in the contrary direction. SLiD was prone to produce factor collapse, while biFAD was impacted by cut-off point selection and the number of factors involved. Guidelines for interpreting bi-factor quality indexes are provided.
To uncover cause of measurement invariance is critical to improve validity of cross-cultural comparisons from an ecological approach to validation (Zumbo, 2017). New developments in quantitative methods like "alignment" method can help researchers in understanding the absence of invariance. "Alignment" results can make it easier to link quantitative results with qualitative findings within a mixed methods research framework. The Indigenous Social Desirability Scale (ISDS) (Domingez & van de Vijver, 2014), was developed from an emic perspective for Mexican culture. The aim of this study is to explore how to integrate quantitative evidence of approximate measurement invariance obtained by "alignment" with qualitative evidence of cognitive probes for a multi-national research project with country samples from Mexico, Colombia and Spain. 967 participants responded to the ISD scale: Mexican (257), Spanish (513), and Colombian (197). The ISD scale consists of 14 polytomous items to capture the positive and negative dimensions of social desirability. A robust maximum likelihood fixed alignment analysis was conducted to test approximate measurement invariance. At the same time, a literature review was conducted to propose a mixed research design to integrate alignment results with cognitive probe findings. 4 intercepts and 1 loading for the Spanish sample, and 1 intercept for the Mexican sample, out of the 14 items are not invariant across the country groups. The percentage of non-invariant parameters support quality of alignment results. At the same time, a mixed research design including cognitive probe formats was developed for off and online cognitive interviewing. Alignment results support approximate metric invariance across ISDS country samples. In addition, we discuss benefits of the mixed methods research to investigate into possible causes of the absence of invariance for cultural-bound constructs like social desirability from an ecological view of validity.
Covariate adjusted treatment effects are commonly estimated in non-randomized studies, either with ANCOVA or Propensity score methods. It has been shown that measurement error in a covariate can bias treatment effect estimates, if it is not appropriately accounted for. So far, delineations on the bias of a fallible covariate primarily assumed a true data generating model that included just the respective latent covariate. It is, however, more plausible that the true model consists of more than one covariate. By intuition, an additional covariate that is correlated with the latent covariate can be helpful, if only a fallible measure of the latent covariate is available for adjustment - the correlation might partly compensate for the bias due to measurement error. We disentangle when it is advisable to include a correlated covariate for adjustment. For this aim, we analytically investigate a true model with two covariates that are correlated and derive the bias when only a fallible measure of one of the covariates is available for adjustment. With the fallible covariate, it is not always advisable to include the additional covariate in the adjustment model, as it can extensively increase the bias, even if it is highly correlated to the latent covariate. We point out the distorting effects of fallible covariates and discuss adjustment for latent covariates as a possible solution.
SESSION Multilevel Modeling

Chair: Wouter Talloen

Friday, July 27, 11:30 - 12:50
Salon Hölderlin

**Measurement Error and Unmeasured Confounding in Multilevel Mediation Models**
WOUTER TALLOEN, Tom Loeys, Beatrijs Moerkerke

**Same Same but Different?! Measuring of Local Sex Ratios**
ANDREAS FILSER, Richard Preetz

**Assessing Structures of Prejudice in Europe with Multilevel Latent Class Analysis**
ALICE BARTH

**Detecting Selection Bias in Meta-Analyses with Dependent Effect Sizes: A Simulation Study**
BELÉN FERNÁNDEZ-CASTILLA, Wim Van Den Noortgate
Multilevel 2–1–1 mediation models are frequently used in educational research when intervention is measured at cluster (e.g. class) level and mediator and outcome at unit (e.g. student) level. In such settings, different indirect effects may be of interest. The within indirect effect measures the indirect effect through the mediator at the unit-level while the between indirect effect measures this effect at cluster-level. The latter can be estimated by adding an aggregated unit-level mediator as a predictor for the outcome. Manifest approaches assume that observed group means are perfectly reliable measurements of this aggregated unit-level mediator while latent approaches such as multilevel structural equation modeling take into account measurement error. In a first step, we study the impact of measurement error on estimators for between- and within indirect effects obtained by manifest and latent approaches. In a second step, we focus on the impact of unit- and cluster-level unmeasured mediator-outcome confounding on causal effects. In a third step, we study the combined effect of measurement error and unmeasured confounding. A simulation study is conducted to compare bias, precision, coverage rate and bias-variance trade-off for each causal effect. Estimators for the within indirect effect obtained by manifest and latent approaches perform similarly, only unmeasured confounding at unit-level induces bias for the within indirect effect. The estimator for the between indirect effect in both approaches is affected by unmeasured confounding at cluster-level. For manifest approaches, bias for the between effect estimator also depends on the strength of the intraclass correlation for the mediator and unit-level unmeasured confounding.
Imbalanced numbers of men and women in societies or social groups (i.e. sex ratios) have been linked to a variety of social consequences. Studies report associations with relationship formation patterns and timing, divorce rates, fertility timing and rates, sexual norms, female labour market participation as well as violence and aggression. Theoretical arguments commonly start from a social exchange perspective, considering imbalanced sex ratios as a factor that shapes individuals’ dyadic power on the partner market and within relationships. However, theoretical reasoning remains unclear on whether behavioural consequences result from individuals’ deliberate adjustment of partner market strategies or unconscious endocrinal or normative variations. Previous studies use population register data to refer local sex ratios on county or municipality-level for a wide range of age, mostly from 16 up to 40 years old adults. These studies are based on the implicitly assumption that individuals reflect those imbalances deliberate, resulting in specific daily life actions. But, there is no study which tests this assumption empirically. Based on combined data from a representative German longitudinal survey (pairfam) and population register, we analyse how local sex ratios, measured by register data, are perceived conscious by individuals. Our empirical approach contains two parts: First, we test correlations between local sex ratios and perceived subjective sex ratios. Second, we analyse the transition from singlehood to partnership as an example for the consequences of imbalanced sex ratios and the (competing) influence of local and perceived sex ratio. Our results suggest that correlations between local sex ratios from register data and perceived sex ratios by individuals are very low. Our longitudinal example also indicates that local and perceived sex ratios are independent while subjective perceived sex ratios seems a better indicator to explain partner market outcomes.

FILSER, ANDREAS 1
Preetz, Richard 1

Same Same but Different?!
Measuring of Local Sex Ratios

1 University of Oldenburg, Germany

andreas.filser@uol.de

Keywords:
Measurement, Multilevel Modeling, Partner Market, Sex Ratio
Analyses of prejudice against out-groups often focus on a specific group, e.g. Muslims, immigrants, or homosexuals. However, Gordon Allport already argued in the 1950s that “people who reject one out-group will tend to reject other out-groups” (1954, p. 68), and many studies have since shown that different types of prejudice are interrelated (e.g., Zick et al. 2008; Reeskens 2013). In this presentation, we propose to assess structures of prejudice in Europe by means of multilevel latent-class analysis (ML-LCA). For this purpose, we use a variant of the Bogardus Social Distance scale that was part of the European Values Study (EVS) in 2008: Survey respondents were asked to sort out any groups (e.g., drug addicts, right wing extremists, Jews, immigrants…) they would not like to have as neighbours. Applying ML-LCA to this question, we identify both an individual-level typology of perceived out-groups and segments of countries with different patterns of prejudice. On the individual level, five latent classes are distinguished that differ with respect to the characteristics of out-groups (cultural minorities, people with deviant behaviour, and political extremists) and the amount of resentment towards groups-as-neighbours in general. The latter is also a main differentiating factor between countries. In a second step, predictors for class membership on the individual and country level are investigated. Substantial and methodological implications of the ML-LC approach towards prejudice structures will be discussed.
In meta-analysis, it is common to find primary studies that include multiple effect sizes, generating dependence among them. Although there are several techniques available for dealing with dependent effect sizes, the assessment of selection bias in this context (i.e., publication bias and selective outcome reporting bias) has not yet been thoroughly scrutinized. Therefore, the aim of this study is to explore, by means of a simulation study, the performance of commonly used methods for detecting publication bias in situations where primary studies include multiple effect sizes. To that end, meta-analytic datasets were generated under a variety of realistic conditions. Next, three different types of bias were induced: publication bias, selective outcome reporting bias, and the combination of both. Datasets unaffected by any type of selection bias were also considered. Afterwards, six different methods for detecting publication bias were applied: Begg’s Rank Correlation test (using variance and sample size), the Trim and Fill method (R0 and L0 estimator), Egger’s Regression and Funnel Plot test. These two last methods were adapted by using multilevel three-level models to account for within-study dependency. These methods were evaluated in terms of Type I error and power. Results indicated that Begg’s Rank Correlation test (using both variance and sample size), the Trim and Fill (L0) method and Egger’s Regression test lead to Type I error rates that are too high in most conditions, whereas the Funnel Plot test and the Trim and Fill (R0) method lack power. Also, results showed that concluding that there is selection bias when four out of the six methods indicate the presence of selection bias, leads to controlled Type I errors across conditions. However, this approach is still unsatisfactory in terms of power. We conclude that the studied approaches have serious flaws, and that other approaches should be explored.
SESSION Bayesian Statistics

Chair: David Kaplan

Friday, July 27, 11:30 - 12:30
Salon Novalis

An Approach to Addressing Multiple Imputation Model Uncertainty Using Bayesian Model Averaging
DAVID KAPLAN, Sinan Yavuz

Bayesian Meta-Analysis of Studies Using Cohen's d in R
SHERI KIM

Hypothesis-Testing Demands Trustworthy Data – A Simulation Approach to Inferential Statistics Based on the Research Program Strategy
FRANK ZENKER, Antonia Krefeld-Schwalb, Erich H. Witte
An Approach to Addressing Multiple Imputation Model Uncertainty Using Bayesian Model Averaging

1 University of Wisconsin - Madison, USA

david.kaplan@wisc.edu

This paper considers the problem of imputation model uncertainty in the context of missing data problems. We argue that so-called "Bayesianly proper" approaches to multiple imputation, although correctly accounting for uncertainty in imputation model parameters, ignore the uncertainty in the imputation model itself. We address imputation model uncertainty by implementing Bayesian model averaging as part of the imputation process. Bayesian model averaging accounts for both model and parameter uncertainty, and thus we argue is fully Bayesianly proper, in the sense of Schafer (1997). We apply Bayesian model averaging to multiple imputation under the fully conditional specification approach. An extensive simulation study motivated by real data considerations is conducted comparing our Bayesian model averaging approach against choosing the imputation model with the highest posterior model probability, and against normal theory-based Bayesian imputation not accounting for model uncertainty. The results reveal a small but consistent advantage to our Bayesian model averaging approach under MCAR and MAR in terms of Kullback-Liebler divergence. No procedure works well under NMAR. A small case study is also presented. Directions for future research are discussed.
Bayesian meta-analysis has several key advantages over frequentist meta-analysis. First, a Bayesian framework theoretically utilises the correct conditional probability, and practically allows evidence for the null hypothesis. Second, the posterior distribution and credible intervals are intuitively interpretable. Third, data can be added as new participants or studies appear, which is particularly important in living meta-analyses (Elliott, et al., 2017; Simmonds, Salanti, McKenzie & Elliott, 2017). There already exist examples of how to apply a Bayesian framework to meta-analysis (e.g., Scheibehenne, Jamil & Wagenmakers, 2016; Smith, Spiegelhalter & Thomas, 1995; Sutton & Abrams, 2001). However, these examples only utilise Bayes factors and odds ratios or risk differences. There are no studies that demonstrate how to apply effect size meta-analysis on commonly used effect sizes in psychology such as Cohen's d. This paper proposes a Bayesian fixed-effects meta-analysis of studies that use Cohen's d. The meta-analysis results in an overall effect size (an estimate of the population effect size) and its credible interval. The analysis is conducted using Stan in R. It will also provide practical guidelines about how to interpret the results of the Bayesian meta-analysis.
In psychology as elsewhere, the main strategy to obtain empirical effects remains null-hypothesis significance testing (NHST). However, recent attempts have failed to replicate “established” effects that allegedly were well supported. Hence, NHST retains too many errors. For otherwise far more such effects should have been successfully replicated. This makes trusting even results that top-journals publish a difficulty.

We advocate the research program strategy (RPS) as superior to NHST. Employing both Frequentist and Bayesian tools, we show by means of data-simulation that RPS’s six steps—leading from making a discovery against a random model, to statistically verifying a hypothesis—retain far fewer errors than a standard usage of NHST. Therefore, RPS-results deserve far greater trust than NHST-results. Our simulations moreover estimate the expectable proportion of errors among published results.

Where test-power is unknown, NHST constitutes the RPS’s first step, where probabilities serve to discover an effect preliminarily. By contrast, if we know test-power, then a substantial discovery may arise (step 2). Moving beyond discoveries, steps 3 to 6 concern the justification of hypotheses (falsification and verification). These steps presuppose the use of likelihoods, and demand data of high induction quality (test-power) for such data to test hypotheses. We employ Wald’s criterion (the ratio of test-power and significance level) to preliminarily or substantially falsify the H0 (steps 3, 4), and to preliminary verify the H1 (step 5). Finally, if the ratio of the likelihoods for the H1 and the H0 exceeds Wald’s criterion, while the maximum-likelihood-estimate of data lies close to the H1, then this substantially verifies the H1 (step 6).
SESSION Item Response Theory

Chair: Timo Bechger

Friday, July 27, 14:00 - 15:20
Saal Friedrich Schiller

DIF Methods in Dexter
IVAILO PARTCHEV, Timo Bechger

The Great Dexperiment: Psychometrics With Observed Variables
TIMO BECHGER

Bayesian Estimation of Item Response Models to Account for Learning During the Test
JOSÉ H. LOZANO, Javier Revuelta

A Probabilistic IRT Model for the Joint Assessment of Objects and Persons in Fully Crossed Designs
GEORG HOSOYA
Arguably, DIF is a shrinking problem as item writers strive towards, and get better at, producing DIF-neutral items. Notwithstanding, methods for detecting DIF are proliferating. We discuss two methods included in the R package, dexter (Partchev, Bechger, Maris & Koops, 2017): one exploratory, based on the relative difficulties of pairs of items (Bechger & Maris, 2015); and one confirmatory, related to latent profile analysis (Verhelst, 2012).
The Dexter package is based on the principle that psychometric models are vehicles to answer questions about observed variables: Users are not supposed to see parameters or (estimates of) latent abilities. My intention is to discuss how this principle is implemented illustrated with a number of real live examples.

BECHGER, TIMO ¹

The Great Dexperiment: Psychometrics With Observed Variables

¹ Cito, The Netherlands
timo.bechger@cito.nl

Keywords:
Item Response Theory
In the present work, several explanatory item response models are proposed to account for the learning that takes place during the execution of a test due to the repeated use of the operations involved in the items. The models include a difficulty component derived from the cognitive operations involved in solving the item, as well as a learning component derived from the use of said operations in previously answered items. Six different models are proposed taking into account the type of response in which the model establishes that learning occurs (i.e., in correct responses only, in correct responses and errors indistinctly, or in correct responses and errors in different measure) and whether or not the model considers the existence of individual differences in learning. Based on the above, a simulation study was conducted in order to test whether Bayesian goodness-of-fit procedures allow identifying the model used to simulate the data. The data were generated from the six proposed models plus the LLTM and the Rasch model. Additionally, three different sample sizes were used (i.e., N = 250, N = 500, and N = 1000). Thus, the combination of models and sample sizes resulted in an 8 × 3 factor design. One hundred data sets were simulated for each of the 24 design points. For each data set, 20 dichotomous responses were simulated based on a weight matrix of five components. The eight models were estimated from each simulated data set using Bayesian inference. Specifically, parameters were estimated via Markov Chain Monte Carlo (MCMC) using the Stan computer language. The fit of the models to the data was assessed with three deviance measures based on the information theory: the deviance information criterion (DIC), the widely applicable information criterion (WAIC), and the leave-one-out cross-validation (LOO). As expected, the results indicated that the model used to generate the data for each design point minimized the discrepancy statistics. The results therefore support the ability of the proposed models to detect learning effects during the test.
In psychological aesthetics, it is of interest to assess an artwork’s (emotional) impact on the perceiver. From the perspective of differential psychology, it is of interest to assess interindividual differences of perceivers in displaying or reporting a certain response. The methodological question that arises in this context is: how to assess perceivers and artworks simultaneously with regard to a psychological construct? The answer is relatively simple: perceivers are presented with artworks in a fully crossed design and the task is to judge the objects on, for instance, the items of the AESTHEMOS questionnaire. In a next step, the manifest responses are projected onto latent dimensions using a Bayesian probabilistic IRT model which allows for disentangling the effects of the artworks as well as the effects of the perceivers. While models of this type are certainly not uncommon, the proposed model differs from known approaches as it is an extension of Master’s partial credit model migrated to a Bayesian MCMC setting. Using the Bayesian approach allows readily for the identification of posterior distributions of the perceivers’ individual characteristics. In addition, it is possible to evaluate the artworks’ overall tendencies to evoke certain responses. In this talk, the theoretical and methodological underpinnings of the proposed model are discussed. Results from a real world application of the model to the task of judging paintings are presented. In addition, it is highlighted that the model’s application is not limited to artworks but also persons could be subjects of the judging process.

HOSOYA, GEORG

A Probabilistic IRT Model for the Joint Assessment of Objects and Persons in Fully Crossed Designs

1 Freie Universität Berlin, Germany

georg.hosoya@fu-berlin.de

Keywords:
Aesthetics, Bayesian Statistics, Categorical Data, Fully Crossed Designs, Item Response Theory
SYMPOSIUM Challenges in Interdisciplinary Research Methodology: The Study of Complex Systems

Chair: Hilde Tobi

Friday, July 27, 14:00 - 15:05
Salon Schlegel

Using Simulation Models to Measure Resilience
GUUS TEN BROEKE

A Modest Step Toward Bringing Unity in Interdisciplinary Research
JARL KENNARD KAMPEN

Innovation Modelling in Engineering and Scholastic Philosophy
JULIA P. A. VON THIENEN, Chiara Paladini, Christoph Meinel

Mapping Validity in Modelling for Interdisciplinary Research
HILDE TOBI, Guus ten Broeke
Using Simulation Models to Measure Resilience

Many social-ecological systems (SES), such as fisheries, land-use systems, and agricultural systems are under pressure from human activities and environmental changes. Thus, it is important to study their resilience against such pressures. Resilience may be generated by various mechanisms, such as stabilising feedbacks, spatial interactions, diversity of underlying units (like agents), and mechanisms for adaptation. Simulation models are an important tool for assessment of resilience. Simulation models help us to test the effects of various assumptions on interactions and feedbacks within the system on its resilience. But, not all types of simulation models may be equally suitable for this purpose. We compare two commonly used model types for describing SES, namely ordinary differential equation (ODE) models and agent-based models (ABMs). As test-case, we consider a system in which consumers compete for a renewable common-pool resource. The system is modelled both as an ABM and as an ODE model. The ABM is spatially explicit and dynamic with respect to time. Agents can move in search of resource, and can harvest from their present location. The ODE model is dynamic, but non-spatial. We examine to what extent the ODE model can be fitted to the ABM. We investigate how both models respond to external shocks, and apply resilience measures such as return time to quantify resilience. The results show that the ODE model can reproduce the behaviour of the ABM, if some mechanisms relevant for resilience are excluded. Specifically, the ODE model does not capture effects of agent adaptation, or local differences in space. We conclude that the most suitable modelling approach depends on the system. If resilience is caused by system-level feedbacks, then ODE models may be suitable for assessing this resilience. If, in contrast, agent adaptation or localised actions of agents contribute to resilience, then ABMs may be more suitable.
Research communication in interdisciplinary research projects requires a way of demarcation of theory and knowledge that is easy to communicate, inconsequential for the framework of concepts, results, and procedures within existing scientific disciplines, and abstains from trying to resolve the dispute of (neo)positivist and constructivists. In this essay, a simple way of demarcation is proposed that only secures the vocabulary needed to comprehensively discuss research methodology and findings in interdisciplinary research contexts. It starts from the notion of language independent, and language dependent, reality. Language, when it’s not sheer fantasy, is at best after-fact. All language is instruction tied to specific senses or acts which definitions have to be known for the instruction to be possible. A possible instruction can be carried out, an impossible instruction cannot be carried out (logically or empirically, temporarily or permanently). Knowledge is to know which instructions are predictive of a demonstrable result, state or situation in language independent reality. Knowledge decreases outcome space by pointing out possibilities and impossibilities. Any theory that contains one or more impossible instructions is not knowledge. Any theory that doesn’t reduce outcome space (typically, by not pointing out impossibilities) is not knowledge. Any theory with wrong predictions is not knowledge. Any theory falling short of a demonstration is not knowledge.
A phenomenon that affects all domains of human affair is that of innovation. In the course of innovation, someone makes a new contribution in a subject domain, causing more or less perturbation in the field. Today, when a novel car-model appears on the market, the customers obtain a new choice (additive effect), but the overall transportation infrastructure remains mostly unchanged. By contrast, when cars made their first appearance on the market, transportation by horse was gradually replaced and respective industries, offering horse food, equipment, medicine etc., had to give way (subtractive effect). Innovation domains are complex adaptive systems. Stakeholders in the domain can embrace or mobilize resistance against innovation depending on the effects they anticipate. The modelling of innovation patterns is an interdisciplinary challenge par excellence. A basic model is sought that should be applicable to all domains of human affair, while parameters in the model might change depending on the field of application. To develop and test a basic model, as well as to identify domain-specific parameters, inter- and cross-disciplinary studies are required. We apply an innovation model recently developed in the field of engineering to Scholastic Philosophy. This application has three important advantages: (i) Scholastic Philosophy is clearly different from engineering. (ii) High quality data is available for long time spans; we analyse changes in philosophical theory from Augustinus to Aureoli, thus covering almost 1000 years of theorising. (iii) Scholastic authors are highly explicit as to which claims of their predecessors they endorse or reject, and which novel claims they propose, so that additive versus subtractive effects can be easily identified. In sum, the basic model that is tested fits the data well. However, modelling challenges are also identified, especially pertaining to the operationalization of community resistance against innovation.
Computer simulations are a promising methodology in the interdisciplinary study of complex systems, such as socio-ecological systems and socio-technical systems. One quality criterion of all empirical research, regardless of its inter-disciplinarily, is validity. Actually, validity is not one criterion as different kinds of validity are usually distinguished (e.g. content validity, external validity). Validity is also used as a quality criterion in the context of simulation modelling. Here, validity pertains to different aspects of the model built. To understand and assess the quality of interdisciplinary research in which models are designed and used, a thorough understanding of the different meanings of ‘validity’ is needed. In this paper, we first review concepts of validity and validation of models of complex systems. Then we review validities and validation procedures in interdisciplinary research with an emphasis on research involving both the social sciences and the natural sciences. Looking at both the purpose of the model and the input of empirical sciences, these two strands (models of complex systems, and interdisciplinary empirical research) are synthesized into one map of different validities and validations in modelling for interdisciplinary research. With this map, we propose unambiguous terminology for validity assessment in modelling for interdisciplinary research.

TOBI, HILDE

ten Broeke, Guus

Mapping Validity in Modelling for Interdisciplinary Research

1 Wageningen University & Research,
The Netherlands

hilde.tobi@wur.nl

Keywords:
Dynamic Modeling, Interdisciplinary Research Methodology
SESSION Latent Class Models

Chair: Ana Gomes

Friday, July 27, 14:00 - 14:40
Salon Hölderlin

Internet Use in the European Union: A Multilevel Latent Class Analysis
ANA GOMES, José G. Dias

Using Latent Variable Models to Evaluate Test Quality Criteria of Tests Measuring Nominal Constructs
HENDRYK BÖHME
Multilevel data structures are quite common in the social and human behavior sciences and new analytical techniques have to be applied to these specific data sets. In this particular case, the Multilevel Latent Class Model (MLCM) becomes a viable alternative to the conventional Latent Class Model (LCM). The Multilevel Latent Class Model (MLCM) considers not only the individual level 1 (Level 1), but also an upper level (Level 2) that defines a nesting or hierarchical structure (Kimberly & Muthén, 2010). The MLCM decomposes the existing heterogeneity between countries and within countries (individuals), resulting into homogeneous segments of countries and individuals. The data set comes from the Eurobarometer (TNS Opinion & Social, 2013) and contains information on the 28 countries of the European Union (n = 26680 citizens). The average age of the respondents is 46.82 years (s.d. = 1.9) and varies between 15 and 98 years. At the individual level (Level 1), three variables were used to identify individual segments in Europe, taking their Internet usage pattern into account: frequency of access to the Internet, means of access, and online activities. Six sociodemographic variables were introduced to characterize the latent classes, namely: gender, age, literacy, marital status, occupation, and type of community. At the second level of analysis (Level 2), countries were introduced as contextual predictors, allowing the grouping of individuals into segments based on the similarities found.

In psychological testing, measured constructs are often assumed as being continuous latent variables, and thus are modeled accordingly. In the case of continuous latent variables, there is a variety of methods for evaluating test quality criteria. However, many of these methods are not applicable if the latent variable is constructed as being nominal. By using the example of the nominal construct sophistication of conditional reasoning (reasoning with if-then-propositions), analysis procedures are presented that allow an evaluation of important test quality criteria like reliability and validity. These analysis procedures (e.g. latent class analysis, latent transition analysis, multinomial logistic regression analysis with latent variables) provide adequate information and well interpretable results when it comes to developing and evaluating psychological tests measuring nominal constructs.

BÖHME, HENDRYK

Using Latent Variable Models to Evaluate Test Quality Criteria of Tests Measuring Nominal Constructs

1 Department of Psychological Research and Development, Federal Employment Agency, Germany

hendryk.boehme@arbeitsagentur.de

Keywords:
SESSION Applied Statistics

Chair: Alrik Thiem

Friday, July 27, 14:00 - 15:00
Salon Novalis

**Interpretation of Main Effects for Moderated Regression Models**
JULIE LORAH

**Topic Modeling As a Type-Forming Process of Social-Ecological Education Research**
THOMAS PRESCHER

**Small Act, Huge Effect: Algorithmic Sources of Publication Bias in Political Science Research**
ALRIK THIEM, Tim Haesebrouck
Moderated regression models include an interaction, or product term and can be used to assess whether the relationship between a given independent variable (IV) and dependent variable (DV) depends on a third moderator variable (MV). Literature exists regarding interpretation of a significant moderator effect as well as guidance for interpreting the main effects in the presence of a significant moderator effect. Typically, researchers recommend either ignoring main effect completely, or carefully interpreting them as conditional effects. However, when the interaction effect is not significant, recommendations indicate that the typical interpretation of main effects as average effects is appropriate. The present study challenges this claim since lack of significance may be due to lack of power rather than no true population effect. To explore this idea, a simulation study is conducted. A moderated regression model with one predictor, \( Y = a + b_1X + b_2M + b_3X*M + e \), is estimated based on simulated data varying the sample size, effect size of the interaction term, and the centering of X. Preliminary results indicate that for typical data sets this model may be underpowered to detect moderation effects, which is consistent with the literature. Examining the distribution of slope coefficient estimates for the main effects (\( b_1 \) and \( b_2 \)) indicates that even for models with no significant interaction effect, interpreting these main effects as average effects could be very misleading. Recommendations for applied researchers include using model selection procedures that can provide evidence for either the more or the less complex model, such as Bayesian Information Criterion (BIC); routinely mean-centering predictors to guard against particularly misleading main effect interpretation; and conducting post hoc power analyses for non-significant interaction effects before proceeding with interpretation of main effects.
Title: Topic modeling as a type-forming process of social-ecological education research: Program analysis on education for sustainable development in and by companies

The latent semantic analysis of qualitative data is presented on the basis of a programme analysis on education for sustainable development in and by companies. It deals with the question of how a topic modeling can be applied to a qualitative data set in the context of a classical program analysis. And the question is being pursued, what is the possibility of validating the result? To this end, the concrete evaluation and analysis steps of computer-aided analysis are traced in detail with the "MAchine Learning for LanguagE Toolkit" (MALLET) based on the algorithm of the "latent dirichelet allocation" (LDA). As a result, a Topic Modeling with 10 topics will be presented and interpreted as a central framework for a sustainability-oriented learning culture with regard to ecological education as a problem solving and life-world-oriented cognitive process in the context of a required legitimization management.
Meta-analyses in political science continue to demonstrate the pervasiveness of publication bias, the reasons for which are said to lie with authors, reviewers, journal editors and project sponsors. In this article, we reveal an as yet undiscovered source of publication bias. More specifically, we demonstrate why the uncritical import of the Quine-McCluskey algorithm (QMC) from electrical engineering into social-scientific data analysis with Qualitative Comparative Analysis (QCA) in the late 1980s had to lead to considerable publication bias. Drawing on complete replication material for 160 studies from political science that have employed QCA, we also measure the extent this problem has assumed in empirical research. Last but not least, we present a solution that is guaranteed to eliminate this source of bias. It consists in a redefinition of the objective function under which optimization algorithms such as QMC operate in QCA. Besides contributing to the scientific study of publication bias, our article thus also underlines the importance of evaluating the adequacy of foreign methods before putting them to uses which they were not originally designed for.
STATE-OF-THE-ART Systematic observation of human behavior from a methodological perspective (M. Teresa Anguera)

Chair: Milica Miočević
Friday, July 27, 15:30 - 16:00
Saal Friedrich Schiller

Although once considered by some researchers to be on the fringe of conventional methodology, systematic observation has been progressively incorporated into diverse areas of psychological research. Psychological science is increasingly focusing on the study of everyday behavior, and studies applying systematic observation methodologies can now be found in mainstream interdisciplinary psychology journals (e.g., Frontiers in Psychology and Psicothema) and methodology journals (e.g., Behavior Research Methods and Quality & Quantity).

In this state-of-the-art lecture, I will discuss core aspects of systematic observation as a scientific method, with a focus on the profile of this approach and the specific processes it involves. Observational methodology is characterized by high scientific rigor and flexibility throughout its different stages and allows the objective study of spontaneous behavior in natural settings.

The study of spontaneous behavior is characterized by a richness of information that can only be captured by video or sound recordings, without elicitation. Furthermore, the tools now available to explore this richness, often hidden within the deeper layers of the data, have been greatly enhanced by recent technological advances. Quantification in observational methodology is particularly robust and observational studies applying this methodology deserve consideration as mixed methods research.

One particularly interesting area is the use of indirect observation of everyday behavior in natural settings based on textual material, such as conversations, blog posts, tweets, etc. This approach involves 'liquefying' original or transcribed texts into a format in which the original qualitative data can be quantified and analyzed using techniques based on the order or sequence of events rather than on traditional frequency counts.

I will conclude this lecture by providing an overview of the broad range of applications of scientific systematic observation and highlight the possibilities of systematically studying spontaneous behavior in natural settings.
Randomized experiments provide the strongest warrant for causal inference. However, randomized experiments assume full treatment adherence for a proper estimate of the causal effect. One form of treatment non-adherence is binary in which participants in the treatment group do not accept treatment. Biases resulting from binary treatment non-adherence using several traditional analysis approaches (e.g., per protocol analysis) are briefly illustrated. Newer approaches that compare the effect for those participants in the treatment group who received treatment with those participants in the control group who would have accepted treatment if offered are presented (average treatment effect on the treated). Another form of non-adherence is partial adherence in which participants in the treatment group receive only a proportion of the intervention (e.g., 5 sessions of a 10-session intervention program). Confounder adjustment and instrumental variable approaches that provide estimates of the treatment effect conditional on the proportion of the full treatment that was received are presented. We discuss the assumptions of these approaches and the conditions under which they may be most useful. The usefulness of binary and partial adherence approaches as supplements to intention to treat analyses as an estimate of the causal effect will be discussed.
### Index

Abad, Francisco J., 231
Agache, Alexandru, 46
Akuro, Anta, 142
Aliri, Jone, 76
Ana, Hernández, 160
Anderson, Mark, 44
Anguera, M. Teresa, 69, 72
Anguera, María Teresa, 73
Angulo-Brunet, Ariadna, 49, 63
Appelbaum, Sebastian, 148
Arasil, Patricia Flor, 203
Arce, Constantino, 55
Aritzeta, Aitor, 76
Ark, Andries van der, 131
Arnold, Manuel, 223, 224
Asún, Rodrigo, 68
Auerswald, Max, 154
Bachbauer, Nadine, 53
Baghaei, Purya, 97
Balluerka, Nekane, 76
Banaszak, Slawomir, 51
Barendse, Mariska, 127
Barrada, Juan Ramón, 99
Barrios, Carolina, 232
Barth, Alice, 238
Bauer, Johannes, 147
Bechger, Timo, 245–247
Becker, Benjamin, 177
Bengs, Daniel, 201
Benitez, Isabel, 232
Berchtold, André, 218
Beretvas, S. Natasha, 184
Bilhler, Lilly, 46
Blasius, Jörg, 105, 106
Bliem, Harald R., 43
Blázquez, Desirée, 54, 198
Bonnard, Claire, 161
Born, Sebastian, 147, 199, 200
Borrucco, Marta, 63
Botella, Juan, 54, 198
Braeken, Johan, 200
Brandt, Holger, 44, 78, 94
Brauer, Kay, 83, 86
Brefeld, Ulf, 201
Brunner, Martin, 142
Buchholz, Janine, 179
Böhme, Hendryk, 259
Bürkner, Paul, 165
Camerino, Oleguer, 69
Castañer, Marta, 69
Chacón-Moscoso, Salvador, 72, 73
Charpentier, Emanuel, 165
Chen, Hui-Fang, 197
Coballes del Río, Soraya, 70
Cruz, Jaume, 49
Cuevas, Ana María, 57
Damian, Elena, 107
de Francisco, Cristina, 55
De la Fuente-Sánchez, Leticia, 79, 84
de la Torre, Jimmy, 99
De Neve, Jan, 196
De Roover, Kim, 227
De, Tamal, 82
De, Tamal Kumar, 71, 217
Debeer, Dries, 177
Debelak, Rudolf, 145, 149
Dehaene, Heidelinde, 195, 196
del Río, F. Javier, 56–58
Devlieger, Ines, 128
Domínguez, Alejandra del Carmen, 232
Doncel, Pablo, 232
Driver, Charles C., 225
Eid, Michael, 114, 216
Elipe Miravet, Marcel, 182, 205
Engelen, Ron, 96
Erosheva, Elena, 102
Estrada, Catalina, 232
Fan, Qianqian, 47, 74
Fazel, Mina, 226
Felderer, Barbara, 59
Fernandez-Alonso, Ruben, 143
Fernández García, María Paula, 70, 75, 77, 81
Fernández-Castilla, Belén, 239
Ferron, John M., 184
Filser, Andreas, 237
Fink, Aron, 147
Flor Arasil, Patricia, 182, 205
Frey, Andreas, 147, 212
G. Dias, José, 258
Gantner, Julia, 92
Garcia-Garzon, Eduardo, 231
García Sedeno, Manuel Antonio, 56–58
Garcia-Garcia, Juan, 79, 84
Garrido, Luis E., 231
Giret, Jean-Francois, 161
Gische, Christian, 91, 93
Goldhaber, Dan, 102
Goldhammer, Frank, 179
Gomes, Ana, 257, 258

Gonzalez, Trinidad, 171
González González-Mesa, Carmen, 70
González-Romá, Vicente, 116
Gorostiaga, Arantxa, 76
Grimm, Kevin, 117
Grund, Simon, 144
Gu, Xin, 164
Guo, Jiin-huarng, 80
Haesebrouck, Tim, 264
Hahn, Christiane, 86
Hardt, Katinka, 61, 225
Hartig, Johannes, 119–121, 213
Hecht, Martin, 61, 225
Hedges, Larry, 89
Helm, Jonathan, 166, 189, 190
Hernandez, Ana, 111
Hernández, Ana, 116
Herrero Machancoses, Francisco, 182, 205
Ho, Moon-Ho, 47, 74
Hochleitner, Margarethe, 43
Hochweber, Jan, 213
Hoijtink, Herbert, 163, 164
Holling, Heinz, 89
Hoppe, Kevin, 62
Hosoya, Georg, 249
Jaenicke, Johannes, 103
Jamshidi, Laleh, 184
Jara Jiménez, Pilar, 182, 205
Jin, Kuan-Yu, 197
Juola, James F., 198
Kampen, Jarl Kennard, 253
Kamphuis, Frans, 95, 96
Kaplan, David, 241, 242
Karpe, Karolina, 45
Kelava, Augustin, 114
Khorramdel, Laleh, 178
Kiefer, Christoph, 170
Kim, Sheri, 243
Klaassen, Fayette, 219, 220
Klopp, Eric, 192, 193
Klößner, Stefan, 192, 193
Koch, Tobias, 113, 114
Komlenac, Nikola, 43
Koopman, Letty, 133
Kossi, Yann, 161
Kraft, Volker, 181, 183
Krefeld-Schwalb, Antonia, 244
Krieger, Ulrich, 59
Krishna, Anand, 108
Kroehne, Ulf, 179
Köhler, Carmen, 120, 121
Köhler, Hanna, 147
König, Christoph, 211, 212
Lakhtar, Nakul, 206
Langenberg, Benedikt, 115
Lanz, Margherita, 50
Lawes, Mario, 215, 216
Lek, Kimberley, 221
Leyendecker, Birgit, 46
Livacic-Rojas, Pablo, 81
Livácic Rojas, Pablo, 75
Loeys, Tom, 236
Lorah, Julie, 262
Lorente, Sonia, 60
Losada, José Luis, 73
Losilla, Josep-Maria, 60
Lozano, José H., 248
Lozano, Luis Manuel, 85, 232
Lozano-Lozano, José Antonio, 73
Luh, Wei-ming, 80
López-López, José, 226
Lüdtke, Oliver, 144, 146
M. Dolores, Hidalgo, 160
Martinkova, Patricia, 101, 102
Mayer, Axel, 109, 115, 139, 170
Meinel, Christoph, 254
Meuleman, Bart, 107
Michiels, Bart, 71, 82, 217
Michta, Mikolaj, 45
Milica Miočević, Daniel, 265
Miocevic, Milica, 167
Moerbeck, Mirjam, 209
Moerbeek, Mirjam, 122, 123, 187
Moerkerke, Beatrijs, 236
Moeyaert, Mariola, 184
Moshagen, Morten, 154
Muela, Alexander, 76
Mukhopadhyay, Prabir, 206
Mulder, Joris, 164
Musow, Stephanie, 213
Mutz, Rüdiger, 172
Muñiz-Fernández, José, 143
Nadri, Mahsa, 97
Naumann, Alexander, 120, 213
Naumann, Felix, 214
Nestler, Steffen, 152
Ni, Haifang, 222
O’Rourke, Holly, 117
Oliver-Rodríguez, Juan Carlos, 123
Onghena, Patrick, 71, 82, 217
Onwuegbuzie, Anthony J., 69
Ortega-Campos, Elena, 79, 84
Ortiz, Camila, 68
Ostermann, Thomas, 148
Paap, Muirne, 200
Padilla, Jose-Luis, 85, 229, 232
Paladini, Chiara, 254
Pallares, Susana, 63
Pandkar, Sushant, 103
Partchev, Ivailo, 246
Pasikowski, Slawomir, 42
Pearson, Rebecca, 226
Peter, Sebastian Maximilian, 108
Pilarska, Justyna, 159
Ploetner, Jan, 173
Pohl, Steffi, 62, 141, 142, 175, 176, 233, 267
Pokropek, Artur , 178
Portell, Mariona , 72, 73
Preetz, Richard , 237
Prescher, Thomas , 263
Proyer, René , 83, 86
Rabe-Hesketh, Sophia , 209
Ramis, Yago , 63
Rdz-Navarro, Karina , 129
Rettig, Tobias , 59
Revuelta, Javier , 248
Robitzsch, Alexander , 144, 146
Roman, Zachary , 78
Rosel Remírez, Jesús F . , 182, 205
Rosseel, Yves , 127, 128, 164, 196
Rouzinov, Serguei , 218
Ruiz, F . Javier , 56–58
Röseler, Lukas , 48
San Martín, Ernesto , 171
San Martín, Ernesto , 202
Sanduvete-Chaves, Susana , 72, 73
Scharf, Florian , 151, 152
Schermelleh-Engel, Karin , 129
Schneider, Lennart , 149
Schultze, Martin , 216
Schuurman, Noémi , 185
Schweizer, Karl , 155
Schäfer, Thomas , 157, 158
Sengewald, Marie-Ann , 233
Serido, Joyce , 50
Serrano, Celia , 85
Shim, Soyeon , 50
Sijtsma, Klaas , 132
Siller, Heidi , 43
Smits, Niels , 98
Sorgente, Angela , 50
Soroa, Goretti , 76
Sorrel, Miguel A . , 99
Spoden, Christian , 147, 212
Steyer, Rolf , 92, 139, 169, 173
Strobl, Carolin , 149
Suero, Manuel , 54, 198
Suvarnan, Thirdev , 206
Suárez Coalla, María Paz , 77
Sánchez, Yolanda , 58
Sánchez-Algarra, Pedro , 69
Sánchez-Romero, Elisa Isabel , 55
Tagliabue, Semira , 50
Taloen, Wouter , 235, 236
Tanious, Renné , 82
Tanious, René , 71, 217
Tarka, Piotr , 230
ten Broeke, Guus , 252, 255
Thiem, Alrik , 261, 264
Tilling, Kate , 226
Tobi, Hilde , 251, 255
Tomasik, Martin J . , 145, 148
Tomás, Inés , 116
Tuero Herrero, Ellían , 75, 81
Ulitzsch, Esther , 62, 176
Vallejo Seco, Guillermo , 75, 81
Valor, Inmaculada , 85
Van Den Noortgate, Wim , 239
Van den Noortgate1, Wim , 184
van der Ark, Andries , 134
Van Oorschot, Wim , 107
Veen, Duco , 220, 222
Verdugo, Laura , 56
Vermunt, Jeroen K . , 227
Viladrich, Carme , 49, 60, 63
Villalobos, Cristobal , 68
Vinzuda, Vipul , 206
Vives, Jaume , 60
Voelkle, Manuel , 93
Voelkle, Manuel C . , 61, 225
Vogelsmeier, Leonie V.D.E . , 227
von Davier, Matthias , 176, 178
von Thienen, Julia P . A . , 254
Vilchez, María del Pilar , 55
Wallot, Sebastian , 52
Washbrook, Liz , 226
Weirich, Sebastian , 177
Widaman, Keith , 125, 126, 153
Wilcke, Juliane , 191
Wind, Stefanie , 135
Winter, Sonja , 221
Witte, Erich H. , 244
Woitschach, Pamela , 143
Yavuz, Sinan , 242
Zaldívar-Basurto, Flor , 79
Zeleis, Achim , 149
Zenker, Frank , 244
Zhao, Maggie Yue , 204
Zondervan-Zwijnenburg, Mariëlle , 222
Zumbo, Bruno , 143
Zyluk, Natalia , 45
Zúñiga, Claudia , 68
Impressum
Lehrstuhl für Methodenlehre und Evaluationsforschung
Institut für Psychologie
Friedrich-Schiller-Universität
Am Steiger 3, Haus 1
D-07743 Jena

Organizing Committee
Prof. Dr. Rolf Steyer
Andreas Neudecker
Katrin Schaller
Marcel Bauer

Abstract Book
Andreas Neudecker
Matthias Gehlert