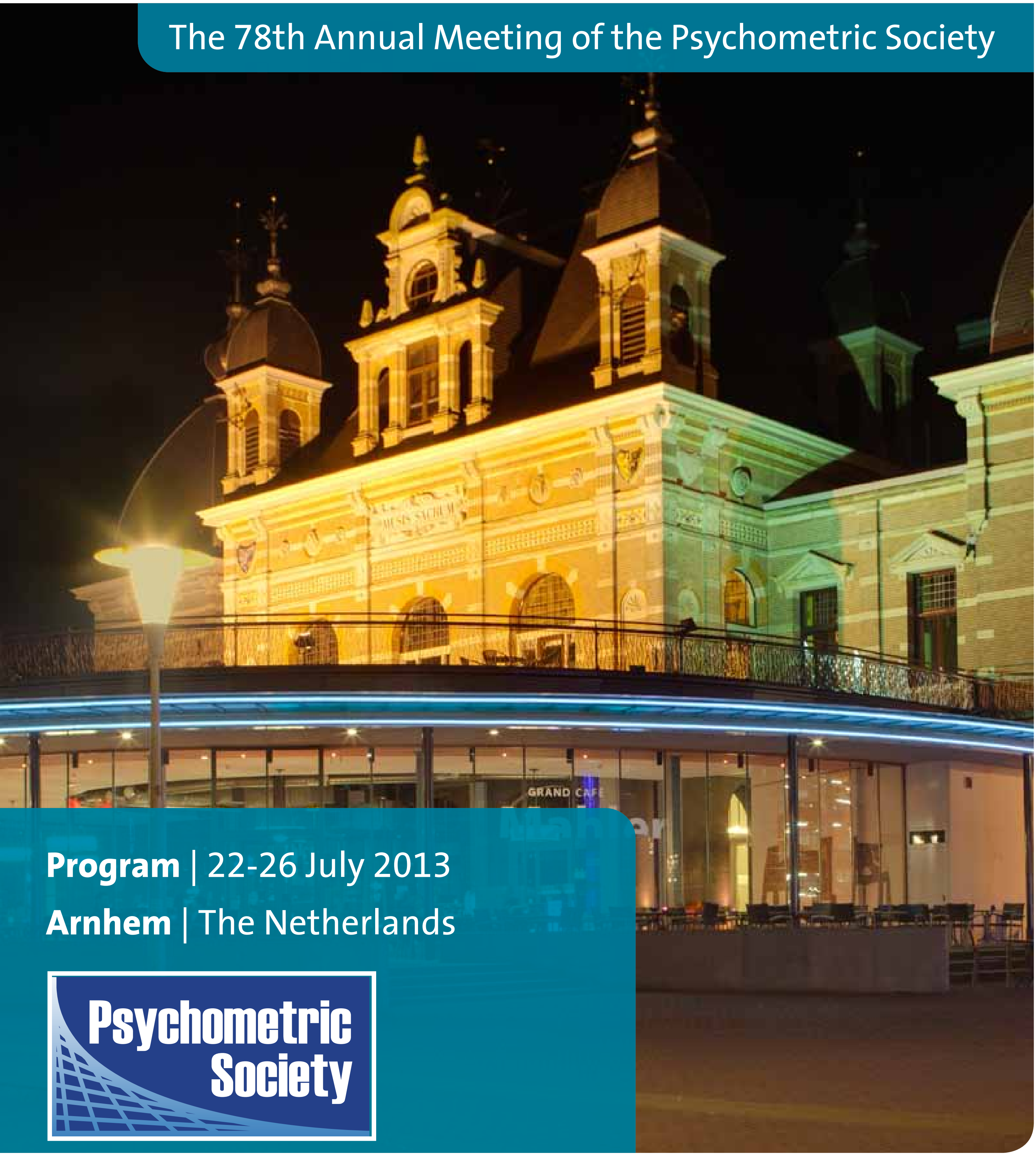
IMPS 2013

The 78th Annual Meeting of the Psychometric Society



Program | 22-26 July 2013

Arnhem | The Netherlands

Psychometric Society

78th Annual Meeting of the Psychometric Society

Arnhem, The Netherlands

Pre-conference workshop

July 22, 2013

Annual Meeting

July 23-26, 2013

Psychometric Society Officers

Officers of the society (August 2012 –July 2013)

President: Hua Hua Chang, University of Illinois at Urbana-Champaign, USA

President-Elect: Alberto Maydeu-Olivares,University of Barcelona, Spain

Past President: Mark Wilson, University of California-Berkeley, USA

Secretary: Terry Ackerman, University of North Carolina at Greensboro, USA

Treasurer: Luz Bay, Measured Progress, Dover, USA

Program Committee

Hua-Hua Chang, University of Illinois at Urbana-Champaign, USA

Terry Ackerman, University of North Carolina at Greensboro, USA

Alberto Maydeu-Olivares,University of Barcelona, Spain

Mark Wilson, University of California-Berkely, USA

Alina von Davier, Educational Testing service, Princeton, USA

Andries van der Ark, Tilburg University, the Netherlands

Hong Jiao, University of Maryland - College Park, USA

Anton Béguin, Cito Institute for Educational Measurement- Arnhem, the Netherlands

Local Committee, Cito Institute for Educational Measurement- Arnhem, the Netherlands

Anton Béguin

Patricia Gillet

Herbert Hoijtink

Ellen Omvlee

Saskia Wools

(and all the other people who did smaller parts of the work)

Table of Content

Program at a Glance - IMPS 2013

Monday: July, 22

Pre-conference workshops	9:00-17:00
Registration and Information Desk Open	14:00-17:00

Tuesday: July, 23

Registration and Information Desk Open	8:00-17:15
Welcome *Concertzaal*	8:45-9:05
Keynote - Wim van der Linden *Concertzaal*	9:05-9:55
State of the art lectures *Concertzaal & Parkzaal*	10:00-10:25
Break *Ravelijn*	10:25-10:45
Parallel session A	10:45-12:05
Lunch break Poster session I *Promenoir near the Balkonzaal*	12:05-13:20
Parallel session B	13:25-14:45
Invited talks *Concertzaal & Parkzaal*	14:50-15:30
Break *Ravelijn*	15:30-15:50
Parallel session C	15:50-17:10
Welcome reception Cito *Cito, Amsterdamseweg 13*	17:30-19:30

Wednesday: July, 24

Registration and Information Desk Open	8:30-13:00
Invited talks *Concertzaal & Parkzaal*	8:30-9:10
Parallel session D	9:15-10:35
Break *Ravelijn*	10:35-10:55
Parallel session E	10:55-12:15
State of the art lectures *Concertzaal & Parkzaal*	12:20-12:45
Social event: National Park: Hoge Veluwe and Kröller Müller	12:45-17:45
Conference diner - lifetime achievement award & travel awards	19:00

Thursday: July, 25

Registration and Information Desk Open	8:30-18:00
Dissertation award talk: Dylan Molenaar *Concertzaal*	8:30-9:10
Parallel session F	9:15-10:35
Break *Ravelijn*	10:35-10:55
State of the art lectures *Concertzaal & Parkzaal*	10:55-11:20
Keynote: Zhiliang Ying *Concertzaal*	11:25-12:15
Lunch break Business meeting *Jubileumzaal* Student luncheon *Balkonzaal*	12:20-13:35
Parallel session G	13:40-15:00
Break *Ravelijn*	15:00-15:20
Presidential Address: Hua Hua Chang *Concertzaal*	15:20-16:20
Parallel session H	16:25-17:45

Friday: July, 26

Registration and Information Desk Open	8:30-17:00
Parallel session I	8:30-9:50
Panel Discussion: Everything you always wanted to ask to senior people in the field *Concertzaal*	9:55-10:55
Break *Ravelijn*	10:55-11:15
Parallel session J	11:15-12:35
Lunch break Poster session II *Promenoir near the Balkonzaal*	12:40-13:55
Parallel session K	14:00-15:20
Invited talks *Concertzaal & Parkzaal*	15:25-16:05
Keynote: Lawrence Hubert *Concertzaal*	16:10-17:00
Closing Ceremony: best junior presenter & best poster awards *Concertzaal*	17:00-17:10

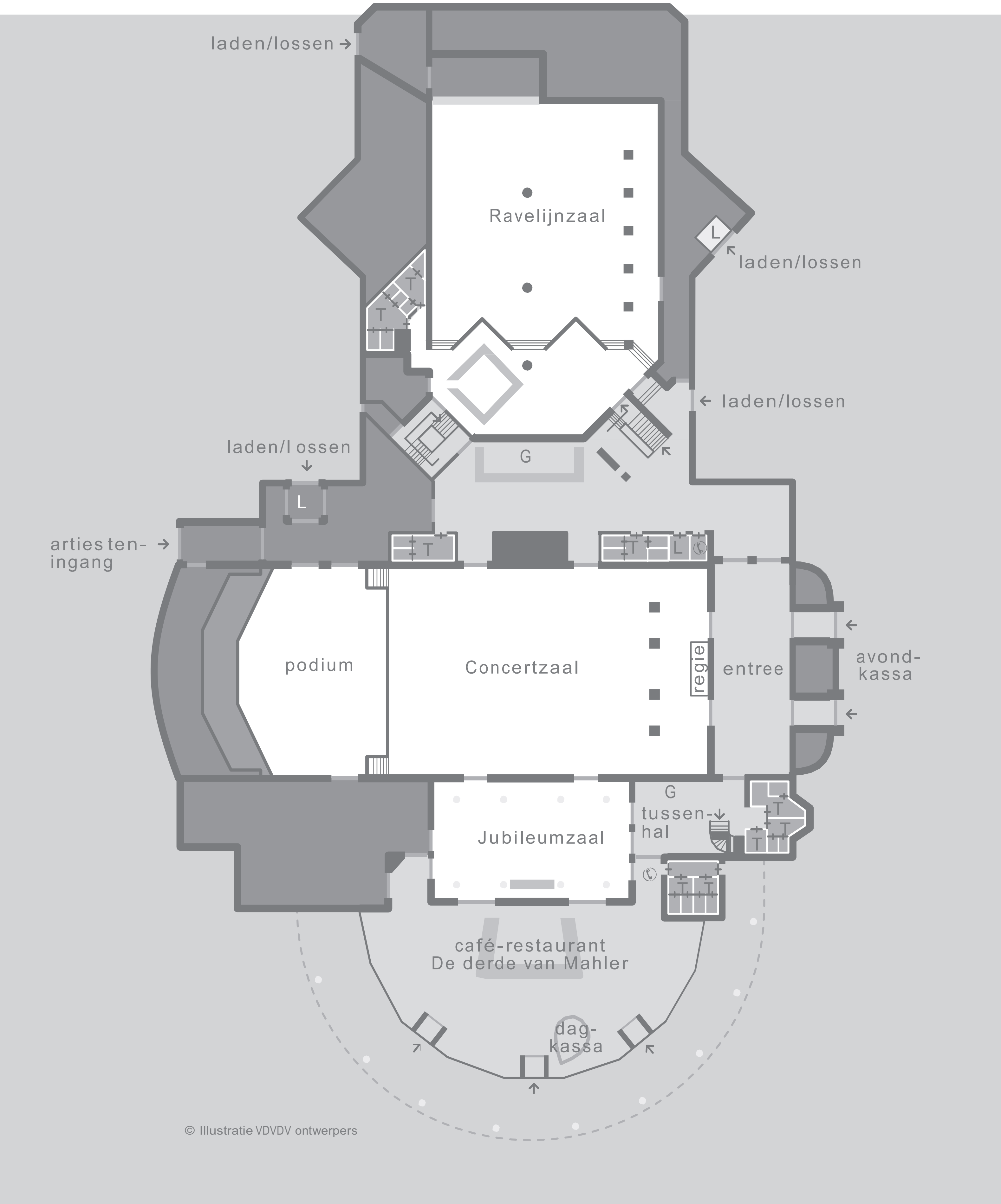
	Monday: July, 22					
9:00-17:00	**Pre-conference workshops**					
	Tuesday: July, 23					
8:45-9:05	**Welcome**					
9:05-9:55	**Keynote:**	Wim van der Linden				
10:00-10:25	**State of the art lectures:**	Doug Steinley	Chun Wang			
10:25-10:45	break					
10:45-12:05	**Fraud in the social sciences:**	**Variance Components Analysis**	**Differential Item Functioning -1**	**MIRT-1**	**(CCC) Classification, Clustering and**	**Applications 1**
	Jelte Wicherts	Rolf Steyer	Carol Woods	Huey-Min Wu	Choulakian Vartan	April Galyardt
	Marcel van Assen	Gregor Kappler	Ian Carroll	Hyesuk Jang	Hsiu-Ting Yu	Markus Nussbaum
	Marijtje van Duijn	Burhanettin Ozdemir	Jared K. Harpole	Jimmy de la Torre	Pieter Schoonees	Ray Y. Cheung
	Patrick Groenen	Eren Halil Özberk	Mian Wang	Jing-Ru Xu		Hendrik Straat
12:05-13:20	lunch break + poster session I					
13:25-14:45	**Bayesian Statistical Inference**	**IRT interpretation**	**Exploratory data analysis**	**SEM 1**	**Issues in assessment**	**MDS**
	David Kaplan	Alberto Maydeu-Olivares	Chin-Kai Lin	Renfen	Chun-Yen, Cheng	Kensuke Okada
	Peter Ebbes	Ernesto San Martin	Sacha Epskamp	Robert Cudeck	Michelle LaMar	Mark de Rooij
	Qiwei He	Peter van Rijn	Satoshi Usami	Taehun Lee	Rianne Janssen	Tomoya Okubo
	Xin Gu	Shanshan Qin	Yury Dodonov		Ya-Hui Su	Yoji Yamashita
14:50-15:30	**invited talks:**	Eric-Jan Wagenmakers	Steven Culpepper			
15:30-15:50	break					
15:50-17:10	**Invited symposium: Collaborative problem solving**	**Panel discussion: Challenges in Publishing Psychometric Manuscripts**	**Application or modeling of response times**	**SEM 2**	**Modeling data with differences in response behaviour**	**IRT**
	Alina A. von Davier	Chang	Ingmar Visser	Hao Luo	Danhui Zhang	Bor-Chen Kuo
	Peter F. Halpin		Jonathan Weeks	Xinyuan Song	Hong Jiao	Can Guerer
	Yoav Bergner		Usama S. Ali	Yves Rosseel	Marianthi Tzislakis	Jiyoung Jung
	Ilya Goldin		Yi-Hsuan Lee		Marie-Anne Mittelhaeuser	Terry Ackerman
17:30-19:30	welcome reception Cito					
	Wednesday: July, 24					
8:30-9:10	**invited talks:**	Theo Eggen	Daniel Bolt			
9:15-10:35	**Characterization of intra-individual dynamical processes**	**Equating-1**	**Patient-reported outcomes**	**Test design and simulation**	**Differential Item Functioning -2**	**Modeling**
	Ellen Hamaker	Björn Andersson	Abhijit Chatterjee	Angela Verschoor	Hongyun Liu	Hailemichael Metiku Worku
	Ingrid van de Leemput	Hyun-Woo Nam	Ruslan Jabrayilov	Njål Foldnes	Yao Wen	Marcel van Assen
	Sacha Epskamp	Ivailo Partchev	Wilco H. M. Emons	Takashi Akiyama	Yu-Wei Chang	
	Raoul P. P. P. Grasman	Marie Wiberg		Tzu-Yao Lin	Zhushan Li	
10:35-10:55	break					
10:55-12:15	**Invited symposium: Computerized Multistage Testing: Theory and Applications**	**Equating-2**	**Modeling and estimation of diagnostic models**	**SEM for LDA**	**Modeling growth and its consequences**	**IRT estimation 1**
	Chun Wang	Bartosz Kondratek	Chia-Yi Chiu	Axel Mayer	Jinah Choi	David Magis
	Wim van der Linden	Sayaka Arai	Hans-Friedrich Koehn	Sy-Miin Chow	Liping Sun	Luning Sun
	Duanli Yan	Won-Chan Lee	Karen Draney	Xin Tong	Ronli Diakow	Megan Kuhfeld
	Peter van Rijn	Yoshinori Oki	Mark Hansen	Zhenqiu (Laura) Lu		Sedat Sen
						Angela Verschoor
12:20-12:45	**State of the art lectures:**	Ying Cheng	Jingchen Liu			
12:45-17:45	**Social event: National Park: Hoge Veluwe and Kröller Müller**					
19:00	**Conference diner+ lifetime achievement award & travel awards**					

	Thursday: July, 25					
8:30-9:10	**Dissertation award talk:** Dylan Molenaar					
9:15-10:35	**Applied Psychometric Methods and Techniques: Achievements and**	**Bayesian IRT modeling**	**Modeling survey data**	**Factor analysis 1**	**Linking with parameter drift**	**Applications 2**
	Alina A. von Davier	Wang, Wen-Chung	Eunike Wetzel	Guangjian Zhang	Anton Béguin	Semirhan Gokce
	Jiahe Qian	Jin, Kuan-Yu	Jonathan M. Lehrfeld	Kohei Adachi	Meng Ye	Gonca Usta
	Mo Zhang	Matthew D Zeigenfuse	Joost R. van Ginkel	Takashi Murakami	Rui Guo	Khurrem Jehangir
	Jorge González	Bengt Muthen	Kristina Schmidt	Xiaoling Zhong	YoungKoung Kim	Yu Jiang
10:35-10:55	break					
10:55-11:20	**State of the art lectures:**	Ke-Hai Yuan	Irine Moustaki			
11:25-12:15	**Keynote:**	Zhiliang Ying				
12:20-13:35	lunch break + business meeting + student luncheon					
13:40-15:00	**Psychometric Modeling of Responses and Response Times**	**MIRT 2**	**Dif in diagnostic models**	**(CCC) Classification, Clustering and Correspondence Analysis -2**	**Missing data**	**Modeling issues**
	Dylan Molenaar	ChunLai Wu	Guaner Rojas	Ali Ünlü	Ashley Lawrence	David Torres Irribarra
	Maarten Marsman	Mariagiulia Matteucci	Jung Yeon Park	Gertjan van den Burg	Iris Eekhout	Shaobo Jin
	Paul de Boeck	Mark D. Reckase	LI Xiaomin	Michio Yamamoto	Yoonsun Jang	Ting Hsiang Lin
	Wim van der Linden		Zhuoran Wang	Joe Grochowalski		
15:00-15:20	break					
15:20-16:20	**Presidential Address:**	Hua Hua Chang				
16:25-17:45	**Reliability 1**	**IRT modeling**	**Measurement of change and growth**	**Understanding the individual: modeling dynamics & development**	**SEM 3**	**Computerized Adaptive Testing**
	Jules L. Ellis	Edward H. Ip	Rivka de Vries	Abe Hofman	Walter Herzog	M.M. van Groen
	Kappler, Gregor	Johan Braeken	Roger E. Millsap	Claudia van Borkulo	Fang Luo	Ping Chen
	Rashid S Almehrzi	Vicente Ponsoda	Tanja Krone	N.K. Schuurman	M.T. Barendse	Shiu-Lien Wu
	Matthijs J. Warrens	Yuanchao Bo	Zhen Li	Silvia Rietdijk		Usama S. Ali
	Friday: July, 26					
8:30-9:50	**Ordinal inference and latent variable models**	**SEM 4**	**Testing and measurement invariance**	**Cognitive diagnostic assessment 1**	**Forced choice**	**Reliability 2**
	L. Andries van der Ark	Iris A. M. Smits	Joran Jongerling	Ehsan Bokhari	Anna Brown	Tom Benton
	Robert Zwitser	Scott Monroe	Josine Verhagen	Lihong Song	Chia-Wen Chen	Saori Kubo
	Jules L. Ellis	Suzanne Jak		Wenyi Wang	Daniel Morillo Cuadrado	Tyler Hunt
	Rudy Ligtvoet	Yao Wen			Yin Lin	
9:55-10:55	**panel discussion** with Paul de Boeck, Alina von Davier, Cees Glas, Michael Kane en Jimmy de la Torre. Everything you always wanted to ask to senior people in the field					
10:55-11:15	break					
11:15-12:35	**Invited symposium: Network Psychometrics**	**Cognitive diagnostic assessment 2**	**Differential Item Functioning -3**	**IRT estimation 2**	**GLM**	**Factor analysis 2**
	Denny Borsboom	Laine Bradshaw	Carolin Strobl	Carl F. Falk	Ji Seung Yang	Elif Bengi Ünsal Özberk
	Gunter Maris	Matthias von Davier	Julia Kopf	Clemens Draxler	Minjeong Jeon	Karl Schweizer
	Angélique O.J. Cramer	Shuliang Ding	Mei Ling Ong	Frank Rijmen	Renske E. Kuijpers	Prathiba Natesan
	Francis Tuerlinckx	Tao Xin	Muhammad Naveed Khalid	Tammy Trierweiler	Rüdiger Mutz	Steffen Grønneberg
12:40-13:55	lunch break + poster session II					
14:00-15:20	**Cognitive diagnostic assessment 3**	**Missing data in IRT**	**Differential Item Functioning -4**	**Factor analysis 3**	**(CCC) Classification, Clustering and Correspondence Analysis**	**Problems with the use of NHST**
	Jianzhou Zhang	Cees Glas	Hui-Fang Chen	Marcel van Assen	David Torres Irribarra	Coosje Veldkamp
	Marcus Waldman	Dries Debeer	Jungkyu Park	Po-Hsien Huang	Marije F. Fagginger Auer	Marjan Bakker
	Shuliang Ding	Maria Bolsinova	Timo M. Bechger	Stella Bollmann	Zsuzsa Bakk	Michèle B. Nuijten
	Xin tao	Joost R. van Ginkel	Zairul Nor Deana Md Desa	Vincent Kieftenbeld	Stephen Aichele	Rink Hoekstra
15:25-16:05	**invited talks:**	Paul de Boeck	Herbert Hoijtink			
16:10-17:00	**Keynote:**	Lawrence Hubert				
17:00-17:10	**closing ceremony + best junior presenter & best poster awards**					

First Floor

Concertzaal, Jubileumzaal, café-restaurant ‘Mahler’, Ravelijnzaal

G | Cloakroom **T** | Toilets **L** | Elevator ☎ | Telephone

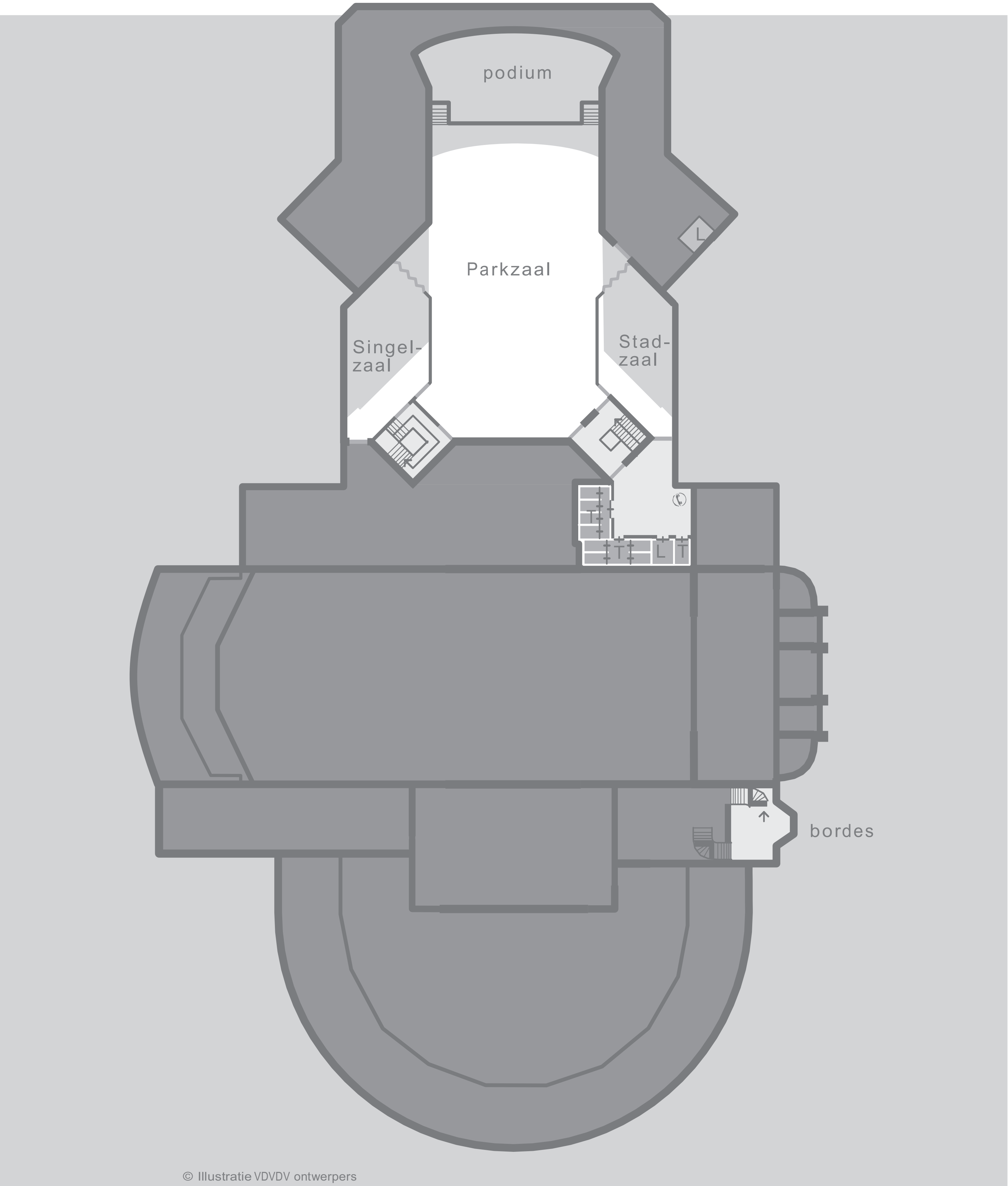


© Illustratie VDVDV ontwerpers

Middle Floor

Parkzaal, Stadszaal, Singelzaal

T | Toilets **L** | Elevator ☎ | Telephone

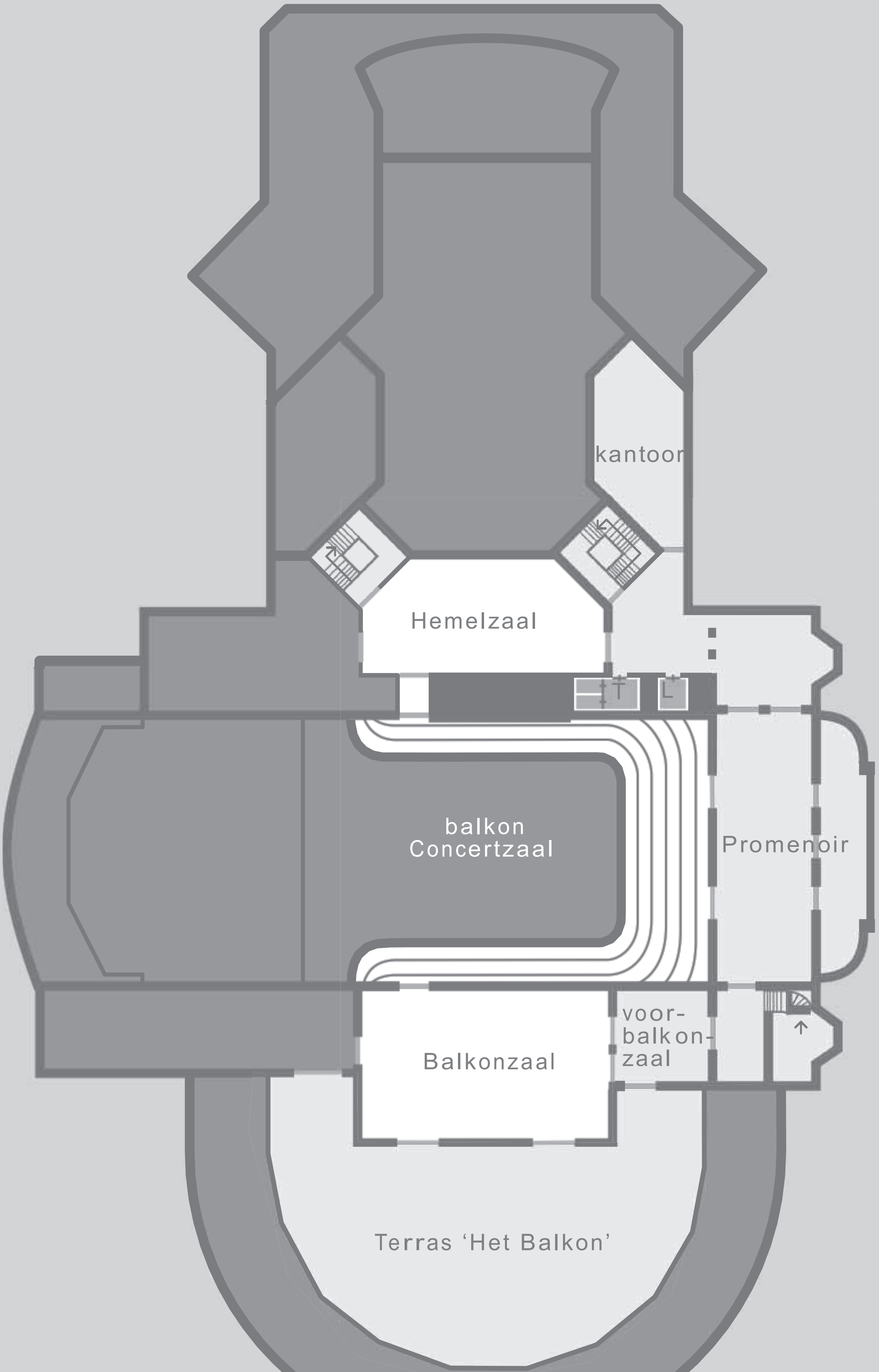


© Illustratie VDVDV ontwerpers

Second Floor

Concertzaal, Balkonzaal, Promenoir, Terras ‘Het Balkon’, Hemelzaal

T | Toilets **L** | Elevator



© Illustratie VDVDV ontwerpers

Monday July 22

Registration for Pre-Conference Workshop Attendees

08:30 – 13:30 Registration and Information Desk Open
Cito

Registration and Information Desk

14:00 – 17:00 Registration and Information Desk Open

Pre-Conference Workshops

09:00 – 13:00 **Pre-Conference Workshop A**

Mixture models and hidden Markov models
Ingmar Visser, *University of Amsterdam*
Maarten Speekenbrink, *University College London*

There are many situations in which one may encounter distinct types of entities, such as different animal species, and different states in which these entities may exist, for example motivational states like hunger. Cognition is sometimes also best understood in terms of discrete types and states. For example, forms of cognitive development can be characterized as the acquisition of increasingly complex rules [2]. Effectively, these rules constitute different types of reasoning and associated response patterns in reasoning tasks. And rather than a gradually shifting trade-off, people may switch rapidly between distinct decision-making modes favoring either speed or accuracy [1]. The idea that cognitive processes are guided by qualitatively different strategies underlies a wide range of theories concerned with topics such as word recognition, cognitive development, categorization, and decision making, to name but a few [4].

As types and states are generally not explicitly labeled, appropriate statistical techniques are required to identify them. This tutorial will focus on mixture and hidden Markov models, which are the basis of such techniques. In the context of MMs, a type or state (e.g., a motivational state or cognitive strategy) is formalized as a probability distribution over observables. Because a dataset may contain different types, the overall distribution is a mixture of such individual component distributions. As the component distributions need not be of the same parametric family (e.g., Gaussian distributions can be mixed with other distributions), MMs allow for considerable flexibility in the definition of types and states. HMMs are a natural extension of MMs, allowing switches between states over time. For example, these models are useful when people can switch between cognitive strategies during a task. In addition to identifying the different states, HMMs allow one to also focus on the process underlying state transitions.

While mixture models (MMs) and hidden Markov models (HMMs) are widely used in fields such as computational biology (e.g., for DNA sequence analysis) and machine learning (e.g., for speech recognition and estimation of topic models), their use in the analysis of cognition and behavior is relatively rare. This is unfortunate, as MMs and HMMs are ideally suited to test and explore important theoretical ideas in psychology. The objective of this tutorial is to provide researchers with an accessible introduction to MMs and HMMs.

13:30 – 17:00 Pre-Conference Workshop B

On assessing the dimensionality and the number of latent traits

Marieke E. Timmerman, *University of Groningen*
Urbano Lorenzo-Seva, *Rovira i Virgili University*

Identifying the number of latent traits underlying a set of observed items is a commonly encountered goal in psychological research. It is important in evaluating measurement instruments, and in research that explicitly aims at identifying distinguishing traits, like for example found in personality psychology. The simplicity of the task –just identifying a number – contrasts to the difficulties in selecting a suitable approach to perform the task. Many formal criteria exist, which may indicate different numbers for the same data. Furthermore, in spite of much research and debate, consensus is lacking about the most suitable approach. This workshop aims to provide the theoretical underpinning to value the different available approaches, and to sensibly select an approach in empirical practice. We further aim to stimulate the debate on this topic.

To achieve an understanding of the theoretical foundations underlying the various available criteria, we provide a theoretical framework. We categorize each criterion according to its associated model (e.g., common factor analysis) and to its definition of dimensionality (i.e., strict dimensionality versus major factors). We discuss that the dimensionality of a data set crucially depends both the subjects and the items involved, and on the specific model involved. We elaborate upon the meaning of the dimensionality, number of latent traits and the number of item sets associated with a single trait. We discuss its implications for the evaluation of measurement instruments.

Considering the framework, we discuss the properties of currently popular and / or well-performing criteria. In our overview, we include and relate criteria associated with Principal Component Analysis (as the Kaiser criterion, scree test, Horn's Parallel Analysis (PA)), with common factor analysis (as PA for common factors and/or ordinal variables, HULL), model selection approaches commonly applied in structural equation modeling (as likelihood ratio tests and goodness-of-fit criteria, like the RMSEA) and nonparametric item response theory (as a procedure in Mokken scale analysis).

Based on the theoretical properties of the criteria considered, we outline a strategy to proceed in empirical practice. We illustrate the use of the strategy with data collected from a normative sample of caretakers (n = 1,448) to evaluate a screening instrument for the detection of feeding problems in young children. We conclude with an open discussion.

Target audience:
The intended audience of the workshop includes anyone interested in identifying the number of latent traits and dimensionality assessment, who has some familiarity with factor analysis or item response theory.

Tentative program
1. Theoretical framework of dimensionality and the number of latent traits Break.
2. Overview of specific Criteria Break.
3. Strategy to proceed in empirial practice – empirical example – Discussion

09:00 – 17:00 Pre-Conference Workshop C

Developing and delivering CAT using open-source software
Michal Kosinski, *University of Cambridge*

During this one day hands-on workshop you will learn how to build, benchmark, and deliver an Computerized Adaptive Test using real and simulated data. You will use R, an open-source language for statistical computing, to simulate participants, responses, and item parameters, use real and simulated samples to build item banks based on dichotomous and polytomous Item Response Theory models, and simulate CAT to choose parameters, including the length of the test, starting and stopping rules, item selection criterion, etc. Finally, you will learn how to use Concerto, an open-source R-based adaptive testing platform, to deliver CAT to your respondents.

Requirements:
Laptops will not be supplied, so please bring your own and make sure that the conference's Internet connection is properly configured. BEFORE the workshop please download and install R (http://cran.rstudio.com/) and R studio (http://www.rstudio.com/). We will be available 15 minutes before the workshop to help you with that if necessary.

If you are new to R, we strongly recommend reading (and trying out the examples!) first 10 chapters of an official introduction to R (http://cran.r-project.org/doc/manuals/R-intro.html). We are going to briefly introduce R, so it is not a requisite**,** *but it takes only 2 hours* or so, letting you to focus on CAT and not R itself.

Programme:
Morning session:
1. Brief refresher on IRT models, IRT scoring methods, and CAT approaches
2. Brief Introduction to R
3. Simulating participants, responses, and item banks
4. Using real and simulated samples and item banks to simulate CAT in order to choose appropriate parameters
5. Presenting the results and choosing CAT parameters

Afternoon session:
6. Introduction to Concerto, an open-source adaptive testing platform
7. Building a linear test
8. Building an adaptive test based on item bank and CAT parameters developed during the morning session.

09:00 – 17:00 Pre-Conference Workshop D

Bayesian Structural Equation Modeling Using Mplus
Herbert Hoijtink, *Cito / University Utrecht*

Mplus (www.statmodel.com) is a software package that allows data analysts to use Bayesian structural equation modeling (BSEM). This one day preconference workshop will teach the participants how to use BSEM for data analysis. The workshop consists of four lectures. The structure of each lecture is the same: first of all an important feature of Bayesian data analysis will be introduced in a non-technical manner (formulas and derivations are virtually absent from this course); secondly, it will be elaborated how this feature is implemented in Mplus (Mplus code snippets will be given and explained); and thirdly, examples in which the feature plays an important role in data analysis will be presented.

The four lectures subsequently focus on the following features:

Lecture 1, 9.00am – 10.30am, Bayesian estimation using non-informative and informative prior distributions.
Lecture 2, 11.00am - 12.30pm, Bayesian estimation in the presence of missing data via multiple imputation.
Lecture 3, 14.00pm – 15.30pm, Bayesian model selection using the BIC and the DIC.
Lecture 4, 16.00pm – 17.00pm, The computation of error probabilities in the context of model selection.

Course Materials
In the first week of July 2013 the slides that will be presented during this course can be downloaded from http://tinyurl.com/hoijtink they can be found at the bottom of the page under sideline activities. Note that references to books and papers are included in the slides.

Participants
The course focuses on participants who want to learn about important features of BSEM when to goal is to use BSEM for data analysis. This course elaborates concepts and features that are important for the application of BSEM, it does not elaborate the technical foundation of BSEM (although references to technical elaborations will be given).

09:00 – 17:00 Pre-Conference Workshop E

Parametric and non-parametric tests of the Rasch model
Norman D. Verhelst, *Eurometrics*
Ivailo Partchev, *Cito Institute for Educational Measurement*

In this one day workshop, a number of different approaches to testing the Rasch model will be discussed. The theoretical backgrounds of existing tests will be considered, as well as the practical difficulties of their implementation, with special attention to the power of the tests.

Two approaches to test the Rasch model will be discussed in detail with the opportunity for participants to apply the technique to their own data set(s). The first approach is a new technique to generalize DIF analysis, called profile analysis, where hypotheses about differential functioning of various categories of items – instead of individual items – in an arbitrary number of groups can be tested. The second approach is the construction of non-parametric tests of the model. It uses the basic fact that in the Rasch model all data matrices with the same marginals are equiprobable. Whence it follows that, if we can draw a random sample of matrices with the same marginals as the observed one, we can approximate the sampling distribution of any univariate or multivariate statistic, where the accuracy of the approximation only depends on the number of matrices drawn. To draw such a sample, the R-package RaschSampler can be used. In the workshop, backgrounds of the RaschSampler will be explained, and participants will have the opportunity to build their own tests.

Norman Verhelst has been a senior researcher from 1985 to 2010 at the National Institute for Educational Measurement (Cito) in Arnhem, the Netherlands. His main interest was the development of IRT models belonging to the exponential family and being more general than the Rasch model. Since 2011 he is retired, but continues work in his own one-man company Eurometrics.

Ivailo Partchev studied statistics and sociology in Sophia, Bulgaria. He worked for some years for Bulgarian governmental institutions, and from 1999 to 2009 as researcher at the university of Jena (Germany). Since 2011 he is working at the National Institute for Educational Measurement (Cito) in Arnhem, the Netherlands. He is closely involved in implementing IRT models in the R-language.

Tuesday July 23

08:00 - 17:15 Registration and Information Desk Open

08:45 - 09:05 **Welcome**
Concertzaal

09:05 - 09:55 **Keynote session**

Concertzaal **Multidimensionality in Item Response Theory**
Wim van der Linden, *CTB / McGraw-Hill*

Generalizing item response theory to include multidimensional ability parameters seems like an obvious next step. But should we want these parameters to represent compensatory or conjunctive abilities? Or should the abilities be modeled to drive the success of individual steps in a cognitive process? And do we really need the generalization to test examinees on multiple dimensions? Your presenter struggles with these issues and would like to tell you why.

10:00 - 10:25 **State of the art lecture**

Concertzaal **Combining discrete and continuous latent variable modeling to better understand the structure of data**
Doug Steinley, *University of Missouri, Columbia*

This talk will explore combining aspects and cluster analysis and factor analysis to fully understand the structure of a data set. Particular attention is given to the robustness, or lack thereof, present in each modeling approach depending on the nature of the variables included in the data set. As such a more cautious approach (rather than the traditional "kitchen sink" approach) is recommended for both extracting classes and/or factors.

Parkzaal **Jointly Analyzing Response Times and Accuracy -- From Parametric Models to Semi-Parametric Models**
Chun Wang, *University of Minnesota at Twin Cities*

In computer-administered tests, response times can be recorded conjointly with the corresponding response accuracy, which broadens the scope of potential modeling approaches. Current models for response times, however, mainly focus on parametric models that have the advantage of conciseness, but may suffer from a reduced flexibility to fit real data. In this talk, I will present two new types of semi-parametric models that combine the flexibility of nonparametric modeling with the brevity and interpretability of the parametric modeling. The two models are (1) Hierarchical proportional hazard model, which adopts the hierarchical structure suggested by van der Linden (2007) with the well-known Cox proportional hazard (PH) model in survival analysis; (2) Hierarchical linear transformation model, which is a further extension of the Cox PH model. The linear transformation model represents a rich family of models that includes the almost all parametric models as special cases. Bayesian model estimation methods and model fit checking procedures will be briefly presented, and the applicability of both models will be demonstrated with a real response/response time data set. Going beyond the educational testing context, the potential of the proposed models in analyzing data arising from cognitive experiments will also be discussed.

10:25 - 10:45 ***Break***
Ravelijn

10:45 - 12:05 Parallel session A

Concertzaal **A.1.-Fraud in the social sciences: detection methods and cases**

Exposing Dr. Evil; An overview of methods of (detecting) data fabrication

Jelte Wicherts, *Tilburg University*

In this talk, I relate several methods to detect potential data fabrication to different types of data fabrication. I argue that the choice of method of fraud detection depends on the expected modus operandi of the perpetrator. The difficulties faced by fraudsters like “Dr. Evil” in fabricating data lie with (1) estimating realistic sizes of the hypothesized effects, (2) adding sampling error, (3) using a realistic psychometric model, and (4) mimicking typical behaviors by participants (e.g., failure to follow procedures, dropout). Many cases of exposed fraud involved typing-in of data by hand instead of the use of statistical software to simulate data, which often leads to a lack of sampling error and awkward raw data distributions. In other cases, perpetrators copied or altered existing data, which may lead to overly consistent results across supposedly independent samples. Other fabrication methods can be detected because they involve the making up of summary results without the fabrication of the underlying raw data, in which case the detection method should focus on finding inconsistencies in the reported results. I illustrate different detection methods on the basis of actual data from several cases of scientific misconduct.

Fraud and questionable research practices in the work of Stapel

Marcel van Assen, *Tilburg University*

On behalf of the 'Committee Levelt' I examined 46 papers of Stapel and six dissertations to detect evidence of fraud and questionable research practices. I will present the framework of the investigation and some procedures that were particularly useful in the Stapel case. The framework and procedures boil down to finding inconsistencies in the pairs 'research materials-data', 'research materials-article', 'data-article', or detecting extremely unlikely patterns in the data. I will also discuss the implications of the fraud investigation for the state of social psychological research in particular and empirical research in general.

Implications of fraud detection for the practice of reviewers and statistics education

Marijtje van Duijn, *University of Groningen*
Wendy Post, *University of Groningen*
Ruud Koning, *University of Groningen*
Don van Ravenzwaaij, *University of New South Wales*

While investigating the fraud into Stapel's scientific publications, it became clear that the distinction between questionable research practice (called “sloppy science” by the Stapel investigation committee) and fraud was often not clear, due to inconsistencies in the data and analysis as well as results that were “too good to be true”. The ‘mistakes’ can be categorized as omissions in the procedural details of the experiment, incorrect descriptive statistics, incorrect test statistics, and unlikely cell means. One might argue that if good research practice had been performed, the fraud could have been detected much earlier. Likewise, with the proper training, some incorrect and unlikely results could have been spotted by critical reviewers or colleagues.
Therefore, we propose two routes to reduce the risk of scientific fraud. First, we sketch an education program for all levels of scientific education, focusing on the the consequences of sloppy science in terms of validity and statistical issues. Second, we propose a checklist for reviewers to help them verify the correctness of results presented in a manuscript in order to improve assessment of its scientific contribution.

Simulation approaches in the detection of scientific fraud

Patrick Groenen, *Erasmus University Rotterdam*

In the recent past, there have some cases of scientific misconduct in the Netherlands in the area of social psychology. In this presentation, I will focus on one such case from experimental psychology. It is built around the ideas of random assignment of individuals to experimental conditions so that there should be enough variation in means amongst conditions with a similar predicted outcome. Based on ideas outlined in Simonsohn (in press) simulation studies have been performed for this case. I will outline the methodology that was used by the scientific integrity committee in this case and discuss some methodological concerns

Parkzaal

A.2.-Variance Components Analysis

Nonorthogonal Analysis of Variance (ANOVA) Revisited

Rolf Steyer, *University of Jena*
Gregor Kappler, *University of Vienna*
Axel Mayer, *University of Jena*
Lisa Dietzfelbinger, *University of Jena*

Nonorthogonal ANOVA has been discussed in many papers in the 1960ties and 1970ties. This discussion led to three methods (type I, II, III of sums of squares) that are nowadays implemented in all major statistical packages. If, for simplicity, we consider only *n x m*-factorial designs, all three methods yield tests of the two main effects and a test of the interaction between the two factors. It is argued that, if one factor, say X, represents the treatment (intervention, exposition) of interest and the other one, say Z, ‘only’ serves as a control that may affect the response variable Y and/or modify the effects of X on Y, then, except for very special cases, none of the three methods

(a) analyzes the effects of interest, i.e., the (Z=z)-conditional effects of X on Y, and
(b) tests what we would expect if we are interested in the main effect of X on Y,

namely the average of the (Z=z)-conditional effects.
We present a new method that remedies these critical points. This method is based on the distinction between fixed and stochastic regressors and can be implemented in any SEM program allowing for nonlinear constraints and for estimating the probabilities for the *n x m* cells of the design.

Nonorthogonal ANOVA with GLM, SEM and Bayesian models: a simulation study.

Gregor Kappler, *University of Vienna*
Steyer, Rolf, *University of Jena*
Mayer, Axel, *University of Jena*
Dietzfelbinger, Lisa, *University of Jena*

In quasi-experiments, treatment (X) and block factors (Z) are typically nonorthogonal and cell frequencies are usually not fixed but stochastic. We present models and tests of the following hypotheses:

1) The (Z=z)-conditional (or simple) effects of X on response Y are zero for all levels z of Z.
2) The (Z=z)-conditional effects of X on response Y is zero for a specific level z of Z.
3) The (Z=z)-conditional effects of X on response Y are identical for all levels z of Z.

4) The average effect of X on response Y (averaging the (Z=z)-conditional effects across the distribution of Z) is zero.

We report simulations for combinations of research designs and several sample sizes, analyzed by GLM and SEM with fixed and stochastic cell frequencies. Additionally to classical significance tests we also present Bayesian analyses of the SEM models. These simulations demonstrate that GLM results in extreme type I error rates for hypothesis 4. In contrast, Wald tests with SEM and Bayesian model tests are valid. Furthermore, SEM and Bayesian models allow for heterogeneous variances. Finally, we show how to extend these models for multiple categorical and continuous covariates.

Partitioning variance into constituents in multiple regression models: commonality analysis

Burhanettin Ozdemir, *Hacettepe University*

Commonality analysis is a method of partitioning the explained variance in a multiple regression analysis into the variance constituents associated with each independent variable uniquely and the variance associated with the common effects of one or more independent variables in various combinations. By partitioning variance, commonality analysis helps to determine accurately degree of multicollinearity between the independent variables, suppresor variable (if there is) and related importance of independent variables in a model. İn addition, commonality analysis provides regression affects (R2) of all possible simple and multiple regression models that can be constructed by given independent variables and thus it helps researchers decide the most appropriate regression model. The purposes of this study are to (a) provide a general overview of multiple regression analysis and its application, (b) explain how to conduct a commonality analysis in a regression model and (c) to determine the degree of multicolinearity between independent variables and suppressor variables in the model by means of commonality analysis results. For these purposes, OBBS data set, which was collected during a project in Turkey is, used to provide a heuristic example. In this example, three independent variables (interest=student's interest to turkish lesson, perception= student's perception of success and social science=students social science score) that are assumed to predict students turkish academic performance (dependent variable) were selected to create model and then multiple regression analysis and commonality analysis were conducted. It was found that social science and perception variables made significant contribution to model whereas interest did not. Because unique and common variance related to interest were very low. Although, regression analysis results showed no sign of multicollinearity, according to commonality analysis results there was multicollinearity between perception and social science variables. Uniquevariance of perception was small while common variance of perception and social science was very high that indicates multicollinearity between variables. In addition, all commonality coefficients found to be positive which means there is no suppressor variable in the model.

Comparing different coefficients in generalizability theory decision studies

Eren Halil Özberk & Selahattin Gelbal
Hacettepe University

The aim of this research is to compare reliability and validity using the two different variance component estimation procedure which rules proposed by Wiley (2001) for estimating standard errors held up across ANOVA and bootstrap procedures. Various coefficients and indices were used to interpret reliability and validity which recommended by Brennan (2001) and Kane (1999) in Generalizability Theory Decision Studies. Simulated data was generated for single facet conditions and standard error estimates were calculated for each estimated variance component and relative and absolute error variance across an ANOVA method and a variety of bootstrap procedures for each combination of conditions. It was found that, signal noise ratios produced adequate estimations for non-normal and dichotomous data in procedure. However, error tolerance ratios produced adequate estimations for non-normal and dichotomous data in   procedure. Thus, gives more information about validity and   gives more precise estimation of universe scores in G studies. This study provides support for the use of bootstrap procedures for interpreting reliability and validity in Generalizability Theory with estimating standard errors of estimated variance components when data are not normally distributed.

Jubileumzaal **A.3.-Differential Item Functioning -1**

The effect of item residual heterogeneity on three DIF testing methods

Carol Woods & Jared Harpole
University of Kansas

The assumption of item residual homogeneity is often untenable and generally unattended to in applications, but implicit in most methods for testing differential item functioning (DIF). For example, the underlying item response process may be more homogeneous for a reference group of Caucasian native-English-speaking Americans than for a group of non-native-English-speaking “Asians,” grouped together to make an adequate sample size for the analysis, which may refer to people from any part of China, Japan, Korea, etc. On psychological questionnaires, people with cognitive symptoms of disorders may experience more internal distractions like intrusive thoughts than people without disorders, leading to greater item residual variance. A simulation study will be presented in which Type I error and statistical power are evaluated in conditions with and without item residual heterogeneity, for binary logistic regression, two-group item response modeling (IRT-LR-DIF), and crossing-SIBTEST. Results will be presented showing that item residual heterogeneity caused severe inflated Type I error for DIF testing with all methods.

MIMIC DIF Testing When the Latent Variable Variance Differs Between Groups

Ian Carroll, *University of Kansas*

Multiple indicators multiple causes (MIMIC) models (Joreskog & Golberger 1975) can be employed in a psychometric context to test for differential item functioning (DIF) between groups on the measurement of a latent variable (Muthen 1989). MIMIC DIF models can be attributed some favorable properties when compared to alternative DIF testing methods (i.e., Item Response TheoryLikelihood Ratio DIF) such as having generally small sample size requirements while simultaneously maintaining reliably low Type 1 error rates and sufficient DIF detection power (Woods 2009). The mechanism by which MIMIC models test for DIF is to regress a latent variable and its nonanchor indicators onto an exogenous (grouping) variable. This allows the model to account for differences in the mean of the latent variable across groups, while also testing for uniform DIF in individual items. However, the model does not allow heterogeneity in the covariance structure of the latent variables themselves—it is assumed to be equal across groups.

A simulation study was conducted to examine the consequences of violating this assumption for the MIMIC DIF model. In this simulation, the following characteristics were varied: sample size, DIF effect magnitude, heterogeneity in latent variance between groups, magnitude of the group mean difference on the latent variable, and the ratio of focal group size to reference group size. Preliminary results suggest that violating the model’s equality of latent covariance structure assumption leads to systematically biased parameter estimates on factor loadings and estimates of the latent group mean difference, inflated Type 1 error in DIF detection, and several other undesirable statistical sideeffects.

This presentation will provide a more thorough explanation of the results for the present simulation, as well as discussing how these results can be used to inform sound methodology in substantive research.

The Effect of Local Dependence when Testing for Differential Item Functioning

Jared K. Harpole & Carol M. Woods
University of Kansas

In unidimensional item response theory (IRT) the assumption of local independence (LI) is imperative to proper item and person parameter estimation. When using IRT to assess the fairness of a psychological measure, tests of differential item functioning (DIF) are carried out. Before testing an instrument for DIF it is important to establish the presence of LI to ensure a valid DIF assessment. Chen and Thissen (1997) were first to define surface local dependence (SLD), which occurs when item pairs are similar in content or location. Studies indicate that psychological questionnaires may have violations of LI when item content is similar and item locations are adjacent (Stucky et al., 2011; Steinberg, 1994). Yet, currently there is no research on the effect of SLD on DIF testing. The present study seeks to assess the influence of SLD when testing for DIF using a variation of Lord's (1980) Wald χ^2 test implemented in flexMIRT (v. 1.04.3; Cai, 2012) with binary items. Results of a simulation study will be presented in which Type I error and power are evaluated in 36 conditions varying with respect to sample size, DIF magnitude, LI violation magnitudes, and LI or non-LI anchor sets).

Anchor Selection for Differential Item Functioning Using Wald Test

Mian Wang, *University of Kansas*

Methods for testing differential item functioning (DIF) require that the scale for the reference and focal groups is linked using group-invariant anchor items. Past research has shown that contamination of the anchor set results in biased parameter estimates and inflated Type I error. Several anchor-selection strategies have been introduced in an item response theory framework. However most of them use either iterative purification procedures or likelihood ratio testing with all others as anchors (AOAA). Iterative purification procedures are often cumbersome to carry out, and AOAA also requires multiple model fittings. The current study introduces a new strategy for anchor selection using a modified version of the Wald χ2 test that is implemented in flexMIRT (Cai, 2012). The accuracy of identifying DIF-free items and false alarm rate will be presented. Limitations of the study and suggestions for future research are also discussed.

Balkonzaal **A.4.-MIRT-1**

The Performance of the MIRT Plausible Values Method under BIB and NEAT Designs

Huey-Min Wu, *National Academy for Educational Research*
Tian-Wei Sheu, *National Taichung University of Education*
Bor-Chen Kuo, *National Taichung University of Education*
Wan-Ning, Chen, *National Taichung University of Education*

The purpose of this paper is to explore the performance of plausible values method under BIB and NEAT designs based on simulated data. The major focus of large-scale assessments is always on the population statistics, such as means and standard deviations, and the plausible value method is usually used to estimate the population parameters. For large-scale assessments the spectrum of subject matter is usually wide, but the testing time is short. Therefore, in order to cover the proficiency domain sufficiently, multiple booklets are used. Balanced incomplete block design (BIB) and non-equivalent groups with anchor test design (NEAT) are two popular test equating methods for this condition. The experimental results show that the estimating method based on plausible values estimate better than that of other methods in equating designs, and as the test length increase, population parameters (means and standard deviations) are well estimated.

Exploring the Estimation of Examinee Locations Using Multidimensional Latent Trait Models under Different Distributional Assumptions

Hyesuk Jang & Mark Reckase
Michigan State University

Multidimensional item response theory (MIRT) has been developed to meet the requirements of the data from real tests. One MIRT model, a bi-factor model has been designed and used for analyzing empirical data. This research evaluates the bi-factor model to determine how well the model works in various empirical contexts. While the distributions of the latent traits are often assumed to be normal distributions on the trait continuum, the distributions observed from empirical data do not always show the normal distribution. Also, despite the advantages of the bi-factor model, it has restrictions when the model is constructed: The latent traits are orthogonal. The results from this research will provide information on the estimation properties of the bi-factor models under conditions when assumptions about such as distributions and orthogonality are not met. Also the influence of item parameter is shown. Based on the information, the results can be applied to analyses using models of multidimensional latent traits and be significant in terms of providing information on measurement error for data analysis.

Multidimensional Higher-Order IRT Model with Multiple Groups

Jimmy de la Torre, Yan Huo & Eun-Young Mun
Rutgers University

The higher-order IRT (HO-IRT) model developed by de la Torre and Hong (2010) and de la Torre and Song (2009) assumes that multidimensional domain-specific latent traits can be linked to a single (i.e., unidimensional) higher-order latent trait via a certain linear function. Additionally, the HO-IRT model is applicable only to a single group. The current study proposes a multidimensional HO-IRT model to accommodate multiple higher-order latent traits and multiple groups. In this model, the various dimensions of domain-specific latent traits are partitioned into multiple clusters corresponding to the number of higher-order latent traits. The proposed model allows for the characteristics of the latent traits (i.e., means, variances, covariances, relationships between the higher-order and domain-specific traits) to vary by group. In addition to the model formulation, we developed Markov chain Monte Carlo (MCMC) algorithms to estimate the parameters for the multidimensional HO-IRT model. A simulation study was conducted to test the viability of the multidimensional HO-IRT model and the corresponding MCMC algorithms. The simulation study involved three groups with different mean vectors and covariance structures, two correlated higher-order latent traits, as well as six domain-specific latent traits, and preliminary results are promising.

The Evidence for a Subscore Structure in a Test of English Language Competency for English Language Learners

Jing-Ru Xu & Mark Reckase
Michigan State University

The research hypothesis is that a test that is well fit by a unidimensional model within a group of English native speakers, or for individuals that are very different in language background and instructional approach, might still support meaningful subscores for those of a particular language background, or those who received instruction in a particular way. The purpose of the research is to evaluate the data from a particular English language test to determine if it supports this hypothesis. That is, can meaningful and useful subscores be identified for a test whose data are well fit by a unidimensional item response theory model? Multidimensionality item response theories (MIRT) was applied to investigate the dimension structure of the test data. Missing data imputation, factor analysis, parallel analysis with 100 simulations, cluster analysis, reference composite computation using MIRT, correlations and corrections for attenuation between clusters identified based on subscores for different language groups were analyzed respectively for both highest frequency items across all 164 test forms and four independent highest-frequency forms to validate the dimensionality of subscore structure. The research proved the evidence of reporting subscores on different constructs in different language groups for a test fit by a unidimensional model using multidimensional IRT.

Stadszaal **A.5.-(CCC) Classification, Clustering and Correspondence Analysis -1**

Taxicab Correspondence Analysis of Ratings

Choulakian Vartan, *University of Moncton*

Let Y be an nxQ ratings dataset, where Q represents the number of items, and n represents the number of rated objects or the number of individuals expressing their opinions on the Q items. In the correspondence analysis (CA) literature on ratings 3 kinds of codings have been proposed: Benzecri's personal equation mapping of the ratings followed with barycentric fuzzy coding, the doubled dataset Ydouble of size nx(2Q) and the dataset Ynega of size nx(Q+1), where a column named nega is added representing the cumulative complementary columns. It is well known that CA applied to these 3 codings of the same dataset produces different results, and there is no equivalence relationship between any 2 of them.

Taxicab correspondence analysis (TCA) is a L1 variant of CA. We show the following new result: TCA of Ydouble is equivalent to TCA of Ynega , if and only if the first factor score of both analyses are a linear function of the sum score of the ratings. This will imply that the Q items point to the same direction; and following Sir D. Cox(2006) who titled his talk ``In praise of the simple sum score'', the simple sum score statistic of ratings can be used to summarize the underlying latent unobserved variable.

We also note the following important point: Often the 2nd and 3rd TCA factor scores are also interpretable, thus uncovering other aspects of the data which are not related to the sum score statistic that characterizes the first TCA factor; such as raters' response styles which represent the raters' personal equations. Examples will be provided.

Developing Measures of Dependency for Models with Multilevel Discrete Latent Variables

Hsiu-Ting Yu & Jungkyu Park

McGill University

Intraclass Correlation Coefficient (ICC) has been used as a measure of the amount of dependency due to nested data structure in multilevel models. The variance of the outcome variable is decomposed into two independent components: variance of the lower-level errors and variance of the higher-level errors. ICC is the measure of the proportion of the variance attributable to including the higher-level units. However, ICC as a measure of dependency defined in the random effects regression models cannot directly apply to Multilevel Latent Class Models (MLCM) since discrete latent variables are assumed at both higher and lower levels. Several measures of dependency that are suitable for MLCM are proposed. The statistical criteria unitized to develop these measures include: information on the classification errors, classification likelihood, homogeneity of the observations within each class/cluster, heterogeneity of classes/clusters, and entropy criterion. Numerical experiments on simulated and real data are used to illustrate the properties of the developed measures of dependency in MLCM. The application of these measures as model selection tools to determine the number of latent clusters and classes in MLCM will also be discussed.

Constrained dual scaling for detecting response styles in categorical data

Pieter Schoonees, *Erasmus University Rotterdam*

Dual scaling is a multivariate exploratory method equivalent to correspondence analysis when analysing contingency tables. However, for the analysis of rating data different proposals appear in the dual scaling and correspondence analysis literature. It is shown here that a peculiarity of the dual scaling method can be exploited to detect differences in response styles. Response styles occur when respondents use rating scales differently for reasons not related to the questions, often biasing results. A spline-based constrained version of dual scaling is devised which can detect the presence of four prominent types of response styles, and is extended to allow for multiple response styles. An alternating nonnegative least squares algorithm is devised for estimating the parameters. The new method is appraised both by simulation studies and an empirical application.

Hemelzaal

A.6.-Applications 1

Modeling Student Metacognitive Strategies in a Intelligent Tutoring System

April Galyardt, *University of Georgia*

When students use an Intelligent Tutoring System (ITS), they may utilize a variety of metacognitive learning and problem solving strategies. They may believe they know how to solve the problem and attempt it straight away. They may think they know how to proceed, but feel uncertain and ask the computer for a hint just in case. They may ask for several hints in a row, leaving us to wonder whether they are trying to learn the material from the hints or if they are simply abusing the hints. The log files stored by the ITS contain the records of all student actions.

I will present the results of using a Mixed Membership-Markov Chain Model to discover the strategies that students use. The sequence of actions that a student takes on a particular problem can be treated as a Markov chain: e.g., hint, long attempt, correct answer. Mixed membership models are an extension of latent classes that allow individuals to have mixed membership in multiple profiles. The mixed membership structure allows us to model how students switch between these metacognitive strategies. I fit this model as an unsupervised model to allow us to learn the patterns that students use from the data.

Implementing a Multilingual Selection Model across 27 European Countries

Markus Nussbaum, *European Personnel Selection Office*

The European Personnel Selection Office (EPSO) delivers a staff selection service on behalf of the institutions of the European Union. For each selection process, candidates from the 27 member states are assessed in order to select the best for possible recruitment as EU officials. The EPSO initiated a major overhaul of its selection processes under the EPSO Development Programme (EDP). From a psychometric point of view, one of the key features was the introduction of comprehensive item analyses on the basis of the Rasch model.

In March 2010, the first competition for graduates under this model was launched, and over 37,000 candidates were assessed against several competencies, including verbal, numerical, and abstract reasoning. For the following competitions for graduates in 2011 and 2012, over 30,000 candidates per year were assessed against the original competencies along with an additional situational judgment test. All of the competencies were measured through computer-based tests. The challenge in implementing the new selection model for these high-stakes exams was to ensure fair and equal treatment across the 27 member states and for the 23 different official languages. By equating across the different language versions, EPSO can ensure equal opportunities for all candidates.

Comparing Different Statistical Methods in Analyzing Multiple Informant Data
Ray Y. Cheung & Wai Chan
The Chinese University of Hong Kong

Multiple informant data (MID), which are collected from more than one person when measuring a construct, are becoming increasingly popular in research. However, there is no consensus among researchers in analyzing in present literature. Common methods to utilize MID include aggregation by weighted and unweighted means, combination by the "OR" and the "AND" algorithms, separation of measurement score into various components by mixing and matching contexts and perspectives, structural equation modeling and multilevel analysis. In light of this, a simulation study was conducted to systematically compare the performance of the methods mentioned across different condition. Practical guidelines will be provided to applied researchers based on the results obtained.

An IRT-based Extension of Angoff's Method for Standard Setting
Hendrik Straat & Gunter Maris
Cito Institute for Educational Measurement

Test developers use standard setting procedures to define cut scores between different levels of achievement or proficiency based on expert judgments of the test difficulty. One of the most-frequently used methods is Angoff's method, in which the main question is to predict how many theoretical minimally competent candidates would correctly respond to each item. In the literature, many modifications have been proposed with respect to iterative processes, presentation of item difficulties, and weighting the items with respect to their content validity. Nevertheless, the method and its modifications have been criticized because experts may not accurately predict the performance of minimally competent candidates on single test items and the consistency within and between raters is not evaluated. In this presentation, we propose an IRT-based extension of Angoff's method. We use IRT results to graphically represent the expected scores on clusters of test items on the total score scale. This graphical representation enhances intrarater consistency by providing visual feedback about the ratings over the item clusters. Finally, we evaluate the consistency of the raters by comparing the cumulative observed cut score distribution with the theoretical cumulative score distribution of the proposed minimally competent candidate.

12:05 - 13:20 Lunch break + Poster session I

1. **Discrete Time Survival Mediation Analysis: A Structural Equation Modeling Approach.**
   Amanda J. Fairchild, *University of South Carolina*
   Amanda Gottschall, *University of South Carolina*
   Katherine Masyn, *Harvard University*

2. **So How Well Do American Undergraduate College Students Know Themselves?**
   David V. Rudd, *Lebanon Valley College*

3. **A calculation method of the encounter probability with the uncollected kind of words in free descriptions**
   Kotaro Ohashi, Hideki Toyoda & Kazuya Ikehara
   Waseda University

4. **A Bayesian Hierarchical Model of Eye-tracking Data with Implication for the Development of Online Lexical Processing**
   Chansoon Lee, David Kaplan & Jan Edwards
   University of Wisconsin–Madison

5. **A study on the classification consistency of standard-setting methods for the speaking and writing test**
   Yongsang Lee, SeulKi Koo, HwangKyu Lim, KiJa Si & DoYoung Park
   Korea Institute for Curriculum and Evaluation

6. **An Automated Essay Scoring System in Korea**
   KiJa Si, DoYoung Park, Yong Sang Lee, Seul Ki Koo & HwangKyu Lim
   Korea Institute for Curriculum and Evaluation

7. **An option-based partial credit model parameter estimation using Metropolis-Hastings within Gibbs**
   Yuanchao Bo & Charles Lewis
   Fordham University

8. **The computerized testing tools for arm-leg coordination based on theory of sensory integration**
   Chin-Kai Lin & Huey-Min Wu *National Taichung University of Education*
   Bor-Chen Kuo & Chen-Yu Lin, *Graduate Institute of Educational Measurement and Statistic, National Taichung University of Education*
   Wen-Ching Su, *Department of Early Childhood Education, National Taichung University of Education*

9. **The Empirical Bayes Approach in Correspondence Analysis: Adjustment for Low Counts that Misrepresent Configurations of Estimates**
   Joe Grochowalski & Se-Kang Kim
   Fordham University

10. **The Relationship Between Degree of Attributes and Tests' Dimensionality and Difficulty**
    Feng LI, *Jiangxi University of Economy and Finances*

11. **Bivariate dependence patterns and Copulas: Model discrimination and Robustness**
    Lianne Ippel & Johan Braeken
    Tilburg University

12. **Exploration vs. Exploitation in the Estimation of Treatment Heterogeneity**Maurits Kaptein, *Tilburg University*

13. **Latent class analysis of attribute prevalence in cognitive diagnostic models**
    Jung Yeon Park & Matthew Johnson
    Columbia University

14. **A mixed model of moderated mediation and mediated moderation in career selection anxiety study**
    Zhonghua Liu, *University of Cambridge*

15. **Utility of the Summed Score and Weighted Summed Score in the Generalized Partial Credit Model**
    Daphna Harel, *McGill University*

16. **Estimating mediation effects with Mediators Being Survival Data**
    Jenn-Yun Tein, David P. MacKinnon & Yu Liu
    Arizona State University

17. **A Comparison of Angoff, Yes/No and Ebel Standard Setting Methods Using Generalizability Theory**

Ceylan Gundeger, *Hacettepe University*

18. **Regression analysis of additive hazards model with latent variables**
Deng Pan, *The Chinese University of Hong Kong*

19. **Applicable Distributional Conditions of Satorra-Bentler Scaled Test Statistic under Model Misspecification in Structural Equation Modeling**
Tzu-Yao Lin & Li-Jen Weng
National Taiwan University

20. **Multivariate Discrete and Continuous Maximum Entropy Distributions with Moments Constraints**
Yen Lee & David Kaplan
University of Wisconsin Madison

21. **Bayesian Causal Mediation Analysis**
Soojin Park & David Kaplan
University of Wisconsin-Madison

22. **Item Selection Optimization in CAT Early Stage with the Nominal Response Model**
Jing-Ru Xu, *Michigan State University*

23. **The Effects of the Model Selection and Anchor Item Format on the Equating of Testlet-Based Tests Score Under The Common Item Non-equivalent Group Design.**
Hwanggyu Lim,*Korea Institute for Curriculum and Evaluation*
Stella Yun Kim, *Yonsei University*
Yong Sang Lee, *Korea Institute for Curriculum and Evaluation*
Sang Wook Park, *Korea Institute for Curriculum and Evaluation*

24. **An Examination of Initial Condition Specification in the Structural Equations Modeling Framework**
Diane Losardo, Lu Ou & Sy-Miin Chow
Wireless Generation

25. **Reliability Evaluation through Comparison of SEM from Cut-off Score and Classification Consistency of Angoff and Bookmark Method**
Hee Won Yang & Guemin Lee
Yonsei University

26. **Effects of Inter-trait Correlation on Parameter Estimation in the MIRT Within-item Design**
Kyungtae Kim & Jwa K. Kim
Middle Tennessee State University

13:25 - 14:45 Parallel session B

Concertzaal **B.1.-Bayesian Statistical Inference**

Bayesian Model Averaging for Propensity Score Analysis
David Kaplan & Jianshen Chen
University of Wisconsin

An internally consistent Bayesian framework for model building and estimation must account for all forms of uncertainty. Although much of the focus of Bayesian inference is on parameter uncertainty, it is also understood that there is uncertainty in model selection. The current approach to addressing the problem of uncertainty lies in the method of Bayesian model averaging. Hoeting, Madigan, Raftery, and Volinsky (1999). This paper considers Bayesian model averaging as a means of addressing uncertainty in the selection of the propensity score equation for propensity score analysis. Given the full list of covariates in the propensity score equation, we provide a fully Bayesian approach to model selection allowing priors to be implemented for each selected model. A detailed simulation study of our approach examines the differences in the causal estimate when incorporating non-informative versus informative priors in the model averaging stage. In addition, we examine the impact of varying the size of Occam's window – a method used to narrow down the range of possible models. The application of our procedure to a small sample size problem is also presented. Our results show that the fully Bayesian model averaging approach provides excellent recovery of the true treatment effect, with more accurate intervals.

Attribute-Level Heterogeneity

Peter Ebbes, *HEC Paris*
John C. Liechty, *Penn State University*
Rajdeep Grewal, *Penn State University*
Matthew Tibbits, *Penn State University*

Modeling consumer heterogeneity helps researchers understand market structures and devise effective marketing strategies. In this research the authors study finite mixture specifications for modeling consumer heterogeneity when each regression coefficient has its own finite mixture, that is, an attribute finite mixture model. An important challenge of such an approach to modeling heterogeneity lies in its estimation. A proposed Bayesian estimation approach, based on recent advances in reversible jump Markov Chain Monte Carlo (MCMC) methods, can estimate parameters for the attribute-based finite mixture model, assuming that the number of components for each finite mixture is a discrete random variable. Attribute specification has several advantages over traditional, vector-based, finite mixture specifications; specifically, the attribute mixture model offers lower complexity for modeling heterogeneity and a more appropriate aggregation of information than the vector specification. In an extensive simulation study and an empirical application, the authors show that the attribute model can recover complex heterogeneity structures, making it dominant over traditional (vector) finite mixture regression models and a strong contender compared with mixture-of-normals models for modeling heterogeneity.

Combining Textual Analysis and IRT Scale Estimates Using a Bayesian Approach

Qiwei He, Bernard Veldkamp & Cees Glas
University of Twente

Analytic data is generally divided into two categories: structured data and unstructured data. To handle various data categories, different methods can be employed. For instance, item response theory is commonly used to estimate the latent trait of respondents based on structured numeric data that are collected from questionnaires, while text mining techniques seek to extract useful information from unstructured textual data sources through identifying interesting patterns. This paper presents an innovative trial to combine text analysis and IRT scale estimates into one systematic framework using a Bayesian approach. Results from text mining were used to elicit an informative prior for the IRT analyses. The objective of this study was to investigate whether combining the two models could enhance reliability of the test even further. The results showed that the prior information from textual analysis significantly influenced the location and shape of the posterior distribution of latent traits, and the model combination marginally increased the test reliability. In conclusion, the methodology presented here is very promising for use in combining information from both unstructured data and structured data, and a broader range of its application is expected in future studies.

Bayesian evaluation of inequality constrained hypotheses

Xin Gu, *Utrecht University*
Joris Mulder, *Tilburg University*
Maja Dekovic, *Utrecht University*
Herbert Hoijtink, *Utrecht University/Cito*

Bayesian evaluation of inequality constrained hypotheses has become an attractive alternative for the evaluation of null hypotheses, because it enables researchers to investigate their expectations with respect to the structure among model parameters. This presentation will propose an approximate Bayes procedure that can be used for the selection of the best of a set of inequality constrained hypotheses based on the Bayes factor in a very general class of statistical models. A software package BIG (Bayesian evaluation of inequality constrained hypotheses for general statistical models) will be introduced such that psychologists can use it for the analysis of their own data. Execution of BIG renders the Bayes factor that can determine the support in the data for each candidate hypothesis. To illustrate the approximate Bayes procedure and the use of BIG, we evaluate inequality constrained hypotheses in a path model and a logistic regression model.

Parkzaal

B.2.-IRT interpretation

Identifying the source of misfit in item response theory models

Alberto Maydeu-Olivares, *University of Barcelona*
Yang Liu, *University of North Carolina*

When an item response theory model fails to fit adequately, the items for which the model provides a good fit and those for which it does not need to be determined. To this end, we compare the performance of several statistics for item pairs with known asymptotic distributions under maximum likelihood estimation of the item parameters: a) a mean and variance adjustment to Pearson’s X2, b) a bivariate subtable analogue to Reiser’s (1996, 2008) overall goodness-of-fit test, c) a z-statistic suitable for ordinal data, d) Maydeu-Olivares and Joe’s (2006) M2 statistic. The unadjusted Pearson’s X2 and X2 with degrees of freedom heuristically equal to that of the independence model, as suggested by Chen and Thissen (1997) are also included in the comparison. The z-statistic and mean and variance adjusted X2 are recommended based on their Type I error and power results in the simulation study.

On the Statistical Interpretation of the Item Parameters in a Marginal Rasch Model without Resorting on Latent Variables

Ernesto San Martin, *Pontificia Universidad Catolica de Chile*

Rasch models are typically specified through a marginal-conditional decomposition. The marginal model corresponds to the distribution generating the person-specific abilities. The conditional model specifies the distribution of a binary random variable conditionally on both the person-specific ability and the difficulty parameter. In this context, the statistical meaning of the model is discussed using the conditional component of the model. By so doing, it is claimed that the difficulty parameter corresponds to an odd-ratio, whereas the person-specific ability corresponds to a betting-odd. However, the model which is fitted is the statistical model, which is obtained after integrating out the person-specific parameter. This step requires an additional assumption on the distribution generating such abilities.
This panorama raises the following question: Is it possible to make precise the meaning of the item parameters with respect to the statistical model, without resorting on a specific distribution generating the person-specific abilities? In this paper, we provide an answer based on the identification of the item parameters in a marginal

Rasch model, when the distribution generating the abilities is left unspecified. Based on this results, we propose a conditional pairwise mirror ratio method to estimate the item parameters. These ratios are exclusively based on pattern responses.

On the explaining away phenomenon in multivariate latent variable models

Peter van Rijn & Frank Rijmen
ETS Global

Many probabilistic models for psychological and educational measurements contain latent variables. Well-known examples are factor analysis, item response theory, and latent class models. We discuss what is referred to as the "explaining away" phenomenon in the context of such latent variable models (Pearl, 1988). This phenomenon can occur when multiple latent variables are related to the same observed variable, and can elicit seemingly counterintuitive conditional dependencies between the latent and observed variables (Hooker, Finkelman, Schwartzman, 2009). We formally define the explaining away phenomenon in the context of latent variable models and illustrate that the asymmetry in reasoning from latent to observed variables and from observed to latent variables is the main cause for the confusion (Bollen & Bauldry, 2011; Wellman & Henrion, 1993). Through a series of real examples, we aim to demystify this phenomenon for latent variable models. At the same time, we provide some guidance on the subtle differences between different multidimensional latent structures and their implications for practical use.

An Empirical Analysis of A Multilevel Model for Detection of Possible Test Tampering

Shanshan Qin, *University of Georgia*

Erasure analyses in educational accountability testing programs usually investigate potential tampering by teachers or administrators on the base of between-group differences on the prevalence of wrong-to-right (WTR) erasures. Typically, this is done without controlling for the impact of legitimate covariates, such as examinee ability. Methodologies that do exist focus on person level detection and do not appear to provide a probabilistic statement that a test has been tamped. In this study, we extend the IRT-based detecting methods by Wollack, Cohen, and Eckerly (2013) and van der Linden and Jeon (2012) to a multilevel analysis. We include the use of covariates as explanatory variables. These methods are compared and demonstrated using a data from a large-scale high-stakes test.

Jubileumzaal **B.3.-Exploratory data analysis**

Computerized Testing Tool for Visual Motor Integration: A Pilot Study

Chin-Kai Lin, *National Taichung University of Education*
Huey-Min Wu, *National Academy for Education Research*
Yu-Mao Yang, *National Taichung University of Education*
Bor-Chen Kuo & Kai Ching Chen, *Graduate Institute of Educational Measurement and Statistic, National Taichung University of Education*

Visual motor integration is one of the pre-writing skills for children. Scholars have proposed that visual motor integration may predict writing performance, and that it may also useful in evaluating a child's fine motor ability. The purpose of this study was to develop a computerized assessment tool of visual motor integration for children aged four to six years old, on the basis of basic Chinese strokes and structures.
The computerized testing tool for visual motor integration included three subtests: visual motor integration, motor coordination and visual perception. Each subtest consisted of 32 test items, 32 basic Chinese strokes announced by the Taiwan Ministry of Education. A pilot study was administered to three participants. The empirical data consisted of an analysis of the stroke sequence and the stroke length, as well as the proportion, placement, number, stability, pressure, direction and range

of the stroke, which were used to distinguish the performance of visual motor ability of the three participants. Integrating the experts' criterion, the result was that a total of 28 strokes were finally selected for the computerized assessment tool of visual motor integration. This tool could be used to diagnose writing performance problems in the early developmental stage of children.

Network Visualizations of Relationships in Psychometric Data and Structural Equation Models.

Sacha Epskamp, *University of Amsterdam*

I introduce the qgraph package (Epskamp, Cramer, Waldorp, Schmittmann, & Borsboom, 2012) for R, which provides novel visualization techniques, using networks, for psychometric relationships in general and correctional structures in particular. For instance, a correlation matrix can be represented as a network in which each variable is a node and each correlation an edge; by varying the width of the edges according to the strength of the association, the structure of the correlation matrix can be visualized. This technique has many applications, such as allowing a researcher to detect complex structures in a data set, validating the measurement model of a test and comparing individuals on differences in the correlation structure of repeated measures. Extending on the back-end provided by qgraph the semPlot package (Epskamp, 2013) can be used to visualize path diagrams, parameter estimates and implied and observed correlations of structural equation models (SEM). This R package can import the output of several popular SEM packages including Lisrel, Mplus and R packages sem, Lavaan and OpenMx. Finally, semPlot also provides a bridge between these packages, allowing users to construct input for one SEM package based on the output of another.

Generalized Sample Size Determination Formulas for Experimental Research with Hierarchical Data

Satoshi Usami, *Tokyo Institute of Technology*

Hierarchical data sets arise when data for lower units (e.g., individuals such as students, clients,nand citizens) are nested within higher units (e.g., groups such as classes, hospitals, and regions). In data collection for experimental research, estimating required sample size beforehand is a fundamental question to obtain sufficient statistical power and precision of focused parameters.
The present research extends previous research such as Heo & Leon (2008) and Usami (2011), deriving closed-form formulas for determining required sample size to test intervention effects in experimental research with hierarchical data, focusing on both multisite randomized trials (MRTs) and cluster randomized trials (CRTs). These formulas are derived considering both statistical power and confidence interval of effect size of intervention effects, based on estimates from the random intercept model for three-level data that considers both balanced and unbalanced designs. These formulas also address some important results such as the lower bounds of needed units at the highest levels.

Distance-weighted statistics: robust estimation of location, spread and association between variables

Yury Dodonov & Yulia Dodonova
Moscow State University of Psychology and Education

We introduce a new class of statistics based on what we call distance-weighting (DW). In computing distance-weighted statistics, a relative weight for each data point is calculated as the inverse sum of distances between the respective data point and the other points in the data array. Thus, outlying values are down weighted (although not completely discarded); importantly, no distribution parameters, such as mean, are needed as input information to calculate weighting coefficients. We give formulae for DW mean, DW median, DW variance and standard deviation, and DW covariance and correlation. For univariate data, we perform simulations to compare behavior of distance-weighted estimators with that of trimmed and Winsorized statistics and M-estimators under conditions of skewness and presence of outliers. In the case of bivariate data, robustness of DW correlation to outliers is compared to that of percentage bend and skipped correlations, biweight midcorrelation, and minimum covariance determinant and minimum volume ellipsoid estimators. DW estimators are shown to be well-behaved and relatively fast to compute. Advantages of DW-based measures of location, spread and strength of association between two variables, as well as possible extension to multivariate analysis, are discussed.

Balkonzaal **B.4.-SEM 1**

The latent structure of the CSHQ: A latent class analysis

Renfen, Zhangjianxin & Zhoumingjie

Chinese Academy of Sciences

Objective: The goal of this study was to investigate the latent structure of the CSHQ (childhood sleep habit questionnaire) using Latent class analysis. Methods: Chinese CSHQ Version was used to assess 912 adolescents in a nonclinical setting, and technique of Latent class analysis (LCA) was employed to deal with the dataset. Results: One to six class solutions were estimated for the sample. The statistical evidence point to the 3-class solution adequately fit the data from children. Latent Class Analysis revealed three classes of adolescent CSHQ: mild symptoms (n=489, 53.6%) moderate symptoms (n=398, 43.6%), severe symptoms (n=26, 2.9%).The results also showed that the entropy value .856 is large. Conclusion: Three-class model best fit the data for CSHQ symptoms and no significant gender difference of prevalence within each class.

PODs with LVs

Robert Cudeck, *Ohio State University*

Partial one-dimensional functions are used in the study of multiple groups in regression to highlight the fundamental components of the joint distribution of the response and explanatory variables. The graphical display associated with this summary is especially valuable, easy to construct and easy to understand. The same ideas extend directly to latent variable regression models and produce the same benefits. POD functions for multiple group SEMs can be estimated with popular software. The graph of results of a latent variable model is as effective with SEMs as it is with classical regression models.

An Alternative Diagnostic Measure for Detecting Nonlinear Relationships between Latent Variables

Taehun Lee, *University of Oklahoma*

In applications of structural equation modeling (SEM), detecting the existence of nonlinear effects (e.g. latent variable interactions) is an important issue. However, the use of conventional likelihood ratio test and other practical fit indices is problematic because these statistics are shown to be insensitive to the detection of omitted nonlinear terms in the structural part of the proposed model (Mooijaart & Satorra, 2009). Therefore, a researcher could reach an erroneous conclusion that there is no need to model a nonlinear relationship among latent variables when the proposed linear SEM is evaluated by the classical chi-square tests and practical fit indices. To detect the misspecification due to the omitted nonlinearity in the structural part of the model, we propose an alternative diagnostic measure based on a simulation-based model checking method known as posterior predictive model checking (PPMC) in the Bayesian literature (Gelman, Carlin, Stern, & Rubin, 2003). Contrasted with the existing solutions, the proposed method re-cycles the byproducts of model estimation, requiring no further model re-fitting (Mooijaart and Bentler, 2010). Further, the proposed method offers a natural way to conduct formal testing of the linearity assumption without asymptotic arguments (Klein & Schermelleh-Engel, 2010) or bootstrap re-sampling (Pek, Losardo, and Bauer (2011)).

Stadszaal **B.5.-Issues in assessment**

A knowledge Structure based Computerized Adaptive Dynamic Assessment

Chun-Yen, Cheng, *National Taichung University*
Chun-Yen, Cheng, *National Taichung University*
Huey-Min, Wu, *National Academy for Educational Research*
Bor-Chen, Kuo, *National Taichung University*
Chieh-Ting Li, *National Taichung University*

This study aims to explore the effects of different knowledge structure –based computerized dynamic assessments. Previous researches have shown that the performance of knowledge structure –based computerized dynamic assessments outperform than that of traditional computerized dynamic assessments. However, different algorithms to define knowledge structures such as domain experts' knowledge structures or students' knowledge structure were used in these computerized dynamic assessments. The pretest- posttest nonequivalent group design is adopted. The "addition and subtraction of fractions with different denominator" unit of "mathematics" utilized in elementary schools of Taiwan is adopted to develop the computerized dynamic assessments. The subjects are 106 fifth-grade students from 4 classes who are randomly divided into the experimental groups (A and B). In Groups A and B , the students' knowledge structure and domain experts' knowledge structures are incorporated into computerized dynamic assessments respectively. The results indicate as follows: (1) Both computerized dynamic assessments can promote the performance in learning addition and subtraction of fractions with different denominator. (2) The experimental group B is significantly higher than the experimental group A on the scores of the addition and subtraction of fractions with different denominator.

Using Markov Decision Processes to Infer Student Understanding in Complex Tasks

Michelle LaMar, Anna Rafferty & Tom Griffiths
UC Berkeley

Complex performance tasks often involve sequential action planning and experimentation. Such interdependent data are not easily modeled using traditional psychometric models. Taking a cognitive perspective, we propose modeling student decision making in such tasks as a Markov decision process (MDP). The MDP is a dynamic Bayesian model which assumes that the selection of action *a* in state *s* , *p(a|s),* is a function of a cost-benefit metric, *R(s, a)* , and an understanding of the state transition probabilities *T = p(s'|a,s)* within a defined state space S. A hypothesis space of student conceptions and misconceptions about the content material can then be defined as differing transition functions within the MDP, corresponding to different understandings of how their actions affect the system. Given a record of action-state pairs, the posterior distribution over the hypothesis space can be calculated providing both MAP estimates and measures of estimation certainty. We apply this approach to data from an educational game about cell biology to infer student understanding of cell organelle function. A simulation study demonstrates recoverability of latent classes assuming the MDP as the generative model. Application to student data demonstrates the feasibility of the approach for practical use.

Defining achievement levels using a domain-referenced approach

Rianne Janssen, *KU Leuven*

A simple, but common conceptual framework for describing growth in a certain domain refers to a continuous ability scale on which achievement levels are defined as ranges of latent proficiency (e.g., Masters, Adams, & Lokan, 1994). Each achievement level or developmental stage is characterized by a type of items that can be solved by a prototypical student of that level. As an example, the Common European Framework of Reference for Languages (CEFR) describes six achievement levels of learners of foreign languages, which are each defined by a list of “can do” statements of increasing complexity. Several procedures have been proposed to classify examinees on the basis of their test performance into one of the achievement levels of the CEFR. Apart from these procedures, psychometrics offers several approaches for domain-referenced measurement that are based on the a priori classification of the items according to domains (c.q., achievement levels). In the present paper, these different classification procedures will be compared, both conceptually and with regard to an empirical application on language proficiency assessment.

Using a modified multidimensional priority index for item selection under within-item multidimensional computerized adaptive testing

Ya-Hui Su & Yen-Lin Huang
National Chung Cheng University

Computerized adaptive testing (CAT) not only enables efficient and precise ability estimation, but also increases security of testing materials since different examinees are given sets of items from a large item bank. The construction of assessments usually involves fulfilling a large number of non-statistical constraints, such as item exposure control and content balancing. To improve measurement precision, test security, and test validity, the priority index (PI; Cheng & Chang, 2009; Cheng, Chang, Douglas, & Guo, 2009) and multidimensional priority index (MPI; Yao, 2011, 2012) were proposed to monitor many constraints simultaneously for unidimensional and multidimensional CATs, respectively. Many educational and psychological tests are constructed under multidimensional framework. Some of items (multidimensional items) in a test are intended to assess multiple latent traits. For instance, a science performance-based item can be used to assess both scientific declarative and procedure knowledge, a composition task can be used to assess both content understanding and language skills, and an arithmetic item can be used to assess both symbolic representation and calculation. However, Yao’s MPI was developed for the between-item multidimensional framework. When a within-item multidimensional test is assembled, a modified MPI method is necessary. Therefore, the purposes of the study are to derive algorithm of the modified MPI method under within-item multidimensional CATs, and to investigate the efficiency of the modified MPI method through simulations.

Hemelzaal **B.6.-MDS**

A robust Bayesian approach to latent-class multidimensional scaling

Kensuke Okada, *Senshu University*

The probabilistic latent-class multidimensional scaling method is extended. In former methods of probabilistic latent class multidimensional scaling, the latent variables are assumed to have a multivariate normal distribution within each latent class. However, this assumption is often not robust to deal with real-world conditions; outliers in data can sometimes greatly affect the estimates of the parameters in the normal component densities. Therefore, in this study, the use of the mixtures of multivariate t distribution is proposed instead of multivariate normal distribution. This is a natural extension of the traditional normal mixture models.Model selection scheme over the number of latent classes is also proposed. We develop a computational methods for Bayesian estimation of the robust latent-class multidimensional scaling using Markov chain Monte Carlo algorithm. Monte Carlo numerical experiments were conducted to examine the performance of the proposed method. The simulation results show that the proposed method is generally more robust to the non-normal noise.

Multivariate logistic regression in a distance framework

Mark de Rooij & Hailemichael M. Worku
Leiden University

Data with multiple categorical response variables are often collected in empirical sciences such as psychology, medicine, criminology, epidemiology, and life sciences. Latent variable models have been developed and are often applied for the analysis of such data. There are at least three major problems with these latent variable models: 1) Latent variable models make unverifiable assumptions about the data; 2) The latent variables in the models change meaning when a variable is added or deleted from the model; 3) An infinite number of latent variable models may provide the same fit to the observed data leading to an infinite number of conclusions, all based on faith instead of data. We propose a new methodology for this type of data by integrating two domains of statistics: Generalized linear modelling and multidimensional scaling. By combining these two, a methodology is developed that generalizes the standard logistic regression in a natural way and provides a biplot as a graphical aid for interpretation. No unwarranted assumptions are made using this approach. The methodology is illustrated using data from the Netherlands Study of Depression and Anxiety, where internalizing disorders are predicted from personality factors.

An extension of the mixed-preference model

Tomoya Okubo, *The National Center for University Entrance Examinations*
Kensuke Okada, *Senshu University*
Shin-ichi Mayekawa, *Tokyo Institute of Technology*

In this presentation, we discuss an extension of the mixed-preference model that enables us to illustrate intransitive choice in paired comparison judgments. Generally, multidimensional models assume a fixed preference (dominance) order for a subject or group during a session, but this is unable to describe intransitive judgments. However, Tversky (1969) and Montgomery (1977) reported that some subjects systematically violate transitive judgment in multi-attribute choice situations. Following these influential papers, a number of theoretical and empirical studies on the intransitivity of preference judgments have been presented. The mixed-preference model is an extended vector model that allows a subject or group to have multiple preference orders, and permits subjects to change their preference (dominance) order when judging paired comparisons. The proposed model can be considered as the ideal point model version of the mixed-preference model. Further, in this presentation, we will provide some results based on the proposed model.

MDS for series data by using candlestick valued dissimilarity measure

Yoji Yamashita & Hiroshi Yadohisa
Doshisha University

In this paper, we propose a new multi-dimensional scaling (MDS) method considering trends and variability with a series of dissimilarity measures between objects, when a series of data is obtained. Both upward and downward trends of the dissimilarity measures can be described because the dissimilarity measures are expected to change with the period. The variability of the dissimilarity measures is described as the minimum and maximum distances between objects. Our proposed method consists of two phases. In the first phase, we calculated the dissimilarity measure from a series of data. In particular, we use the candlestick valued dissimilarity measure, which consists of opening (beginning) value, highest value, lowest value, and closing(end) value within a given period, because we consider both trends and variability. Here the opening and closing values indicate the trend, and the highest and lowest values indicate the variability. In the second phase, we apply the proposed MDS method to the candlestick valued dissimilarity measures. In symbolic data analysis, we focus on higher-level objects (“ concept ”). In our proposed method, higher-level objects are defined as a pair of an object and several periods. The proposed method has two advantages. First, it can be applied to any type of data series. Second, when analysts focus on changing the distance between objects before and after an event, it is easy to visually understand the effects using the proposed method.

14:50 - 15:30 **Invited speakers**

Concertzaal **The Future of Psychometrics: An Outsider’s Perspective**
Eric-Jan Wagenmakers, *University of Amsterdam*

The goal of this presentation is to encourage a discussion on how Psychometrics can continue to develop as a scientific discipline. I outline several avenues for expansion and collaboration with other fields. In particular, I will argue that psychometricians can gain by becoming Bayesian, by using cognitive process models, and by sending more expeditions into territory currently occupied by disciplines such as mathematical psychology, judgement and decision making, and neuroscience.

Parkzaal **Understanding the reliability of teacher value-added effects**
Steven Culpepper, *University of Illinois at Urbana-Champaign*

The evaluation of teachers' performance in the classroom is an important application of educational testing and psychometrics. Many districts are adopting performance evaluations of teachers that use statistical models, which are referred to as value-added models, to characterize teacher performance in the classroom. This paper examines value-added teacher effects with the common factor model to understand the effect of latent variable characteristics on the reliability of observed teacher effects. Results explore two measures of teacher effectiveness: gain scores and covariate adjusted scores. The results provide evidence that student tracking, measurement bias, slope heterogeneity, and other parameters impact the reliability of value-added effects. The results serve as a guide for researchers, statisticians, and psychometricians who develop and refine methods for evaluating teachers.

15:30 - 15:50 ***Break***
Ravelijn

15:50 - 17:10 **Parallel session C**

Concertzaal **C.1.-Invited symposium: Collaborative problem solving**

Considerations on Including Collaborative Problem Solving Tasks in Standardized Assessments
Alina A. von Davier & Jiangang Hao
Educational Testing Service

Analyzing data from an assessment that includes collaborative problem solving (CPS) tasks involves several modeling aspects that are not encountered in the traditional tests. Two of the assumptions suggested by prior research are that people behave differently when they interact in teams from when they work alone, and that their individual domain skills might not correlate highly with the team’s outcome. Assessing these differences in behavior may lead to the augmentation of the individual domain score in isolation by an individual domain score in collaboration’ and a team score. The data from such an assessment will, therefore, contain process data and outcome data. Traditional assessment issues, such as reliability, validity, comparability of tasks, are discussed. Modeling strategies are also presented. For the process data, the usefulness of dyadic interactions, of dynamic models, and of hidden Markov models is discussed. For the outcome data, we need (a) the individual performance; (b) the group performance; (c) the contribution of each individual to the group performance.

IRT-based models are considered for outcome data. The goals of the study are: (a) the identification of patterns of the dynamics between the team members; (b) the models for the interactions dynamics; (c) the relation of these patterns to individual skills data as collected by the traditional components of the assessment. Advantages and limitations of the methods are described together with their utility to characterize processes in teams and prediction of outcomes.

Psychometric Models for Collaborative Problem Solving Tasks
Peter F. Halpin, *NYU Steinhardt*

Collaborative problem solving (CPS) requires that individuals work together to complete a complex task. The goal of this paper is to propose some broad principles and specific models that can be used to generalize traditional psychometric methods to the CPS context. Letting Xj denote the response pattern of the j-th group member, the factorization p(X1,... XJ) = p(X1) ... p(XJ) is taken to define the performance of a group that demonstrates no collaboration. This provides a baseline “independence model” by which to evaluate groups that perform better or worse than expected if their members had not collaborated. In particular, I propose the following two principles: (a) for a fixed time period, a group that performs better than expected by their independence model demonstrates (ability) collaboration; (b) for a fixed ability level, a group that performs faster than expected by their independence model demonstrates (time) collaboration. This paper also considers some specific models that can be recruited for certain types of CPS tasks. The central feature of these models is that they condition on both the (latent) ability of the individual being assessed, and the (demonstrated) ability of other group members -- i.e., they condition on task history.

Homework Collaboration via Discussion Boards in a Massive Open Online Course
Yoav Bergner, *Education Research Group, MIT*

Massive open online courses (MOOCs) have seen millions of new enrollments in the past year, reinvigorating research into how students use online learning environments. In the first MITx MOOC, 15% of student time was spent in discussion boards, which were the most frequently referenced course component during homework solving. A first step towards understanding which collaborative behaviors are helpful to whom is to identify which discussion threads deserve credit for helping students get unstuck. Data from the MITx course will be converted into a set of event series, in which a student goes from a wrong answer to the viewing of some number of threads to a right answer. With a large number of students consulting many of the same discussion threads, a dynamic Bayesian network model will estimate the helpfulness of each thread, dichotomized as a hit or a miss. By running the estimation model separately on data from students who have individually tested high or low prior to the homework instance, it may be possible to resolve different cognitive “zones” by the set of threads which operate optimally in each population. Posterior analysis of the hit threads will provide an opportunity for validating qualitative models of productive collaboration.

Towards Supporting Students and Instructors with Models of Peer Assessment
Ilya Goldin, *Carnegie Mellon University*

Collaborative problem solving may involve working individually with the benefit of peer feedback. Such problem solving has gained popularity as an instructional technique that is sometimes called Computer-supported Peer Review in Education. Although obstacles remain to wide adoption, psychometric modeling may help overcome these obstacles. One obstacle is that an instructor cannot examine peer-to-peer interactions because of the quantity of feedback generated by student peers, and therefore cannot monitor the quality of the student experience. Another obstacle is that stakeholders worry about the validity of peer assessments, e.g., because student peers are novices. We develop multilevel Bayesian models of peer review and evaluate them on real-world datasets. The models relate instructor scores of student essays to peer scores elicited by different peer assessment rubrics. We show how pooling across students and different representations of rating criteria affect model fit, and how they reveal information about student writing and assessment criteria. We propose that the models overcome some of the obstacles to adoption of peer review: the models may be used by an instructor to gain insight into the quality of the student experience, and the models may assuage stakeholders' concerns through better summaries of peer assessment than naïve averaging.

Parkzaal **C.2.-Panel discussion: Challenges in Publishing Psychometric Manuscripts:** ***Advice from the Editors of Six Journals***

We know all too well how difficult it is to publish a psychometric manuscript. The purpose of the panel discussion is to help younger investigators get much needed advice. The editors of six top tier journals in our field are invited to serve as panel members to offer guidance regarding basic components of a successful paper, insightful information about the review process, and crucial steps to handle revisions, as well as constructive ways of dealing with rejected papers. A question-and-answer session will follow the panel discussion to further equip researchers with the knowledge needed to successfully navigate the paper submission process..

Panel Members:

Hua-Hua Chang	(Applied Psychological Measurement)
Jimmy de la Torre	(Journal of Educational Measurement)
Matthew Johnson	(Journal of Educational and Behavioral Statistics)
Sandip Sinharay	(Journal of Educational and Behavioral Statistics)
Roger Millsap	(Psychometrika)
Matthias Von Davier	(British Journal of Mathematical and Statistical Psychology)
Mark Wilson	(Measurement: Interdisciplinary Research and Perspectives)

Jubileumzaal **C.3.-Application or modeling of response times**

The General Linear Ballistic Accumulator Model
Ingmar Visser, *University of Amsterdam*

The linear ballistic accumulator (LBA) model (Brown & Heathcote, 2008) has proven succesfull in modelling response times from experimental data. This paper presents and extension of this model in which the LBA parameters can be modeled with linear effects to accomodate explanatory variables. An R-package has been developed to fit these models using maximum likelihood estimation. The usefulness of this model is illustrated using developmental data from a Flankers task and a numerosity estimation task.

Using Response Time Data to Inform the Coding of Missing Responses
Jonathan Weeks, Matthias von Davier & Kentaro Yamamoto
Educational Testing Service

The estimation of item responses and examinee abilities depends on the correct coding of item responses and is particularly salient in the context of missing responses. For paper-and-pencil tests, missing response strings at the end of a test are usually indicative of items that were not reached and should be excluded from the likelihood. On the other hand, the interpretation of missing responses in-between non-missing values is less clear. Oftentimes these responses are coded as omits and treated as incorrect or as the lowest category for estimation purposes; however, the underlying response process for all missing items is only presumed. Using the timing information from computer-based tests, determinations regarding the coding of missing responses can be examined more systematically. The goal of this study is to develop a framework for screening item responses based on the speed of non-responses. When implemented prior to parameter estimation, this approach has the potential to increase the accuracy of estimates. When considered after the estimation, missing data patterns can be used to examine issues such as motivation or sub-population differences. The utility of this framework will be presented using literacy and numeracy items from the Program for the International Assessment of Adult Competencies (PIAAC).

Using Response Times for Scoring With Applications to Computer Adaptive Testing

Usama S. Ali & Peter W. van Rijn
Educational Testing Service

The use of computer based assessments has enabled recording of response times (RTs) on test items. In this paper, we explore the use of RTs in scoring rules to improve measurement precision. We have two applications in mind in which RTs can be of value: adaptive testing and formative assessment. If including response times in scoring improves measurement precision, it might be used in item-level adaptive testing (as known as CAT) or multistage testing (MST). Response times do not necessarily have to be used in the final score in CAT or MST, but might improve the item (or item module) selection algorithm (van der Linden, 2008). The second application is in formative assessment, where test items are often purposefully easier. Therefore, we hypothesize that the RTs can make a contribution to the distinction between students of differing abilities in addition to the item response. Our research is aligned with a series of currently relevant studies of using response time in scoring (Maris & van der Maas, 2012; Ranger & Kuhn, 2012). In a preliminary analysis of a set of mathematics items, it is demonstrated that including RT can lead up to a 10% increase of effective test length.

A Mixture Cure-Rate Model for Responses and Response Times in Time-Limit Tests

Yi-Hsuan Lee, *Educational Testing Service*
Zhiliang Ying, *Columbia University*

Many large-scale standardized tests are intended to measure skills related to proficiency rather than the rate at which examinees can work. Time limits imposed on these tests make it difficult to distinguish between the effect of low proficiency and the effect of lack of time. This paper proposes a mixture cure-rate model approach to address this issue. Maximum likelihood estimation is proposed for parameter and variance estimation for three cases: when examinee parameters are to be estimated given precalibrated item parameters, when item parameters are to be calibrated given known examinee parameters, and when item parameters are to be estimated without assuming known examinee parameters. Large-sample properties are established for the cases under suitable regularity conditions. Simulation studies suggest that the proposed approach is appropriate for inferences concerning model parameters. In addition, not distinguishing the effect of low proficiency and the effect of lack of time is shown to have considerable consequences for parameter estimation. A real data example is presented to demonstrate the new model. Choice of survival models for the response times is also discussed.

Balkonzaal **C.4.-SEM 2**

Comparing PIV, RULS, RDWLS and RML in the estimation of structural equation models of ordinal variables

Hao Luo, *Tsinghua University*
Fan Yang-Wallentin, *Uppsala University*
Shaobo Jin, *Uppsala University*

Ordinal variables are commonly encountered in analyzing structural equation models

(SEM). An approach of using polychoric correlations and fitting the models using unweighted least squares (ULS), diagonally weighted least squares (DWLS), or maximum likelihood (ML) is recommended. This paper compares the performance of the newly proposed polychoric instrumental variable (PIV) estimator with ULS, DWLS, and ML in the estimation of both confirmatory factor analysis model and general SEM. We evaluate the estimators' performances under various experimental conditions, including: (a) two types of models, (b) five sample sizes (200,400,800,1600,3200), (c) three levels of non-normality (normal, moderately, and extremely non-normal), (d) three numbers of category, and (e)three levels of misspecification. The average relative bias of the parameter estimates and the standard errors of parameter estimates as well as the power of the goodness-of-fit test statistics are compared.

Recent Development in Nonparametric Structural Equation Models
Xinyuan Song, *Chinese University of Hong Kong*

In the behavioral, social, psychological, and medical sciences, latent variables represent unobservable traits that are measured by multiple observed variables. Latent variable models are useful tools for assessing the interrelationships among latent and observed variables. Due to their wide applications, latent variable models have attracted significant attention from various fields. However, the majority of the existing works in the latent variable modeling are parametric. In this talk I will introduce several recent researches related to nonparametric latent variable models, including nonparametric structural equation models, transformation latent variable models, and varying-coefficient latent variable models. These models can be used to reveal the true relationships among latent and observed variables, analyze non-normally distributed multidimensional data, and examine time-varying effects of explanatory latent and observed variables on outcome latent variables. The Bayesian P-splines approach, together with Markov chain Monte Carlo algorithms, is developed to estimate unknown smooth functions, unknown parameters, and latent variables in the models. A modified deviance information criterion is proposed for model selection. The methodologies developed have been applied to several real-life studies in medical and behavioral sciences.

Lavaan and the history of structural equation modeling
Yves Rosseel, *Ghent University*

For several decades, software for structural equation modeling was commercial and/or closed-source. Three years ago, the lavaan project (http://lavaan.org) was started to create a fully open-source platform for latent variable modeling. In this presentation, I will discuss how the lavaan project attempts to capture and preserve the long and rich (computational) history of structural equation modeling and related methods. In the tradition of software archeology, several legacy SEM software packages were studied in order to understand and recover the (computational) details that were (and often still are) being used. By implementing many of these details into lavaan, we are able to 1) reproduce results reported in older papers and book chapters, 2) explain why we observe many subtle (and less subtle) numerical differences in the output of current SEM programs, and 3) study and compare these computational differences in order to better understand their characteristics.

Stadszaal **C.5.-Modeling data with differences in response behaviour**

Comparisons of Models Applied to Achievement Assessment with Different Curriculum Coverage
Danhui Zhang, *Beijing Normal University*
Matthias Von Davier, *Education Testing Service*
Xiuna Wang, *National Assessment Center of Education Quality*
Yang Cui, *National Assessment Center of Education Quality*

As the large-scale assessments in China has gained more and more attention in recent years, the development of the well designed assessment framework, while all the items, ideally, showing no bias toward or against any participating parties became a critical concern. In China, the diverse curriculum coverage in science presents a challenge for the assessment, especially under the scenario that we are trying to compare the final scores. The current study investigated how the different curriculum coverage and diverse science textbook version might affect 8th grade students' science achievement, and whether they offer advantage or disadvantage to certain group. Multiple group IRT models was developed with the aim to address such questions through allowing partial items function differently across diverse group. Through comparing both Rasch model and 2PL model with (1) including only universal parameters and (2) including both universal parameters and group-specific parameters, it was found out that, (1) while estimating the science ability as one-dimension ability and multi-dimention ability, the item parameters varied across different groups slightly, and (2) introducing a subset of curriculum group-specific parameters to the traditional "one universal group parameter" model will improve the overall model fit and item fit.

Mixture Hybrid Item Response Theory Modeling with Different Functional Forms across Latent Classes

Hong Jiao & George Macready
University of Maryland

Mixture item response theory (IRT) models have been proposed to deal with psychometric issues such as differential item functioning between latent groups (e.g., De Ayala et al, 2002; Kelderman & Macready, 1990), and identification of problem-solving strategies (e.g., Mislevy & Verhelst, 1990). Many of these studies extended the Rasch model or the two-parameter IRT model to a mixture version. The functional forms remain the same across latent classes; only item parameters vary across latent classes. This study proposes a mixture version of hybrid IRT models where functional forms differ across latent classes. The mixture hybrid models are illustrated in identifying latent guessing and slipping group, and the latent group not affected by either guessing or slipping. The unaffected latent group's responses are modeled by the Rasch model, the guessing group's responses are modeled by an extended Rasch model with a lower asymptote, and the slipping group's responses are modeled by an extended Rasch model with an upper asymptote. The model parameter estimation is explored using the Markov Chain Monte Carlo estimation method in OpenBUGS. The latent class membership for each person is determined by comparing the posterior probability in each latent class. Model parameter recovery is evaluated under simulation conditions.

An extended Rasch model to measure the dependence of school performance and motivation

Marianthi Tzislakis, Dominik Wied, Rebecca Hartmann &Nele McElvany
TU Dortmund

Beside pupils' ability and their school performance, there is also a connection between pupils' motivation and their school performance, compare e.g. Haahr et al. (2005).
As far as we know there is no model to measure the dependence of motivation and school performance while taking into account the pupils' ability. To handle this task we present an extended Rasch model including both latent constructs "motivation" and "ability".
For describing "motivation" the "Motivation and Engagement Wheel" defined by Martin (2007) is used. Within this definition motivation is split into adaptive and maladaptive behaviours and thoughts. The definition of school performance depends on the

application context, e.g. we consider mathematical performance.
The motivation characteristics for maladaptive and adaptive dimensions of motivation are pre-estimated and appear as \confounders" in the extended Rasch model. The parameters in the extended Rasch model are estimated by maximum likelihood estimations. We analyse the model in simulations and apply it to a survey dataset (N=376, 50.3% female, pupils aged 9-11), which we have collected recently.
The results will be interesting for educational researchers as well as for teachers to provide individual support for pupils with different motivation characteristics and to help them improve their school performance.

The usefulness of low-stakes anchor items in linking high-stakes tests: A simulation study

Marie-Anne Mittelhaeuser, *Cito/Tilburg University*
Anton Béguin, *Cito Institute for Educational Measurement.*
Klaas Sijtsma, *Tilburg University*

If a test is administered in a low-stakes condition, students might not put their best effort into answering items correctly. Within the item response theory (IRT) framework, lack of motivation threatens the consistency of proficiency and item parameter estimation and thereby the usefulness of the IRT model. Therefore, the use of anchors administered in high-stakes conditions is preferred when linking high-stakes tests. However, due to practical limitations, often linking data is used in examinations and state-wise assessments that is administered in low-stakes conditions. Fortunately, a mixture Rasch model can be used to identify classes in data resulting from different types of response behavior. By constraining a mixture Rasch model, one can identify the latent classes in such a way that one of the latent classes represents high-stakes response behavior while the other latent class represents low-stakes response behavior. A simulation study was conducted to compare performance of IRT linking based on the Rasch model and IRT linking based on the mixture Rasch model, where the link between two test forms is established using anchor items administered in a low-stakes condition. The current study provides useful information about the applicability of a procedure to minimize the condition effect in these instances.

Hemelzaal **C.6.-IRT**

The higher-order item response model with ancillary variables

Bor-Chen Kuo, Hsiao-Chien Tseng & Chun-Hua Chen
National Taichung University

The higher-order item response framework specifying an overall and multiple domain abilities in the same model that is natural in many real situations. Many researches show that incorporating student's ancillary variables such as gender, age, race, and grade level into the estimation process can lead to unbiased and more precise ability estimates. Some multilevel models based on unidimensional and multidimensional item response theories were developed for this purpose. However, there is no study incorporating higher order item response theory with the ancillary information.
The goal of this study is to propose a multilevel higher order item response model in which student's ancillary variables are treated as regressors of the overall ability. Markov chain Monte Carlo algorithm is applied to estimate the parameters in the proposed multilevel model. Simulated data are analyzed to establish the usefulness and feasibility of the proposed model. The simulated data are generated based on HO-IRT models with or without ancillary information. Then, the four models (HO-IRT and multidimensional-IRT models with or without ancillary information) are used to fit the generated data for exploring the performances of different models.

A comparison between the Fusion Model and the Mixed Rasch Model applied to educational data

Can Guerer & Clemens Draxler

Ludwig-Maximilians University Munich

This study is concerned with an empirical comparison of two very general approaches of psychometric modelling in the context of educational research, the Fusion Model (Hartz, 2002) and the Mixed Rasch Model (Rost, 1990).
The Fusion Model assumes a set of dichotomous variables, each called a mastery, with a certain combination of these being called a strategy. Further, it is assumed that these strategies then have an impact on the response probability for each item. The Fusion Model primarily focuses on the value of these dichotomous masteries for each respondent.
When considering two masteryvariables there are exactly four groups of mastery/non mastery combinations, which in turn may also be modelled by a Mixed Rasch Model with four groups.
This research compares the empirical adequacy of these two models using data of 300 students' reports. The reports are rated dichotomously regarding XY questions, which respectively are regarded as the items of a questionaire.
The models' parameters are estimated differently. While the Fusion Model utilizes an MCMC procedure, estimating item- and person-parameters simultaneously, a cML-approach is applicable within the Mixed Rasch Modell.
The models will mainly be compared based on information criteria (BIC, AIC, CAIC).

A hierarchical IRT model for identifying DIF in TIMSS 2007

Jiyoung Jung, *Yonsei University*
Yongsang Lee, *The Korea Institute for Curriculum and Evaluation*

As large-scaled academic achievement assessment drew attention, the fairness of the assessment has been an important issue to compare students' ability. Especially in an international comparison study such as the Trends in International Mathematics and Science Study (TIMSS), the fairness of the assessment should be investigated to verify the its findings. In Psychometric point of view, item differential functioning has been intensively discussed to deal with this assessment fairness, and consequently various methods have been developed to investigate the Differential Item Function (DIF). DIF is a statistical analysis to test the fairness on the item level. To detect DIF, various approaches have been applied and designed. One of these methods is the hierarchical IRT model, IRT model with a hierarchical structure on the item side. This approach would make it possible to predict DIF on a conceptual level beyond individual items, the final goal is to explain DIF by showing that it can be reduce by introducing appropriate context variables or by restricting the applicability of the test to certain subgroups or conditions. The purpose of this study is to investigate equivalence of mathematics test between Korean and Singarpore having similar property from achievement result in TIMSS 2007.

Graphical Representations of Items and Tests that are Measuring Multiple Abilities

Terry Ackerman & Bob Henson
University of North Carolina

This talk explores ways to graphically represent aspects of multidimensional item response theory (MIRT) models. Most practitioners are very familiar with graphs of item characteristic curves (ICCs), test characteristic curves (TCCs), information functions, etc. However, when tests are multidimensional item characteristic curves become item response surfaces, test characteristic curves, become test characteristic surfaces, and information functions depend upon the composite being measured. In addition, newer MIRT models common referred to as diagnostic classification models (or cognitive diagnostic models) have increased in popularity, which assume underlying discrete latent classes. Although some graphical representations of the models have been presented, little has been done to broaden these representations to diagnostic models as well. This presentation will examine graphical representations for

several different MIRT models including the compensatory, the noncompensatory and the diagnostic classification models. Graphical representations can provide greater insight for measurement specialists and item/test developers about the validity and reliability of the multidimensional tests. They can provide a link between quantitative analyses and substantive interpretations of the score scale and provide a feedback loop that can inform the test development process.

17:30 - 19:30 Welcome reception Cito

All participants are cordially invited to attend the opening reception at Cito.

Address: Amsterdamseweg 13, Arnhem
(Next to the railway station, exit Sonsbeekzijde)

Wednesday July 24

08:30 – 13:00 Registration and Information Desk Open

08:30 - 09:10 **Invited speakers**

Concertzaal **Developments in computerized adaptive testing in education**
Theo Eggen, *Cito / Twente University*

Computerized adaptive testing (CAT) has evolved from a psychometric tool for the efficient estimation of ability of persons to a testing mode that can serve test several purposes meeting practical conditions. This is especially the case in the field of education. For different goals of testing different algorithms have been developed and are applied in summative but also in formative settings. In the presentation a sketch will be given of the historical developments in available CAT algorithms and the attention of researchers to CAT will be reviewed. Furthermore, two examples of CAT development will be treated. Attention will be paid to item selection in an environment where learning is the main purpose of testing. Finally the approach of multi segment adaptive testing serving a diagnostic purpose of testing will be presented.

Parkzaal **Measurement and Control of Response Styles Using Anchoring Vignettes: A Model-Based Approach**
Daniel Bolt, *University of Wisconsin-Madison*

Anchoring vignettes provide an attractive tool for the measurement and control of response styles in psychological tests. In this presentation, a multidimensional item response theory (MIRT) model is proposed that uses information from vignette responses to control for response style effects in the measurement of a psychological construct. The model is presented as a generalization of a multidimensional nominal response modeling approach (Moors, 2003; Bolt & Johnson, 2009) that through anchoring vignettes can accommodate many additional forms of response style. The result is a framework in which to evaluate the extent to which variability in response style can be understood in relation to particular response style types. An illustration is provided using data from a cross-national study of self-reported conscientiousness by Mõttus et al (2012).

09:15 - 10:35 **Parallel session D**

Concertzaal **D.1.-Characterization of intra-individual dynamical processes**

A critique of the cross-lagged panel model
Ellen Hamaker, *Utrecht University*

The cross-lagged panel model is believed by many to overcome the problems associated with the use of cross-lagged correlations as a way to study causal influences in longitudinal panel data. In the current presentation it will be shown that if stability of constructs is to some extent time-invariant, the autoregressive relationships of the cross-lagged panel model fail to adequately account for this. As a result, the lagged parameters that are obtained with the cross-lagged panel model do not represent the actual within-person dynamics, and this may lead to erroneous conclusions regarding the presence, predominance, and direction of causal influences. Several numerical examples are presented to demonstrate the spurious results that may arise. A modeling strategy to avoid this pitfall is proposed and illustrated with empirical data. The implications for existing and future cross-lagged panel research are discussed.

Critical slowing down as an early warning indicator of transitions in mood
Ingrid van de Leemput, *Wageningen University*

Approximately 17% of all humans experience at least one episode of major depression. It is largely unknown what processes govern the transition from healthy emotional states to psychopathology and vice versa. Currently, we have no way of knowing whether such transitions on the continuum of depression are imminent. Here we show that the relevant transitions are preceded by rising autocorrelations, stronger correlations between emotions, and increased variance in mood fluctuations. These are all indicators of the general phenomenon of Critical Slowing Down (CSD), that occurs as a dynamical system approaches a tipping point. CSD has been shown for ecosystems that shift from a desired to a degraded state, such as lakes and the climate system. The underlying theory, however, applies to all dynamical systems that have internally reinforced tipping points. Our results provide support for the hypothesis that within-person changes from less to more depressed states and vice versa are separated by tipping points. Moreover, our results suggest a novel toolbox for assessing the likelihood that patients will fall into a more depressed state, as well as the chances of being tipped out of this state.

Estimation and Implications of Logistic Vector-autoregression Models
Sacha Epskamp, *University of Amsterdam*

Given intensive repeated measures in an subject over time, and the assumptions that the state of the measured variables only depend on the previous state of the measured variables, and that this process is stationary, vector-autoregression (VAR) can be used to estimate an underlying network of presumed causal relationships between the variables over time (Bringmann et al., in press).
Often the nodes are binary in that they can either be "on" or "off". In such cases, normal regression as used to estimate VAR models is insufficient. In such systems, the probability of a node activating can be modeled using logistic regression on the activation in the previous time point. Simulation under this model shows interesting properties such as phase transitions and common activation of multiple nodes in a cluster.
In this talk we will explain simple ways to estimate (logistic) VAR models using generalized linear models in R and how the estimated network can be visualized using the qgraph package in R. Extensions on these models such as multi-level designs on the network parameters will be discussed.

Studying inertia in a multilevel autoregressive model: Should the lagged predictor be cluster mean centered?
Raoul P. P. P. Grasman & Ellen L. Hamaker
University of Amsterdam

How to center level 1 predictors in a multilevel model has received considerable attention. If the within-cluster slope deviates from the between-cluster slope, and the main interest is in the within-cluster slope, the common advice is to use cluster mean centering. Grand mean centering leads to an estimate that is a blend of the two slopes, and is generally less informative. However, we show in a series of simulations that if one has a multilevel autoregressive model in which the predictor is formed by the lagged outcome variable (i.e., the outcome variable at the previous occasion), cluster mean centering will in general lead to a downward bias in the parameter estimate of the slope (i.e., the autoregressive relationship), while grand mean centering results in an unbiased estimate. This is particularly relevant if the actual interest is in (individual differences in) the autoregressive parameter. We illustrate the different approaches, and discuss when cluster mean centering should be preferred.

Parkzaal **D.2.-Equating-1**

Item response theory observed-score equating and local equating with the R package kequate

Björn Andersson, *Uppsala University*

In equating, scores on different versions of the same standardized test are related in order for the scores on the tests to be used interchangeably. The open source R package kequate enables observed-score equating using the kernel method for all common equating designs. The package is available as a free download at http://cran.r-project.org/package=kequate for usage by anyone interested in test equating. New features in the package include more elaborate methods of item response theory observed-score equating and the newly developed method of local equating. Local equating can be conducted in kequate either by conditioning on the anchor scores in a non-equivalent groups with anchor test design or by utilizing information from an item-response theory model with any of the common equating designs. This presentation describes these additions and suggests a way to conduct local equating in practice. To illustrate the new features of kequate, comparisons are made between the traditional observed-score equating methods and item response theory observed-score equating and local equating.

Optimal Prior Information to the Bayesian IRT Equating in Large-Scale Assessment with Matrix-sampled Anchor Items Design

Hyun-Woo Nam, *SoonChunHyang University*

This study was intended to find out the optimal priors for the Bayesian IRT (Item Response Theory) equating in the ‘common plus matrix-sampled anchor items design’.
For the study, the 10th grade students’ English test in the NAEA (National Assessment of Educational Achievement) was used. The item responses sampled by the ‘non-equivalent group anchor test design’ were revised to the ‘common plus matrix-sampled anchor items design’ with 8 blocks at most. Traditional IRT equating methods, CCT(Characteristic Curve Transformation) and FPIP(Fixed Pre-calibrated Item Parameter) in the context of MLE(Maximum Likelihood Estimation), were included to the Bayesian IRT equating as those of flat priors and point priors methods. In addition to those of completely uninformative priors and completely informative priors, immediate (just, high, and low) informative priors in the context of MCMC were introduced to this study.
The traditional IRT equating methods showed their results a little bit different from those of Bayesians in the context of MCMC. The Bayesians were more robust to the increasing number of blocks in the ‘common plus matrix-sampled anchor items design’ than those of traditional methods. Changing the precision of prior distribution with hyper priors could be an efficient way to improve the equating result in the matrix-sampled design.

A Bayesian approach to observed score equating

Ivailo Partchev & Gunter Maris
Cito Institute for Educational Measurement

Bayesian methods offer the advantage over alternative approaches to IRT based observed score equating that they allow for incorporating all sources of uncertainty in the equating standard error. Especially Markov chain - Monte Carlo methods are effective in this respect, if they show favorable operating characteristics (autocorrelation, and computational cost), for realistically complex models and designs. In this paper we present a new Metropolis-within-Gibbs sampler for the Nominal Response Model, and its various special cases, that performs well both in terms of autocorrelation and in terms of computational cost. We demonstrate with real data that with this new Bayesian approach we can effectively and efficiently deal with IRT based observed score equating, taking into account all sources of uncertainty, and taking into account the multilevel structure that is nearly always present and nearly always ignored in equating studies.

Observed-score kernel equating with covariates

Marie Wiberg & Kenny Bränberg
Umeå university

To equate two test forms we need to collect data in such a way that the link between the scales of the two test forms can be estimated. The traditional approach is to use common examinees and/or common items. In this paper we explore the idea of using variables correlated with the test scores (e.g., school grades and education) as a substitute for common items in a non-equivalent groups design. This is done in the framework of kernel equating, and with an extension of the method developed for post-stratification equating in the non-equivalent groups with anchor test design. Real data from a college admission test are used to illustrate the use of the method. The equated scores are compared with equated scores from an equating with no covariates and where the samples are treated as statistically equivalent. The results indicate that the method can improve the equating by adjusting for differences in test score distributions caused by differences in the distribution on covariates.

Jubileumzaal D.3.-Patient-reported outcomes

Metamorphosis of the Diagnostic Triangle

Abhijit Chatterjee, *Burdwan Homoeopathic Medical College & Hospital*

Diagnostic triangle is a standard equilateral triangle with the three corners based on human emotions - Intimacy, Passion and Commitment. It was evolved from Robert J. Sternberg’s “Triangular Theory of Love”. The objective of our study is to establish a cause-effect relationship of emotional behavior of the patient and to represent it in an easy, decipherable diagrammatic manner which gives a at–a-glance emotional picture and a standard protocol which is not affected by physician’s prejudices. In our study psychometric evaluation followed by homoeopathic case taking was done. The emotional feelings were then segregated under the three components and were graded in 1-9 scale. The changes in the triangle differ from the balanced triangle. This was followed by homoeopathic treatment and counseling. After 1-6 months of medication emotional expressions changed, this corroborated with variability of the triangular components. This study was done in 2006, at Fr.Muller Homoeopathic Medical College & hospital – Mangalore, India. Statistical analysis of 30 case studies showed metamorphosis of every triangle of the patient after Homoeopathic treatment. Thus it can be concluded that Diagnostic Triangle is an emerging tool for personality evaluation for proper diagnosis and prognosis.

Comparison of the Accuracy of Person Parameter Estimation Methods in IRT for Reliable Change Assessment

Ruslan Jabrayilov, *Tilburg University*

In clinical psychology, assessing the effectiveness of psychotherapies for individual patients is common practice. According to Jacobson and Truax (1991) an essential part of this assessment is a test of significance of individual change scores. For this purpose, Jacobson and Truax proposed the Reliable Change Index (RCI). In this study, we examined the RCI in the context of item response theory (IRT). Effective use of the RCI for a person v requires accurate estimates of his/her individual change (δ_v) and its standard error 〖(SE〗_(δ_v)). Using simulated data, we studied bias in estimated change scores δ_v and 〖SE〗_(δ_v) for three most commonly used person parameter estimation methods in IRT (i.e., ML, WML and EAP). Each of the methods may produce biased estimates to a certain degree, in particular when the number of items is small. This study focuses on the degree of bias and its impact on sensitivity and specificity with respect to reliable change detection. Data were simulated for polytomous items. We varied the following factors: the amount of change, item discrimination, spread of item difficulties, and test length. Results of the simulations are discussed.

On the Use of Change Scores in Routine Outcome Measurement

Wilco H. M. Emons, *Tilburg University*

The basic premise of Routine Outcome Measurement (ROM) is that if feedback about individual treatment outcomes is routinely provided to clients, clinicians, and stakeholders, further improvement of the quality of mental health care is likely. ROM has gained a strong foothold in the clinical practice. In the Netherlands, for example, many practitioners use ROM data to evaluate their clients' progress in the course of a treatment and branch organizations use ROM data for benchmarking purposes at the institutional level. A key feature of ROM is that it heavily relies on change scores, both at the individual level and at the aggregated institutional level. In this presentation we provide some new perspectives on important, yet neglected psychometric issues with respect to the use of change scores in the context of ROM. These issues include change-score reliability, measurement precision, and norming. We discuss possibilities and limitations. We also explain some common misconceptions and paradoxical results. To illustrate these issues, we use results from both empirical data analysis and simulation studies. Implications for future research are discussed.

Balkonzaal

D.4.-Test design and simulation

Automated Test Assembly with non-existing items

Angela Verschoor & Roel Visseren
Cito Institute for Educational Measurement

Traditionally, Automated Test Assembly (ATA) is applied in situations where a known item pool with calibrated items is available. If infeasibility occurs – the fact that no test that fulfills all test specifications can be assembled – models like the Weighted Deviation Model are the only tools to force the assembly of a test, by altering the test specifications. In many cases an unwanted side effect takes place: most specifications are formulated for well-defined reasons and the resulting test will often be regarded second choice.

In this paper we propose an ATA model with decision variables that do not correspond with existing items in the pool, but these indicate which type of items need to be developed in order to create a feasible test. A multi-objective approach is used to create a partial test with maximum information, while additional to-be-developed items are added against a penalty.

This model has been successfully employed in the development of the computer based high school examinations, where sets of up to 12 equivalent variants were assembled.

How general is the Vale-Maurelli simulation approach?

Njål Foldnes & Steffen Grønneberg
BI Norwegian Business School

Multivariate data simulation is a main tool in evaluating the finite-sample performance of estimation methods and goodness-of-fit measures in structural equation modeling. A popular and seemingly rather general data generation technique is the use of Fleishman polynomials applied to an underlying multivariate normal vector, as presented by Vale and Maurelli. Here, a Gaussian random vector Z is transformed to a non-Gaussian random vector Q through applying polynomials to each of its marginals. The Vale-Maurelli approach is implemented in popular software packages like Mplus, EQS, Lisrel and lavaan. In this presentation we identify the exact multivariate distribution of a generalization of the Vale-Maurelli transformation. The simulated variable is shown to have a copula that is closely linked to that of the underlying Gaussian variable. This means that the truly multivariate properties of the random vectors generated by Vale-Maurelli approach is close to the Gaussian case, even though it would seem that the resulting highly kurtotic random variables are very far away from the Gaussian case. We present a simple simulation study to numerically illustrate the limitations of the Vale-Maurelli approach.

Method for Choice of Group Number in Item Characteristic Chart on the Basis of Information Criterion

Takashi Akiyama, Hideki Toyoda & Norikazu Iwama
Waseda University

An item characteristic chart is an important tool for analyzing a test item. A composer of the test can detect the characteristics of each item, which are the item difficulty and whether the item reflected an essential concept of the test by inspection of the tool.
If an analyst tries to make an item characteristic chart, he or she has to divide examinees into arbitrary number of groups. However, a criterion, which the analyst decides the appropriate number of groups according to, is not established, and the analyst has to determine the group number arbitrarily, or settles five groups empirically. Therefore, it is useful for the analyst to decide the group number based on a statistical criterion when drawing the item characteristic chart.
In this presentation, we will propose a method for choice of the group number in item characteristic chart on the basis of statistical criterion, an information criterion, and an applicability and advantage of the method will be shown through a simulation and examples of applications of factual test data.

Graphical Extension of Sample Size Planning with AIPE on RMSEA

Tzu-Yao Lin & Li-Jen Weng
National Taiwan University

Sample size planning is a commonly encountered issue in structural equation modeling applications. Rules of thumbs from various perspectives have been suggested. The present study proposed graphical extension with accuracy in parameter estimation (AIPE) abbreviated as GAIPE, on RMSEA to incorporate accuracy in estimation, inferences of model fit, and power consideration for sample size determination. GAIPE simultaneously displays the expected width of a confidence interval of RMSEA, the necessary sample size to reach the desired width, and the RMSEA values covered in the confidence interval. The power values associated with hypothesis tests based on RMSEA can also be integrated into the GAIPE framework. With the capacity of incorporating information on accuracy in RMSEA estimation, values of RMSEA, and power on hypothesis testing on RMSEA in a single graphical representation, the GAIPE extension offers an informative and practical approach for sample size planning in structural equation modeling. The proposed method can be implemented by the GAIPE package in CRAN.

Stadszaal **D.5.-Differential Item Functioning -2**

Testing Measurement Equivalence of Categorical Items' Parameters: A Comparison of SEM and (M)IRT Approaches

Hongyun Liu, Fang Luo & Yan Dong
Beijing Normal University

Two type methods of assessing measurement equivalence of categorical items, namely, multiple group analysis based on structural equation modeling framework, and differential item functioning based on (unidimensional or multidimensional) item response theory framework, were the most popular methods. Unlike the traditional

linear factor analysis, multiple-group categorical confirmatory factor analysis (CCFA) can appropriately model the categorical measures with a threshold structure, which is comparable to difficulty parameters in (M)IRT. In this study, we compared the multiple-group categorical CFA (CCFA) and (M)IRT in terms of their power to detect violations of measurement invariance (also known as DIF) through a Monte Carlo studies. Moreover, noticing that the hypothesis of unidimensional for traditional IRT model, this study extended the DIF test method base on IRT model to MIRT. Simulation study under both unidimensional and multidimensional conditions was conducted for the comparison of DIFTEST method, IRT-LR method (for unidimensional scale), and MIRT-MG (for multidimensional scale) with respect to the power to detect the lack of in-variance across groups.
The results indicated that three methods of DIFFTEST, IRT-LR, and MIRT-MG showed reasonable power to identify the measurement non-equivalence when the difference of threshold was large. For the DIFFTEST method, the Type I errors reached the nominal error rate at 5%, in comparison, IRT-LR and MIRT-MG produced much lower Type I error rates.

DIF for multilevel data: A combined approach of multilevel IRT models and multilevel mixture factor models

Yao Wen & Cindy M. Walker
University of Wisconsin Milwaukee

In educational research, multilevel data structure is common so multilevel models have been studied greatly. However, measurement invariance, or differential item functioning (DIF), has not been fully explored in the context of multilevel models. In particular, since school level variables may account for some of the measurement invariance, using a multilevel model that nests items within students within schools may provide a way to test for DIF that also helps to explain the sources of that invariance. Moreover, since it is widely assumed that DIF causes ability estimation bias, using a multilevel model explicitly models the DIF may provide more robust ability estimates for examinees. Previous studies showed the ability estimates were biased when DIF occurred at school level using the 2PL model. This study will evaluate the ability estimates when DIF occurs at student level, school level, or both levels using multilevel IRT models (ML IRT). A simulation study will be conducted to evaluate estimates of the ability using the 2PL model, the unconditional ML IRT, and ML IRT incorporating DIF. This study is important in that there is a lack of attention to study multilevel DIF completely as well as its influence on ability estimates.

DIF Detection Using Multiple-Group Categorical CFA with Free Baseline Approach

Yu-Wei Chang, *National Tsing-Hua University*
Rung-Ching Tsai, *National Taiwan Normal University*
Way-Gong Huang, *National Taiwan Normal University*

The aim of this study is to assess the efficiency of using multiple group categorical confirmatory factor analysis (MCCFA) and robust chi-square difference test in differential item functioning (DIF) detection for polytomous items under the free baseline strategy. While testing for DIF items, despite of the strong assumption that all but the examined items are set to be DIF-free, MCCFA with such a constrained baseline approach is commonly used in the literature. The present study relaxes this strong assumption and adopts the free baseline approach where item parameters are allowed to differ among groups. Model identification is an issue in the free baseline approach, and we propose a set of identification constraints and have its just-identifiability checked. The robust chi-square difference test statistic with the mean and variance adjustment is utilized as our test statistic. Based on the simulation results, the test statistic is shown to be efficient in detecting DIF for polytomous items in terms of the empirical type I error and power. To sum up, MCCFA under the free baseline strategy is useful for DIF detection for polytomous items.

A Power Formula for the SIBTEST procedure for Differential Item Functioning
Zhushan Li, *Boston College*

A power formula for the SIBTEST procedure for differential item functioning (DIF) is derived. As shown by the derived formula, the power of SIBTEST is related to the item response function (IRF) for the studied item, the latent trait distributions and the sample sizes for the reference and focal groups, and the proportion of each group in the sample. A detailed discussion of the SIBTEST precedure under some practical models such as the 2PL and 3PL item response theory (IRT) models is given. Based on the power formula, a method for calculating the sample size for a desired power level is presented. Monte Carlo simulation studies show that the theoretical values calculated from the formulas derived in the paper are close to what are observed in the simulated data.

Hemelzaal **D.6.-Modeling**

Bivariate and Joint Distance Models for Bivariate Binary Data
Hailemichael Metiku Worku, *Leiden University*

Bivariate binary data are often collected in the social and medical sciences. Statistical models that can
deal with such data and that can answer the important research questions raised in a study, are often needed. The Netherlands Study of Depression and Anxiety (NESDA) data is used as a working example in order to estimate prevalence, comorbidity and association of mental disorders and their relationship with personality traits. Two types of probabilistic multidimensional unfolding models are proposed to deal with such bivariate binary data: a joint probability model and a marginal model. Simulation studies are conducted to evaluate the performance of these models. The application of these models depend on the scientific question we would like to answer. The joint model works fine if the objective of the study is about modeling comorbidity and its relationship with personality traits. We recommend researchers to use the marginal model if (s)he is interested in the prevalence and association of mental disorders. Graphical displays (Biplots) can be obtained from such models to improve interpretation.

Unbiased effect estimation in the presence of publication bias: a new method for meta-analysis and assessing publication bias
Marcel van Assen & Robbie van Aert
Tilburg University

One substantial problem in meta-analysis is publication bias. The problem results in overestimation of the effect size in traditional meta-analysis, particularly in small studies that are ubiquitous in psychology and when population effect size is small. We developed a new meta-analysis method that (i) detects publication bias, and (ii) provides an accurate estimate of effect size whenever there is publication bias. The proposed method is straightforward to implement since it only uses regular study characteristics used in any meta-analysis. We demonstrate the superior performance of the new method compared to traditional meta-analysis methods to detect publication bias and to estimate population effect size.

10:35 - 10:55 ***Break***
Ravelijn

10:55 - 12:15 Parallel session E

Concertzaal **E.1.-Invited symposium: Computerized Multistage Testing: Theory and Applications**

Overview of Multistage Testing: History and Future Directions.

Chun Wang, *University of Minnesota*
Hua-hua Chang, *University of Illinois*
Yi Zheng, *University of Illinois*

Multistage testing (MST) has become increasingly popular in recent years. An MST is a specific assessment design that allows for adaptation of the difficulty of a test to the level of proficiency of a test taker. In this presentation, we will first give a general overview of multistage tests (MST) and briefly discusses the important concepts associated with MSTs. The connections and distinctions between MST and other tests, including linear tests and computerized adaptive tests (CAT), will be deliberated. While MST has many potential benefits, it also generates new challenges for (1) test assembly due to the large number of possible paths through the test; and (2) test security because it is often administered continuously in a testing window. To address the former challenge, we will then present a new paradigm for MST assembly called "assembly-on-the-fly," which borrows well-established item selection algorithms in computerized adaptive testing (CAT) to construct individualized modules for each examinee dynamically. To address the latter challenge, we will introduce a new test security index, namely, the variance of the test overlap rate, to help evaluate the severity of security breach in MST.

Automated Assembly of Multistage Testing Systems.

Wim van der Linden & Qi Diao
CTB/McGraw-Hill

The shadow-test approach to automated test assembly is based on a reconceptualization of the test-assembly processes as the selection of items from a sequence of full-length fixed test forms ("shadow tests"). The approach can be used to automatically generate test forms with any existing format (e.g., fixed form, on-the-fly linear, fully adaptive, and multistage testing) or combinations of elements of these formats (e.g., multistage testing with an adaptive routing test). In addition, the use of a shadow-test assembler allows us to directly evaluate the relative efficiency of different testing formats against each other for any given the set of constraints on their content. In the current presentation, we will demonstrate the use of a shadow-test assembler for a variety of multistage testing formats. In addition, we will present results from an empirical study as to the statistical efficiency of these formats.

Routing and Scoring in Multistage Testing.

Duanli Yan & Charles Lewis
Educational Testing Service

MST routing is the process that routes or classifies test takers to different paths or next stage modules based on their performance on the previous module(s) using selected rules, which can be quite different depending on the purpose and design of the MST. Many methods are considered in MST routing, including IRT-based and CTT-based MST routing algorithms, as well as the approaches for classification, mastery testing, diagnostic testing, and approaches for content optimizations in practice.

There are also many ways in which tests are scored. Again, based on the different purposes and designs of MST, these include IRT-based using maximum likelihood and Bayesian methodologies, CTT-based MST approaches using regression trees, the approaches for classification testing, other models using multi-dimensional IRT,

and models for diagnostic testing.
This presentation illustrates the MST routing and scoring with specific examples.

Reliability for Multistage Testing.
Peter van Rijn, *Educational Testing Service*

Reliability and measurement error are important classical psychometric aspects of educational and psychological tests. In this presentation, we describe the concepts of test reliability and errors of measurement in the context of multistage testing and item response theory (IRT).
In a multistage test, test takers are administered test forms with different items of different difficulty, and test takers' responses are used to determine which level of difficulty they should receive. Therefore, the classical notion of reliability as an indicator of measurement precision of a fixed test to be administered to all test-takers seems difficult to retain for the cases of multistage and adaptive testing because individuals are no longer administered the same set of items. Nevertheless, test reliability can still be useful for these cases, and we will demonstrate how to estimate appropriate reliability measures by making use of IRT methodology. In the first part of this presentation, we describe test reliability and how to compute it in the context of IRT. The second part describes an application of the methods to data from a multistage test pilot study of the National Assessment of Educational Progress (NAEP) program.

An application of Multistage Testing
Angela Verschoor, *Cito*
Theo Eggen, *Cito / Twente University*

When designing an MST, it is frequently the case that we need to: 1) define a test information function target for each module in the MST for automated test assembly, and 2) find a suitable routing through these modules. The optimality of a routing, optimality of module assembly, and if there is a separation of optimizing module assembly and routing are all important concerns for an effective application of MST.

In this paper, a model is presented that simultaneously optimizes the entire MST, both assembly of the modules and the routing. It makes use of simulations that were originally developed to evaluate Computerized Adaptive Tests (CAT). Two different item pools are investigated: an artificial Rasch-calibrated item pool of infinite size, and an item pool developed for an operational Arithmetic Test.

In general, approximately 5% lower RMSEs for the candidates' ability estimations can be obtained, compared to procedures and settings that up to now were regarded as best practice.

Parkzaal **E.2.-Equating-2**

Results of Equating Polish National Examinations Study
Bartosz Kondratek, Henryk Szaleniec, Magdalena Grudniewska, Filip Kulon & Artur Pokropek
Educational Research Institute

The article presents the methodology and results of a research that equated 2002–2011 editions of lower secondary school examinations in Poland. Equating allows distinguishing the variation in examination difficulty from the systematic changes of population ability level. The equating procedure was conducted with a support of data gathered from a random sample of around 10 500 pupils that provided linking information for over 500 items from previous examinations. Multiple-group Item Response Theory model was fit to the data by MIRT software (Glas, 2010). The equating results are presented on the latent variable scale as well as on the observed score scale. The analysis concludes that the ability level for the humanities part of the examination was stable over years, however the ability level for the mathematics with science part exhibits decline throughout the examined time span. The results were validated by comparison to the equated scores of PISA scale for Poland. The equated results are roughly in compliance with PISA.

Simultaneous Equating of Separately Calibrated Polytomous IRT Test Items

Sayaka Arai, *National Center for University Entrance Examinations*
Tomohiro Ohtani, *Tokyo institute of technology*
Shin-ichi Mayekawa, *Tokyo institute of technology*

Item response theory (IRT) models provide many advantages over classical test theory. One of the useful features of IRT is the comparability of test scores obtained from different test forms. To accomplish this, it is necessary for the item parameters of all the test items to be put onto a common scale by so called equating.
Several equating methods have been proposed. A factor analytic method for dichotomous items is one of them. Since it can equate all items included in multiple test forms simultaneously, it is thought to have the advantage in a situation where the set of items are spread in many forms. Comparison studies of equating using simulated data showed that the factor analytic method performed well.
In this study, we extend our factor analytic equating method to the polytomous items including the graded response model and the generalized partial credit model items. The performance of the proposed method will be compared to the existing method using both artificial and real data sets.

Observed-score Equating Methods for Mixed-Format Tests

Won-Chan Lee, *University of Iowa*
Jaime Peterson, *University of Iowa*
Guemin Lee, *Yonsei University*

Equating is an integral part of a testing program when multiple forms of the test are constructed and administered. Comparability and fairness of scores reported to examinees who take different test forms cannot be warranted if equating results are inaccurate. When test forms are composed of items with different format such as multiple-choice (MC) and constructed-response (CR) items, the data will likely be multidimensional due to the item format effect. In such a case, fitting a unidimensional item response theory (IRT) model to multidimensional data would produce inaccurate equating results. We consider two alternative multidimensional IRT frameworks that are considered more appropriate for equating mixed format tests: simple-structure and bi-factor models. These two approaches are compared based on various real data showing different levels of latent correlation between the MC and CR sections.

Simultaneous Equating of Separately Calibrated Multidimensional IRT Test Items

Yoshinori Oki & Shinichi Mayekawa
Tokyo Institute of Technology

In recent years, multidimensional IRT (mIRT) models became popular and seems to be ready for some serious applications. However, for the multidimensional IRT models to be used in the similar way as the unidimensional IRT models, it is necessary to construct item banks, and in order to do so, we must have a method to equate item parameters. In this research, we propose a method to equate separately calibrated item parameters to a common scale simultaneously. The method is based on a variation of so called generalized orthogonal Procrustes rotation in which a set of dimensional weights for each target matrix is employed. As far as we know, all of the available equating methods for mIRT items are pairwise methods in which a chain of equating steps is necessary to equate all the items spread in several forms. On the contrary, our method avoids the accumulation of errors associated with the pairwise

equating methods by directly equating the scales for each test form to a common scale. Therefore, it is expected that this method is preferred to other existing methods. Several simulation studies will be presented to show the performances of our method in comparison with the existing methods.

Jubileumzaal **E.3.-Modeling and estimation of diagnostic models**

Statistical Refinement of the Q-Matrix in Cognitive Diagnosis

Chia-Yi Chiu, *Rutgers, The State University of New Jersey*

Methodologies for cognitive diagnosis have received much attention in recent years. A common connection among these methodologies is the inclusion of a pre-determined Qmatrix that relates items to latent attributes. However, the subjective nature of establishing the Q-matrix for a given test has raised serious validity concerns among researchers. Although the consequences of misspecified Q-matrices have been widely recognized, at present, limited well-established research is available to detect Q-matrix misspecification. In this study, we propose a non-parametric Q-matrix refinement method based on the residual sum of squares (RSS) with the non-parametric classification method (Chiu & Douglas, in press) embedded in the algorithm. A theoretical justification to support the idea that the correct q-vector of an item is identified when the corresponding RSS is minimal is presented. Simulation studies are conducted to evaluate the performance and to examine the robustness of the method. The results show that the Qmatrix refinement method is efficient and can effectively identify and correct the misspecified q-vectors with great accuracy. We also apply the method to real data and demonstrate how the analysis should be done by taking into account the outcome of the statistical procedures and the consultation from domain expertise.

A General Proof of the Consistency of Non-Parametric Classification for Cognitive Diagnosis Models

Hans-Friedrich Koehn, *University of Illinois at Urbana-Champaign*
Chia-Yi Chiu, *Rutgers University*

Estimation of the parameters of cognitive diagnosis models and assignment of examinees to proficiency classes typically relies on maximum-likelihood procedures. Non-parametric classification techniques that do not incorporate fully specified parametric models have been proposed as approximate methods for assigning examinees to proficiency classes. The Asymptotic Classification Theory of Cognitive Diagnosis (ACTCD) by Chiu, Douglas, and Li provided a theoretical foundation for this approach. The ACTCD has been proven for the DINA model. For other cognitive diagnosis models with different statistical properties, individual proofs for each model are required that coverage by the ACTCD is guaranteed a tedious undertaking in light of the numerous cognitive diagnosis models that have been proposed in the literature. A general proof of the ACTCD is presented that relies on the framework of general cognitive diagnosis models.

Extensions of the Saltus model

Karen Draney & Minjeong Jeon
UC Berkeley

The Saltus model (Wilson, 1989) has been developed to address developing proficiency that occurs in Piagetian stages, where movement to a new stage involves the acquisition of a new rule. The original Saltus model assumes a Rasch model for a person at a given stage, but these models differ by shift parameters that depend on stage membership and its effect on items in each item group.
We consider an extension of the Saltus model by incorporating task structures that may be differentially difficult for different stages. In this extension, a linear logistic test model (LLTM; Fischer, 1983) is employed at a given person stage, in which task

difficulties are modeled instead of items. We also consider a multidimensional version of the Saltus LLTM. The proposed models are illustrated using the competence profile test of deductive reasoning –verbal (DRV; Spiel, Gluck, & Gossler, 2001, 2004) that was developed to assess deductive reasoning during the transition from the concrete operational to the formal-operational stages.

A hierarchical item response model for cognitive diagnosis

Mark Hansen & Li Cai
University of California, Los Angeles

Cognitive diagnosis models have received increasing attention due to their potential to provide instructionally or clinically relevant information concerning both respondents and test items. However, the validity of such information may be undermined when these models are misspecified. This study focuses on one aspect of model misspecification: violations of the local item independence assumption. We examine potential causes and consequences of such dependence, with particular attention to those causes that are unrelated to the attributes a diagnostic test is intended to measure. Ignoring such nuisance dependencies can lead to biased estimates of model parameters and misclassification of examinees. We propose a hierarchical diagnosis model in order to address this problem. The proposed model includes random effects to account for dependencies, thus following a strategy already well-established in item factor analysis (serving as the basis for testlet, random intercept, bifactor, and two-tier factor models, among others). The hierarchical model maintains desirable properties of existing diagnosis models while allowing for greater complexity in the underlying response process. Importantly, the model may be estimated efficiently through analytical dimension reduction—even for large numbers of nuisance latent variables. Through simulation study and empirical applications, we examine the utility and limitations of the proposed model.

Balkonzaal **E.4.-SEM for LDA**

On the definition of latent variables: Presentation of a general approach to defining latent effect variables with applications to popular structural equation models

Axel Mayer, *Friedrich Schiller University Jena*

Latent variables are not directly observable, but this does not mean that they cannot be defined. Latent variables are sometimes viewed as diffuse concepts whose meaning depends on imagination. In contrast, this talk follows a research tradition in which latent variables are defined in terms of probability theory. In this research tradition, the random experiment is made explicit and latent variables are defined as random variables that refer to the random experiment considered.
In this talk, I present a general approach to defining latent effect variables based on true score variables or latent state factors. In this approach, contrasts are used to define new latent variables as functions of true score variables or latent state factors. Then, the structural equation model to estimate these latent variables is derived.
The approach can be applied to popular structural equation models. In particular, it can be applied to define latent effect variables in true change models and growth curve models, as well as latent effect variables in models with method effects. In addition, it can accommodate more complex models as will be shown in an application to depression trajectories in cancer patients.

The Cusp Catastrophe Model as a Regime-Switching Mixture Structural Equation Model

Sy-Miin Chow, *Pennsylvania State University*
Katie Witkiewitz, *University of New Mexico*
Raoul Grasman, *University of Amsterdam*
Stephen A. Maisto, *Syracuse University*

Catastrophe theory is the study of the many ways in which continuous changes in a system's parameters can result in discontinuous changes in one or several outcome variables of interest. Catastrophe theory-inspired models have been used to represent a variety of change phenomena in the realm of social and behavioral sciences. Although promising in its own right, the cusp catastrophe model and current approaches of fitting it do not address several practical data analytic problems, such as the presence of incomplete data and categorical indicators, difficulties in performing model comparison, as well as heterogeneous timing of shifts within and across subjects. To account for these data analytic issues, we propose a mixture structural equation model with regime-switching (MSEM-RS) as an alternative way to represent features of the cusp catastrophe model. The proposed model is illustrated using longitudinal drinking data from the MATCH project (Project MATCH Research Group, 1997) and a simulation study.

Outlying observation diagnostics in growth curve modeling

Xin Tong, *University of Notre Dame*

Growth curve models are useful for investigating growth and change phenomena in social, behavioral, and educational sciences and are one of the fundamental tools for dealing with longitudinal data as well as repeated measures. Many studies with real data have demonstrated that data without any outliers are rather an exception, especially with data collected longitudinally. Estimating a model without considering the existence of outliers may lead to inefficient or even incorrect parameter estimates. Therefore, outlier diagnostics become very important in exploratory data analysis for growth curve modeling. The purpose of this study is to compare the performance of five methods in outlier diagnostics through a Monte Carlo simulation study on a linear growth curve model, by varying factors of sample size, number of measurement occasions, number of outliers, and geometry of outliers. Simulation results show that the univariate outlier detection method based on individual growth curve analysis may provide the highest correctly identified rate, while the method proposed by Yuan and Zhang (2012) for model and data diagnostics in structural equation modeling may provide the lowest wrongly identified rate. A real data analysis example is also provided to illustrate the application of the five methods of outlier diagnostics.

Model Selection Criteria Using Bayesian Methods

Zhenqiu (Laura) Lu, *University of Georgia*
Zhiyong Zhang, *University of Notre Dame*

Research in applied areas often involves the selection of the best available model from among a candidate set. As researchers realize the need to account for an increasing number of latent and manifest variables to explain phenomena in applied areas, they have to investigate a large set of candidate growth models, and thus the effective selection of the best model becomes increasingly important. There are many traditional criteria available to the researcher. But this research examines the performance of a series of new proposed model selection criteria.

Given the fact that growth models often combine with the almost unavoidable outliers and attrition, competing models may reflect different missing data mechanisms or data distributions. The estimates from mis-specified models may result in severely misleading conclusions. We focus on complex latent growth models with missing data and outliers. Also, as the Bayesian approach provides many advantages in dealing with complex statistical models with complicated data structure, we use the Bayesian methods. Five simulation studies are conducted. Simulation results show that almost all the criteria can effectively identify the true models. A real data set is illustrated, and related implications of the criteria are discussed.

Stadszaal **E.5.-Modeling growth and its consequences**

Estimating Measurement Error Variance for Student Growth Percentiles using Binomial Assumptions

Jinah Choi, Won-Chan Lee, Stephen B. Dunbar & Catherine J. Welch
University of Iowa

Student growth percentiles (SGPs) have been developed to address the need to demonstrate student growth for educational accountability purposes (Betebenner, 2008). SGPs provide normative growth interpretations by using quantile regression conditional on past test scores, and are used as a growth index in several states in the United States. However, there is little research on SGPs' property related to measurement error and reliability. The primary goal of this research is to present a procedure for estimating conditional (i.e., individual-level) standard errors of measurement (SEMs) for SGPs based on binomial error assumptions (Lord, 1955, 1957; Brennan & Lee, 1999). An estimated conditional SEM can be used to construct a confidence interval for an individual having a reported SGP. Reliability for SGPs can also be computed using estimated conditional SEMs. Real data from multiple grade levels are used to examine the applicability of this procedure. A detailed description of the step-by-step estimation process is also provided in the paper.

Evaluating Teachers' Long-term Effect with value-added models in R

Liping Sun, *Beijing Normal University*

Value-added models have been widely used in evaluating teacher effect, which isolate teachers' contribution from other factors on students. The early models, such as gain score model and covariate adjustment model, focus on teacher effect of the current-year. While with the available of longitudinal data, some models involve in teachers' long-term effect, such as the TVAAS layer model and cross-classified model. To evaluating the persistence of teacher effect, the estimation method shows some new development. Compared to the traditional maximum likelihood estimation, the Bayesian estimation, which has primer information assumption of the parameters and uses Markov chain Monte Carlo algorithms to get approximations to the posterior distributions, has less restriction on the parameters. Hence the Bayesian approach will be helpful to estimate teachers' long-term effect more freely and convenient. Recently, as a free statistical software, R has been applied in many domains. In this article, we will illustrate how to use R software to model teachers' long-term effect on students and to estimate the persistent effect in Bayesian framework.

A longitudinal item response model for differential growth based on initial status

Ronli Diakow, *University of California*

The question of whether an educational program supports learning for all students, and in particular for students with different levels of prior knowledge, is important for thinking about the consequences of educational programs. In the psychometric literature, this question has been called the relationship between initial status and rate of change (Rogosa & Willett, 1985; Willett, 1988; Khoo, 2001; Seltzer, Choi & Thum, 2003). Statistical models for change over time and the relationship between initial status and change were developed in parallel in three modeling traditions: hierarchical linear modeling, structural equation modeling, and item response modeling. In this paper, I present a longitudinal item response model for differential growth based on initial status that incorporates elements from these three traditions. I discuss the identification of the general model and show how restrictions can be used to account for different conceptualizations of prior knowledge. I illustrate the use and interpretation of the model in the context of a new mathematics lesson sequence for integers and fractions in late elementary school. I found that while prior knowledge as

defined by a general intercept term did not predict growth, prior knowledge as defined by over (or under) performance on the pretest did.

Hemelzaal **E.6.-IRT estimation 1**

Equivalence of weighted likelihood and Jeffreys modal estimation of proficiency under polytomous item response models

David Magis, *University of Liège*

This talk focuses on two proficiency level estimators in item response theory (IRT) framework: the weighted likelihood estimator (WLE) and the Jeffreys modal estimator (JME), that is, the usual Bayes modal estimator with Jeffreys' non-informative prior. With dichotomously scored items, the WLE and the JME are completely equivalent under the two-parameter logistic model, while remarkable relationships were established under the three-parameter logistic model. The purpose of this talk is to extend such comparison to polytomously scored items. It is shown that both WLE and JME are also equivalent for two broad classes of polytomous IRT models, including, among others, the (modified) graded response model, the (generalized) partial credit model, the rating scale model and the nominal response model. Parallelisms with dichotomously scored items are drawn. An example from a real data set is used to illustrate this finding.

Application of Adaptive Testing to the Ekman 60 Faces Test

Luning Sun, *University of Cambridge*
Andrew Bateman, *The Oliver Zangwill Centre for Neuropsychological Rehabilitation*
John Rust, *University of Cambridge*

There exist a number of neuropsychological assessment tools, which aim to detect emotion recognition impairment. For instance, the Ekman 60 Faces Test makes use of the Ekman and Friesen Series of Pictures of Facial Affect and asks the participants to identify emotions from facial expressions.

The objective of this study was to explore the potential of applying adaptive testing techniques to individual emotion scales of the Ekman 60 Faces Test. A range of approaches to choosing the starting item were discussed, different criteria to select the next item compared, and various stopping rules investigated. Possible exposure control strategy was also proposed.

194 persons with brain injury and 194 normal controls were administered the Ekman 60 Faces Test. Using half of the sample, item parameters were estimated for each scale based on the Rasch Model, and the results applied to the other half to simulate adaptive testing.

The present study reveals huge potential of applying adaptive testing techniques to the Ekman 60 Faces Test. We are confident that adaptive testing will significantly improve neuropsychological measurement.

The use of Jeffreys prior in IRT estimation of latent traits

Megan Kuhfeld, *University of California*
Li Cai, *University of California, Los Angeles*

The purpose of this study is to investigate the use of the Jeffreys' prior as an alternative to the commonly-used standard normal prior distribution in the estimation of latent ability or trait scores using item response theory (IRT) Bayesian scoring methods. The Jeffreys' prior only requires specification of the item response model and the item parameter values, and therefore is seen as a non-informative prior. For this project, a simulation study was conducted to compare the performance of the standard normal prior and Jeffreys' prior applied to both EAP and MAP estimation methods using dichotomous items and polytomous response items under conditions of normal and non-normal latent generating distributions. Within each condition, plots of empirical coverage rates of true scores and root mean standard error values (RMSE) across range of theta were used to assess differences between the two priors. Preliminary results indicate that there is not a large difference between the normal and Jeffreys' prior in terms of empirical coverage of true scores, but that extreme positive and negative theta values are less likely to be recovered with the normal prior. However, standard errors for theta estimates are larger with the Jeffreys' prior in the extreme theta ranges.

Robustness of MCMC Estimation to Latent Nonnormality in the Mixture IRT Models

Sedat Sen, Allan S. Cohen & Seock-Ho Kim
University of Georgia

Unidimensional item response theory (IRT) models assume that a single model applies to all people in the population. ML estimates of item and person parameters are typically estimated assuming a normal ability distribution; however researchers commonly use the ML method without checking the distributional assumption. Mixture IRT models can be useful when subpopulations are suspected, but that differ along some unmodeled latent variable. The usual mixture IRT model is also typically estimated assuming within class normality. Research on normal finite mixture models suggests that latent classes potentially can be extracted even in the absence of population heterogeneity if the distribution of the data is nonnormal. Empirical evidence suggests, in fact, that test data may not always be normal. In this study, we will examine the sensitivity of mixture IRT models to violations of the normality assumption. Single class IRT data will be generated using different ability distributions and then analyzed with mixture Rasch and mixture 3PL models to determine the impact of these distributions on the extraction of latent classes.

12:20 - 12:45 State of the art lecture

Concertzaal **Classification Decisions in Testing**

Ying Cheng, *University of Notre Dame*

Most testing programs offer scores that will ultimately be used to make classification decisions, e.g., pass/fail, beginning level /intermediate/proficient, etc. A considerable amount of research has been devoted to improving the measurement precision of tests, with the hope that more accurate and precise characterization of the latent trait will lead to more accurate decisions. In this talk, I would like to analyze the factors that influence the accuracy of classification decisions, including the decision rules, underlying item response model, model fit, the number of cut scores and their locations, latent trait density, as well as the interactions among these factors. With simulation studies and real data examples we will illustrate that measurement precision is only a very small piece of the iceberg. The other factors are likely more influential on classification accuracy. Implications on related measurement issues such as standard setting and ROC analysis will be discussed.

Parkzaal **The Self-learning Q-matrix -- Theory and Applications**

Jingchen Liu, *Columbia University*

The Q-matrix, an incidence matrix specifying the item-attribute relationship, is a key element in the specification for many cognitive diagnostic models. It is common practice for the Q-matrices to be specified by experts when items are written, rather than through data-driven calibration. Such a non-empirical approach may lead to misspecification of Q-matrices and substantial lack of model fit, resulting in erroneous interpretation of results. In this talk, we present our recent findings concerning the data-driven construction (estimation) of Q-matrices. Upon writing the model in a regression form, we formulate the Q-matrix estimation to a model selection problem, for which regularized regression estimators are employed. The computation of such estimators is through a combination of the EM algorithm and existing convex optimization methods.

12:45 - 17:45 **Social event: National Park, Hoge Veluwe and Kröller Müller**

19:00 **Conference diner + Lifetime achievement award & travel awards**

Concertzaal **Psychometric Society Travel Awards**

Sedat Sen, *University of Georgia*
Sensitivity of Mixture IRT Models to Distributional Assumptions.

Jinah Choi, *University of Iowa*
Estimating Measurement Error Variance for Student Growth Percentiles Under Binomial Assumptions

Daniel Morillo, *Universidad Autonoma de Madrid*
Theoretical and Simulation Study of a Dominance IRT Model for Forced-Choice Items

ETS Travel Award Winner

Psychometric Society Lifetime Achievement Award Winner

Thursday July 25

08:30 – 18:00 Registration and Information Desk Open

08:30 - 09:10 Dissertation award talk

Concertzaal **Testing distributional assumptions in psychometric measurement model with substantive applications in psychology**

Dylan Molenaar, *University of Amsterdam*

Central to traditional latent variable models, like the linear factor model and the graded response model, is the linear regression of a continuously distributed dependent variable on a continuously distributed latent trait. The dependent variable can represent the observed continuous data (e.g., responses to a line segment) or a hypothetical continuous variable underlying the observed categorical item scores. Commonly, the dependent variable is assumed to be normally distributed, implying 1) homoscedastic and normally distributed residuals, 2) a normally distributed latent trait, and 3) trait-level independent slopes. In this talk, statistical models are presented that enable tests on these specific assumptions. In addition, it is illustrated how such tests are interesting from a substantive point of view, e.g., in the field of intelligence, personality, and behavior genetic research.

09:15 - 10:35 Parallel session F

Concertzaal **F.1.-Applied Psychometric Methods and Techniques: Achievements and Challenges**

Estimation, Calibration, and Scoring in Multistage Adaptive Tests

Alina A. von Davier, *ETS*

Estimation and calibration of model parameters and scoring of item sets and tests are statistical processes required for any practical implementation of multistage testing (MST). Although the details of the analysis depend on the specific MST design, the basic principles are the same. Item response theory (IRT) has traditionally provided the framework for these processes. Within the context of estimation, calibration, and scoring in MST, this presentation considers such issues as the local independence assumption, the estimation method (MLE, EAP, MAP), and the scoring method (IRT based and sum scores). The discussion here is confined to a one-dimensional IRT model and to standardized assessments where the scores are reported at an individual level. Calibration involves an initial phase for data collection in which conventional test administrations are used to build modules and routing rules to begin use of MST administrations. Once MST administrations begin, data are collected to establish scoring rules, develop new test modules, equate the cut-scores for routing, and ensure comparability of tests over time. Discussion involves both direct use of estimated examinee proficiency in routing and scoring and use of sum scores for these purposes.

Optimal Sampling Design for IRT Linking with Bimodal Data

Jiahe Qian, *ETS*

Optimal sampling designs for an IRT linking with improved efficiency are often sought in analyzing assessment data. In practice, the skill distribution of an assessment sample may be bimodal, and this warrants special consideration when trying to create these designs. This study explores optimal sampling design for IRT linking of bimodal data. Our design paradigm is modeled after the work of Berger (1991, 1997) and presents a formal setup for optimal IRT linking. In optimal sampling design, the sample structure of bimodal data is treated as being drawn from a stratified population; the stratum weights will be adjusted to achieve an optimal linking according to a criterion based on a function of the information matrix. The initial focus of the current study is the mean-mean transformation method, though the model of IRT linking under consideration is adaptable to generic methods.

Implementing Sampling Optimization for Automated Scoring Model Calibration
Mo Zhang, *ETS*

Because automated essay scoring is a relatively new field, many measurement issues from both theoretical and practical perspectives have not been adequately examined, one of which is the impact of sampling. In this study, we investigate whether sampling optimization for model calibration can improve the quality of resulting automated scores. Three types of scoring models are compared: one unweighted, one weighted to match the target population, and one statistically optimized. Jackknife replicate resampling approach is used to obtain the variance of several estimators, including the mean squared error of automated scores and agreement between the human ratings and automated scores. To ensure fairness, we also study the impact of optimized sampling on individual population groups with different educational, linguistic, and cultural backgrounds.

A Dependent Bayesian Nonparametric Model for Test Equating
Jorge González, *Pontificia Universidad Católica de Chile*

Equipercentile equating methods are based on the premise that scores on two different tests are equated if the respective cumulative distribution functions of the scores are equal, leading to the traditional equipercentile equating function. The estimation of the equating function has typically involved only the use of tests scores obtained by individuals. The fact that the inference on the equating function is based on the estimation of cumulative distribution functions makes the use of Bayesian nonparametric methods appealing. Moreover, when covariate information is available, a method which accounts for the estimation of the distribution functions depending on those covariates could enhance the accuracy of equated scores. In this paper, we present a novel Bayesian nonparametric model in which the distributions of scores are allowed to depend on covariates. The new dependent Bayesian nonparametric models for test equating are illustrated with both, a simulation study under different scenarios, and the analysis of data from a Chilean assessment test.

Parkzaal **F.2.-Bayesian IRT modeling**

A Framework of Facets Item Response Theory Models
Wang, Wen-Chung & Jin, Kuan-Yu
The Hong Kong Institute of Education

The Rasch facets model has been developed to account for facets data, such as students' responses to essays are graded by raters. In the facets model, ratings given to essays are treated as "virtual" items. In theory, the difficulties of these virtual items can be decomposed into a main effect of actual essays, a main effect of raters, and an interaction effect of essay by rater. To achieve better measurement quality, in the facets model, the interaction effect is constrained at zero (i.e., a rater's severity is constant across essay items). Just like the slope parameters can be added to the simple Rasch model to form the two-parameter model, so too for the Rasch facets model. Furthermore, similar to the linear decomposition on the item difficulties of virtual items, the slope parameters of the virtual items can be decomposed into main and interaction effects. A new framework for facets model is thus developed by setting a series of constrains on the main and interaction effects. The simulation results show that the parameters of the new class of facets models could be well recovered. Three empirical examples were provided. Further model generation is discussed.

IRT Models for Extreme Response Styles

Jin, Kuan-Yu & Wang, Wen-Chung
The Hong Kong Institute of Education

Extreme response styles (ERS) refer to a systematic tendency for a person to endorse extreme options (e.g., strongly disagree, strongly agree) on Likert or bipolar rating scale items. Standard IRT models, not considering ERS, will yield biased parameter estimates when ERS exists. In this study we proposed a new class of IRT models to directly account for ERS so that the tendency of ERS is quantified, and the resulting person measures are free from ERS and thus fair. Parameters of the new class of models can be estimated with marginal maximum likelihood estimation methods or Bayesian methods. In this study, the freeware WinBUGS was adopted. In a series of simulations, it was found that the parameters were recovered fairly well; and fitting a standard IRT model yielded substantially biased parameter estimates and slightly underestimated the test reliability. An empirical example of the 2009 International Civic and Citizenship Education Study was provided to illustrate the implication and applications of the new class of models.

A Rasch Model of Matching Test Performance

Matthew D Zeigenfuse, *Universität Zürich*
William H. Batchelder, *University of California*
Mark Steyvers, *University of California*

Matching tests are a measurement tool utilized in a variety of assessment contexts. In these tests, participants are shown two equal-length lists of test items and asked to associate each element of one list with exactly one element of the other. Because each item of each list is associated with exactly one item in the other list, a given participant's responses to test items are not independent of each other. Thus, matching tests violate local independence, precluding use of the standard Rasch model. Here, we present a novel extension to the Rasch model to matching tests. We also present a Bayesian method for estimating the parameters of our model and demonstrate through simulations that the method can recover its parameters when it is well-specified. Finally, we use our method to apply the model to real data to show that it draws reasonable conclusions.

New methods for multiple-group factor analysis (IRT)

Bengt Muthen, *Mplus*

This talk considers multiple-group factor analysis (IRT) across many groups such as with country comparisons of achievement (e.g. PISA, TIMSS) or in cross-cultural studies. The goal of multiple-group factor analysis is to study measurement invariance and also group differences in factor means and variances. This is typically carried out using confirmatory factor analysis with equality constraints. Four radically different methods are discussed here: Multiple-group analysis using an exploratory factor analysis model, Bayesian analysis with approximate measurement invariance, two-level analysis with random intercepts/slopes, and exploratory measurement invariance analysis. The fourth method has three steps. Step 1 is a configural factor analysis with no across-group restrictions. Step 2 makes an alignment optimization according to a simplicity criterion that favors few non-invariant measurement parameters. Step 3 adjusts the factor means and factor variances in line with the alignment. An application to binary math items in PISA is presented.

Jubileumzaal **F.3.-Modeling survey data**

An alternative way to model population ability distributions in large-scale educational surveys

Eunike Wetzel, *Otto-Friedrich-University Bamberg*
Xueli Xu, *Educational Testing Service*
Matthias von Davier, *Educational Testing Service*

In large-scale educational surveys, a latent regression model based on principal components extracted from background data is operationally used to compensate for the shortage of cognitive information. This method has several important disadvantages. The approach introduced here is to conduct a latent class analysis (LCA) to identify latent nominal variables that can be used to classify respondents with respect to their background characteristics. These classifications are then introduced as grouping variables in a multiple-group item response model that replaces the latent regression with individual vectors of predictors based on principal components. The goal of this project is to explore whether this approach yields similar results as the operational procedure. The LCA approaches differed regarding the number of classes and whether manifest class membership information or class membership probabilities were used as independent variables in the latent regression. Overall, recovery of the operational approach's group means and standard deviations was successful. Furthermore, the posterior means and standard deviations used to generate plausible values derived from the operational approach and the LCA approaches correlated highly. Thus, incorporating independent variables based on an LCA of background data into the latent regression model appears to be a viable alternative to the operational approach.

Detection of Aberrant Responding on Tests with Filter Questions

Jonathan M. Lehrfeld & Ying Liu
Fordham University

Questionnaires with filter items frequently appear in opinion-based surveys, marketing research, clinical interviews, and personality testing. A filter item consists of a chain of individual parts, which may involve one or more questions. If a respondent positively answers one part, he will be directed to follow-up parts; otherwise, he will skip all remaining parts of this chain and proceed to the next chain. Such adaptive design greatly reduces respondents' cognitive load as well as administration time. However, it is especially prone to aberrant responding, because aberrance at any part may lead to a very different subsequent administration sequence, and thus a diverging assessment conclusion. This study extends the work of Levine and Drasgow (1988) and Liu, Douglas, and Henson (2009) on appropriateness measurement, and develops two likelihood ratio test statistics to investigate aberrant responding for filter items. The test statistics follow a null distribution of a 50:50 mixture of χ2(1) and a constant of 0. Simulations suggest that hypothesis tests based on these statistics attain medium-to-high levels of power. We also examine how the performance of these statistics varies with test length, sample size, type and severity of aberrance, and distribution of the underlying latent trait.

Missing Data in Principal Component Analysis of Questionnaire data: A Comparison of Methods

Joost R. van Ginkel, *Leiden University*
Pieter M. Kroonenberg, *Leiden University*
Henk A.L. Kiers, *University of Groningen*

Principal component analysis is a widely used statistical technique for determining subscales in test and questionnaire data. As in any other statistical technique, missing data may both complicate its execution and interpretation. In this study five methods for dealing with missing data in the context of principal component analysis are reviewed and compared: pairwise deletion, the missing-data passive approach, regularized PCA, the EM algorithm, and multiple imputation. Simulations show that all methods give about equally good results for realistic percentages of missing data. Therefore the choice of a procedure can be based on ease of application, or purely the convenience of availability of a technique.

The Effect of Survey Nonresponse on Nonrespondents' Attitudes and Behaviors: An Application of Principal Stratification

Kristina Schmidt, *WHU - Otto Beisheim School of Management*
Walter Herzog, *WHU - Otto Beisheim School of Management*
Maik Hammerschmidt, *Georg-August-Universität Göttingen*

So far, survey nonresponse has primarily been considered as a statistical problem. In this research, however, we explore the psychological and behavioral ramifications of survey nonresponse. More specifically, we investigate the effect of a company's survey invitation on the attitudes and behaviors of non-responding customers. In particular, we estimate the effect of nonresponse on customers' repurchase probability. To do so, we analyze data from a large-scale field experiment and apply the logic of complier average causal effects (Angrist, Imbens, & Rubin, 1996) and principal stratification (Frangakis & Rubin, 2002). Our results provide evidence of a negative nonresponse effect; that is, the mere act of not responding to a company's market research efforts seems to decrease customers' repurchase probability. Additional analyses imply that this effect is triggered by self-attribution processes. Overall, our findings indicate that nonresponse may be more than a statistical problem since it modifies subjects' attitudes and behaviors. Furthermore, our findings have implications for researchers using instrumental variable models to estimate complier average causal effects. Such models usually assume a zero treatment effect for treatment noncompliers (the "exclusion restriction"). However, our findings imply that this assumption may not hold in all contexts and hence, it should be tested carefully.

Balkonzaal **F.4.-Factor analysis 1**

Stability of Rotated Factor Loadings in Exploratory Factor Analysis

Guangjian Zhang & Kristopher J. Preacher
University of Notre Dame

We report a surprising phenomenon: oblique CF-varimax and oblique CF-quartimax rotation produced similar rotated factor loadings and factor correlations but substantially different standard errors in an empirical example. We investigate the phenomenon by systematically manipulating factor loading patterns, levels of factor correlations, and factor rotation methods and observing their influence on standard errors. We discuss implications of the phenomenon on factor rotation in exploratory factor analysis.

Computational Identification of Optimal Confirmatory Factor Analysis Model Using Sparseness Constraint

Kohei Adachi, *Osaka University*
Nickolay T. Trendafilov, *The Open University*

A drawback of the existing confirmatory factor analysis (FA) procedures is that users must specify what factor loadings are constrained to be zero. We address this drawback to propose a procedure for computationally/automatically finding the optimal confirmatory FA model, though it is restricted to an orthogonal factor model and the number of factors is fixed. The proposed procedure consists of two stages. In the fist stage, the loss function of simultaneous FA is minimized over loadings and unique variances under a sparseness constraint, which directly requires a specified number of factor loadings to be exact zeros without any restriction on their locations. This stage is carried out repeatedly with the number of zero loadings changing over all feasible

integers, and the optimal number is selected using an information criterion in the second stage. The proposed procedure is illustrated with real data examples. We finally discuss the differences of our procedure in the first stage to the factor rotation after exploratory FA and to the existing sparse principal component analysis in which penalty functions are used.

A simple exploratory method to obtain the linear composites with a very simple pattern matrix

Takashi Murakami, *Chukyo University*

We will propose a procedure to assist the scale construction from questionnaire items. Consider the r linear composites defined by p variables, and simple regression of each variable on the unspecified composite. The p by r table of regression coefficients in corresponding cells and zeros in remaining cells can be seen as a very simple pattern matrix for a set of oblique factors (in generic sense) when all variables and composites are standardized. If the size of explained variance is maximized for each composite, it is shown that weights of variables for the composites on which they are not regressed are zero. It means that the problem is reduced to the classification of p variables into r exhaustive and mutually exclusive groups in order to maximize the sum of the largest characteristic roots of corresponding correlation matrices. While the proposed method is similar to cluster analysis of variables based on the distance matrix converted from a correlation matrix, it has a convenient aspect that any preprocessing for potential reversed variables, which sometimes leads to a subtle problem, is unnecessary. The efficiency and interpretability of the method compared with principal components analysis were investigated using several sets of real data.

Using purification procedure for choosing reference indicator in multiple-group factor analysis

Xiaoling Zhong, *The Hong Kong Institute of Education*

In the framework of multi-group confirmatory factor analysis (MCFA), the generally preferred strategy for identifying the model is the marker variable strategy, which chooses one of the indicators to be the reference indicator (RI) and fixes its factor loading to be 1.0. This strategy tacitly assumes the invariance of the true factor loadings across groups. When this assumption is violated, other model parameter estimates could be biased, developing inflated Type I errors or overly low power when testing factorial invariance. Therefore, choosing an RI with truly invariant loading across groups becomes the key point of testing factorial invariance, although it has not been fully accounted for so far.

This study proposes a purification procedure to iteratively identify an RI that is most likely to be invariant and studies its behavior when testing factorial invariance. A simulation study is performed when manipulating the type of observed variables (i.e., continuous, categorical, or mixed); the magnitude, proportion, the type of non-invariance (i.e., one-sided or two-sided, balanced or dominated); and the sample size. Results show that the proposed procedure works well as long as the proportion of non-invariance does not exceed 50%. In addition, this study also compares the proposed purification procedure with two existing procedures, the triangle heuristic/factor ratio (TH) method and the modification index (MI) method, in controlling Type I errors and maintaining acceptable power.

Stadszaal **F.5.-Linking with parameter drift**

Vertical Comparison using Reference Sets

Anton Béguin & Saskia Wools
Cito Institute for Educational Measurement

In educational assessment often comparisons are made between students with a different educational background. For example when performance levels are defined across students from different grades or across students following a different track of education. To be able to link the assessments vertical equating procedures are defined. These procedures will take into account that for some items the item characteristics differ between groups of students (DIF). In situations where the proportion of items with differences is large the validity of the vertical linking is challenged. This will especially be the case if designated tests are constructed for the different groups of students. In this paper an alternative procedure is introduced and compared with vertical equating. In this procedure a set of items is constructed that will serve as a basis for comparison between the different groups of students. This set of items consist of samples of items from the test for each of the groups of students and is called a reference set. The content of the reference set represents all aspects of the intended construct and is composed in such a way that none of the groups of students is advantaged. For the total reference set data are collected in each of the groups of students. Subsequently, tests for each of the groups of students can be linked to the reference set separately. For each test all score points are linked to scores on the reference set and as a consequence to a common metric.

Analysis of the effect of Item Parameter Drift on Vertical Scaling: Examination of Different Drift Situations

Meng Ye & Tao Xin
Beijing Normal University

Following the research presented at the 2013 annual meeting of the National Council on Measurement in Education, this paper examine the effect of item parameter drift (IPD) of common items across test levels on vertical scaling with three parameter logistic (3PL) model. The factors examined included numbers of drifted items, difficulty location of the drifted items in the level they come from, and grade pair where IPD occurs. For purpose of evaluation, root mean squared error (RMSE) and bias were computed over the replications for two sets of parameters of interest, ability parameters and the properties of the developmental score scale. The main findings were as follows. (1) When there were IPDs, the results for the grade-specific ability estimates became worse conditional on the grade/grade pair where IPD occur. (2) The effect of IPDs on standard deviations of the grade-specific ability estimates and grade-to-grade variability depended much on the location of the drifted items in the level they come from. (3) IPD could lead to inaccuracy and overestimate of grade-to-grade growth and effect size for the grade pair it occurred. (4) When three items drifted across test levels, IPD might cause a substantial threat to vertical scaling.

Maintaining Scale Stability by Detecting

Item Parameter Drifting

Rui Guo, *University of Illinois*

An important goal of many large scale testing programs is to maintain invariant item parameters across different testing time points. Sometimes, however, item parameters do not remain invariant over testing occasions due to factors other than sampling error. When the invariance assumption does not hold, the item is considered to have drifted from its original parameter values. Detecting item parameter drifting (IPD) has always been an important research topic for paper and pencil (PPT) based assessment. The presence of IPD may cause inaccuracy in linking, which may pose a threat to the validity of score comparisons across different test administrations.
In computerized adaptive testing (CAT), the damage caused by IPD is even more severe as a drifted item can directly affect the quality of the entire item pool. Unfortunately, most current detection methods were originally developed for PPT, which may not be generalized to CAT in a straightforward manner. Since CAT continuously assesses reading, writing, math, and computer skills across years,

maintaining a stable scale by detecting drifted items is of great importance.
The objective of this research is to develop an array of methods that can dynamically detect IPD under the CAT framework. Although several methods have been proposed and can effectively detect drifted items, the traditional methods are mostly based on PPT. Under the CAT environment, the detection of item drifting is even more challenging because few studies have been done and the response data is a spars matrix. Moreover, most of the existing methods are based on detecting individual item drift, which may not be ideal for detecting the drift of the test characteristic curves. Also, certain types of parameter drift by the existing methods, for example, c-parameter drift, are difficult to detect. Inspired by stepwise regression in Statistics, this study introduces a stepwise test characteristic curve (stepwise TCC) method to dynamically detect item parameter drift under the unidimensional 3PL framework without setting an arbitrary cut-off value. More importantly, the study will generalize the stepwise TCC method from PPT to CAT framework.

Examining Comparability for Test Forms by Assessing Equity Properties

YoungKoung Kim, *The College Board*
Stephen Cubbellotti, *The College Board*
Thomas Proctor, *The College Board*
Won-Chan Lee, *The University of Iowa*

When an assessment has two separate operational forms which are equated via two “parent” forms instead of being equated directly to each other, ensuring comparability between the two forms can be a major concern in the equating of the assessment. If examinee performance is quite different between the two forms, the relatively large difference in test taker performance can add more complexity to maintaining the comparability between the two form. If equating is performed properly, scores on the two forms can be used almost interchangeably. Thus, the present study proposes a way to examine the comparability between two forms by assessing first-order equity (FOE) and second-order equity (SOE) using the PSAT/NMSQT(P/N) assessment which is equated via parent SAT forms. If FOE holds, examinees with a given true score will have the same expected scale score on two PN forms. If SOE holds, the standard deviation of the scale score distribution at a given true score will be the same on the two P/N forms. FOE and SOE on the two PN forms are assessed within an IRT framework. The 3PL model is used to calibrate the items from the P/N data. Once items from the Wednesday and Saturday P/N forms are placed on the SAT scale using the Stocking-Lord method, FOE and SOE are examined. To interpret the differences between the two forms at each score level, the practical criterion of the Difference that Matters (DTM; Dorans & Feigenbaum, 1994) are used.

Hemelzaal

F.6.-Applications 2

Comparison of Linear, Computerized Adaptive and Multilevel Adaptive Versions of Mathematics Assessment of the Turkish Pupil Monitoring System

Semirhan Gokce & Prof. Dr. Giray Berberoglu
Middle East Technical University & Cito Turkiye

The purpose of the present study is to compare the computer based linear test results of the Turkish Pupil Monitoring System (TPMS) with the Computerized Adaptive Testing (CAT) and Multilevel Adaptive Testing (MLAT) in mathematics assessment. On the basis of the real responses obtained on the TPMS, different scenarios of CAT and MLAT were tested in post-hoc simulations with various starting rules, termination criteria, and different exposure controls of CAT by the use of Maximum Likelihood and Weighted Maximum Likelihood estimation procedures. Preliminary results indicated that WML with moderate initial item difficulty and fixed test reliability termination along with the item exposure and content control strategies produced defensible results for the CAT. On the other hand, a fixed subtest as the starting point followed by multiple subtests seemed to be producing more valid results in line with the content sampling.

Different scenarios are tested in MLAT to determine the optimum length of subtests and starting rules in predicting the scores obtained in computer based linear test administration.

An investigation on Turkish students' mathematical literacy skills against certain variables according to PIA 2003 results

Gonca Usta, *Cumhuriyet University, Sivas*
Müge Uluman, *Marmara University, Istanbul*
Yurdagül Günal, *Marmara University, Istanbul*

The Programme for International Student Assessment (PISA) developed by the Organisation for Economic Co-operationand Development (OECD) is a survey conducted every three years in order to find out 15-year old children's ability of using school knowledge in their Daily life in leading developed countries. PISA not only measures students' cognitive abilities such as mathematic literacy, reading, applied science sliteracy and problem solving but also collects information about factors affecting students' achievment by means of a questionnaire taking half an hour. In addition, another questionnaire is administered to school managers and parents. On PISA study, one of these cognitive skills is taken as a basis every three years. Mathematics is the course for PISA 2003. Today every body needs to be able touse maths as an instrument. Thus, the concept of "mathematical literacy" comes to the front. It is defined as the capacity of applying to mathematics in order to realize how maths can be used in daily life and to meet needs. In this scope, the aim of PISA 2003 is to identify maths questions in the context of students' real life, toexpressthem in terms of maths problems and to assess the level reached in coping with them. The aim of present study is to investigate the relationship between maths achievment and variables such as "Attitudes towards School", "Student-Teacher Relations", "Interest in Mathematics" and "School Type"in the light of PISA 2003 data. This is a descriptive study and study sample is comprised of PISA 2003 Turkey participants. Study data was obtained from the OECD's web page as well as http://pisa2003.acer.edu.au/ and http://egitek.meb.gov.tr/earged/arasayfa.php?g=87. First, study data was filtered for analysis. Hierarchical linear model was used to examine the analyzed study data at two levels as student and school level. Student level includes variables like "Attitudes towards School", "Student-Teacher Relations", "Interest in Mathematics" and "Mathematical literacy skills" while school level includes "School Type".Results of analysis will be listed as findings.

Exploring the relation between socio-economic status and reading achievement in PISA 2009 through an Intercepts-and-Slopes-as-Outcomes paradigm

Khurrem Jehangir, *University of Twente*

In some countries more than others, factors like parental socio-economic status (SES) can cause inequalities in educational achievement. Here we show how the mechanisms leading to such inequalities can be scrutinized by involving background variables which impact the relation between SES and achievement. We use the Intercept-and-Slopes-as-Outcomes paradigm which recognizes that the outcomes of schooling systems are not only characterized by average achievement (the intercept) but also by the achievement-SES regression slope. We show how background variables moderate the relationship between SES and achievement. As an illustration, we examine the relationship between reading achievement and SES, and how this is moderated by school funding and curriculum. This is done for several countries that participated in the PISA 2009 cycle.

The Impact of Private Supplementary Tutoring on Science Achievement: A Propensity Score Analysis

Yu Jiang & Tao XIN
Beijing Normal University

Shadow education has become much more visible worldwide during the last decade, especially in China. However, it is important but difficult to identify whether shadow education can improve students' academic achievement. The present paper attempts to answer the following questions: (1) What factors influence students' participation in shadow education? (2)Does shadow education improve the students' academic performance, when other effects match equally? If it works, how big the effect is?
The propensity score is the conditional probability of assignment to a particular treatment given a vector of observed covariates. Using data gathered by National Assessment Education Quality(NAEQ) inChina2010, the present research employs the method of propensity score to study the causal effect between students who participant private supplementary tutoring and their Science Achievement. Several observed variants (family socioeconomic status, school environments and individual characteristics of students)are compressed into a composite score according to which two Groups (Students who attend shadow education and who don't)are matched, and then calculates the average difference in outcomes between these two groups. The results of the Propensity score analysis indicate no significant performance advantage between participants and nonparticipants.

10:35 - 10:55 ***Break***
Ravelijn

10:55 - 11:20 State of the art lecture

Concertzaal **Ridge structural equation modeling with large p and/or small *N***
Ke-Hai Yuan, *University of Notre Dame*

Existing methods for structural equation modeling (SEM) are developed using asymptotics by assuming a large number of observations (*N*) and a small number of variables (*p*). However, in practice, *p* can be very large while *N* is always limited due to not having enough participants in data collection. Then the SEM methods as implemented in software become invalid in either failing to run or by giving statistically meaningless output. Currently, there is no effective procedure for SEM with small *N* and/or large *p*. Instead of modeling the sample covariance matrix S, ridge SEM models S + $a\mathbf{I}$, where $a > 0$ is a constant. Our results show that ridge SEM yields more accurate parameter estimates than the normal-distribution-based MLEs even when the sample is from a normally distributed population. In particular, using a technique called Monte Carlo modeling, we develop a test statistic that is better approximated by the nominal chi-square distribution than any existing statistic with real data, including samples with literally singular covariance matrix. Empirical results indicate that ridge SEM works well whenever $N \geq \max(2p, 30)$.

Parkzaal **Recent developments in estimation and testing of Latent variable models for categorical data**
Irini Moustaki, *London School of Economics*

Latent variable models are widely used in social sciences for measuring unobserved constructs such as intelligence, fear of crime, anxiety etc. In the last two decades, latent variable models have been extended to account for categorical responses, multidimensional latent variables, effects of explanatory variables, non-linear relationships, longitudinal data, missing values, outliers and complex survey data. At the same time, those extensions have led to complex models with many parameters in which estimation methods such as maximum likelihood is difficult if not intractable.
In this talk, we discuss weighted and unweighted composite likelihood estimation methods for a large family of latent variable models and recent extensions of limited goodness-of-fit tests statistics under composite likelihood estimators. Finally we will discuss estimation and testing within the composite likelihood framework to account

for complex survey sample designs. Applications using survey data will be used throughout the talk for illustrating the methods.

11:25 - 12:15 Keynote session

Concertzaal **Sequential Experimentation and Adaptive Testing**
Zhiliang Ying, *Columbia University*

Sequential analysis was developed into a branch of statistics during the World War II. This talk reviews and summarizes the classical sequential probability ratio test as well as subsequent and related developments including group sequential methods, stochastic approximation, adaptive control and quantal response experiments. Their connections to clinical trials, toxicity studies, engineering feedback systems and, especially, computerized adaptive testing are discussed. For adaptive testing, focus is on the uni-dimensional IRT models and multi-dimensional diagnostic classification models and on various information functions used for item selection.

12:20 - 13:35 Lunch break
Jubileumzaal Business meeting
Balkonzaal Student luncheon

13:40 - 15:00 Parallel session G

Concertzaal **G.1.-Psychometric Modeling of Responses and Response Times**

Generalized linear item response modeling of responses and response times
Dylan Molenaar, *University of Amsterdam*
Francis Tuerlinckx, *University of Leuven*
Han van der Maas, *University of Amsterdam*

A general item response theory approach to the analyses of responses and response times is outlined. In this approach, generalized linear measurement models are specified for the responses and the response times separate. These two models are subsequently linked by formulating cross relations between them. In this talk it is discussed how popular existing models from the psychometric literature are special cases in this framework depending on restrictions in the cross relations. This allows us to compare existing models conceptually and empirically. For instance, it is shown under what circumstances the hierarchical model of Van der Linden (2007) and the model by Thissen (1983) coincide. In addition, it is discussed how the models for personality data proposed by Ferrando & Lorenzo-Seva (2007) fit into this framework. Various extensions of the traditional models are proposed motivated by practical problems. In addition, some real data analyses are presented.

Using response times in massive Cat: a Bayesian approach.
Maarten Marsman, *Cito - University of Twente*
Timo Bechger, *Cito Institute for Educational Measurement.*
Cees Glas, *University of Twente*
Gunter Maris, *Cito/ University of Amsterdam*

Modern applications of CAT involve large samples of respondents answering frequently items from substantial item banks, over extended periods of time, where both accuracy and response times are registered and with a sophisticated IRT model used to drive the CAT.
For instance, the Maths Garden (www.mathsgarden.com) gathers over 300.000 responses each day and the Chess Tactics server (chess.emrald.net) gathered over 80 million responses from more than 60.000 chess enthusiasts with an item bank consisting of more than 30.000 items, all of which are calibrated continuously.
The estimation of IRT models in real-time with problems of this scale requires

specially designed algorithms. We propose a new class of composition algorithms to sample from the posterior distribution of ability, which can be applied to any marginal IRT model for which it is possible to generate data. These algorithms can be used as part of a Gibbs sampler to tackle large-data applications of marginal IRT models in real-time.
We provide an illustration using data from the Maths Garden, and use the algorithm to estimate both person and item parameters of a complex IRT model for response accuracy and response times called the Signed Residual Time model (Maris & van der Maas, 2012).

The measurement of speed reconceived and what can be learned from a failure
Paul de Boeck, *Ohio State University*
Ivailo Partchev, *Cito Institute for Educational Measurement.*

Given that speed is a relative notion, it cannot be measured through just response times. Speed is distance covered per unit of time. We will present a simple speed-accuracy model with distance measured on the logit scale for accuracy. The model is used for response time and accuracy data from a verbal analogies test. It was found that depending on the item, accuracy was increasing or decreasing with response time. The slope that was supposed to measure speed ranged from negative to positive and was highly correlated with item difficulty. These results are a failure for the model as a way to measure speed. Interestingly, the results are also inconsistent with models based on time homogeneous process assumptions, such as the diffusion model.

How to Measure and Control Test Speededness?
Wim van der Linden, *CTB/McGraw-Hill*

Test speededness is defined as the probability of a test taker running out of time before completing the test. The main determinants of the probability are: (a) the time limit selected for the test; (b) the time intensities of its items; and (c) the speed at which the test taker works. It is shown how to derive the probability for the lognormal model for response times on test items (van der Linden, 2006). In addition, we show how to use the probability to design a fixed-form test, run an adaptive test , or select a time limit for a given test control its degree of speediness.

Parkzaal **G.2.-MIRT 2**

A Study on Parameter Estimation for Multidimensional Item Response Models by Parallel Adaptive Markov Chain Monte Carlo
ChunLai Wu, *Tokyo Institute of Technology*
Shin-ichi Mayekawa, *Tokyo Institute of Technology*
Natsuko Kobayashi, *The Japan Institute for Educational Measurement, Inc*
Norio Hayashi, *The Japan Institute for Educational Measurement, Inc*

Markov Chain Monte Carlo(MCMC) method has become an important numerical tool in statistics. It is crucial to tune the proposal scaling in order to make the algorithm be converged well. It is proposed that Adaptive Markov Chain Monte Carlo(AMCMC) algorithm can handle the tunings automatically. Multidimensional Item Response Theory(MIRT) has developed gradually since its inception (e.g.,McDonald,1967,1997; Muraki & Carlson,1974;Reckase,1985,1997;Samejima,1974).Lately its statistical tractability has been improved considerably, and it is now possible to use several of its models for operational testing. However, it is difficult to estimate item parameters of MIRT when number of dimensions is too large by the use of Maximum Likelihood Estimation (MLE) method. Hence, in recent years, MCMC has been attracting attention as a parameter estimation method for MIRT models. An AMCMC algorithm, Metropolis-Hastings Robbins-Monro (MHRM) (Cai,2010-b,2010-c) has been implemented in a statistics software called IRTPRO (Item Response Theory for Patient-Reported Outcomes).In this study, we propose a Parallel Adaptive Markov Chain Monte Carlo(P-AMCMC) algorithm. P-AMCMC algorithm is able to estimate item parameters automatically in much less time by the simultaneous workings of parallel Markov Chains. We will show the results of comparing the estimation accuracy and the run time between P-AMCMC algorithm and AMCMC algorithm by simulations on multidimensional two-parameter logistic compensatory(M2PL-C) model.

Italian student assessment at the end of lower secondary school: a comparison of IRT multidimensional approaches

Mariagiulia Matteucci & Stefania Mignani
University of Bologna

In Italy, large-scale student assessment is conducted by the National Evaluation Institute for the School System (INVALSI) at different school grades. The aim of the INVALSI test is to evaluate student competencies in language and mathematics, giving a score to each student. This final score is constructed by experts taking into account not only the number of correctly scored items, but also item characteristics. The assessment is treated in a unidimensional context while the test explicitly consists of different subtests. For this reason, we consider a multidimensional approach and, in particular, we compare IRT models with different ability structures to choose the most appropriate model. In fact, IRT multidimensional models (e.g. Reckase, 2009) are able to describe the manifest item responses with an increased degree of accuracy by accounting for the information from responses of correlated abilities. For this class of models, the joint estimation of item parameters and individual abilities is rather complex and we resort to Markov chain Monte Carlo (MCMC) methods and particularly to the Gibbs sampler adopting a fully Bayesian approach. The advantages are the capability of including uncertainties about item parameters and abilities in the prior distributions, and the use of Bayesian model comparison techniques.

When is a paradox not a paradox? When it is good MIRT estimation

Mark D. Reckase & Xin Luo
Michigan State University

In the past few years, an interesting property of the estimation of the location of examinees in a multidimensional space using multidimensional item response theory models (MIRT) has come to light. Under certain conditions, when a correct response to an item on a test is changed from a correct response to an incorrect response, one of the coordinates for the estimate of the location of the examinee in the multidimensional space increases when a casual expectation might be that all coordinates for the estimate of location should decrease. Previous papers have proved that this must occur under certain circumstances. This paper elaborates on the previous work to make very explicit when the “paradox” will occur. The occurrence is related to the form of the likelihood function which is a result of the characteristics of the items used to estimate the location of an examinee in the multidimensional space. The paper includes a demonstration that the set of coordinates with the “paradox” give a better estimate of the location than if the paradox were avoided.

Jubileumzaal **G.3.-Dif in diagnostic models**

Differential Item Functioning detection in DINA model using standardized differences statistics

Guaner Rojas, *Universidad Autonoma de Madrid*
Jimmy de la Torre, *Rutgers University*
Julio Olea, *Universidad Autonoma de Madrid*

This paper proposes two new statistics for differential item functioning (DIF) detection in the DINA model. The DIF detection indices are based on standardized differences in the item parameters between groups of examinees. A simulation study is implemented to evaluate the performance of the proposed procedures in terms of their Type I errors and powers. We also compare the performance of the proposed statistics against that of the Mantel Haenszel method with attribute profiles as matching criterion (MHP). The simulation study includes the factors such as, sample size, DIF size, DIF type, number of attributes per item and quality of the items. The results indicate that new indices showed good control over Type I error rates and high power rates. Further, the new proposed statistics performed better than MHP in detecting DIF.

Identifying Similarities and Differences in Attribute Prevalence across Countries in the TIMSS 2007 Eighth Grade Mathematics Assessment Using a Cognitive Diagnostic Model

Matthew Johnson, Young-Sun Lee, Jung Yeon Park, Jianzhou Zhang,
Ruchi Sachdeva & Marcus Walden
Teachers College, Columbia University

Cognitive diagnostic models (CDM), which are considered to be a popular modeling method to diagnose the cognitive attributes of students, can also be used to compare the knowledge skills of students across countries. As one such modification, ‘multi-group DINA’ (MG-DINA) model is developed and fit 2007 TIMSS Grade 8 Mathematics assessment; specifically we analyze item responses to 88 released items by students in 11 of the high-achieving countries. As the model allows us to estimate attribute prevalence for each country, the aim was to conduct statistical tests whether there exists difference in attribute prevalence across countries. Due to complex sampling design in TIMSS, however, it is not straightforward to perform likelihood ratio test and/or Wald test. Thus, we used Jackknife replicated weights that were provided in TIMSS to approximate covariance matrices of the prevalence estimates for the two tested aforementioned. For doing likelihood ratio test, a parametric bootstrapping method was used to obtain an empirical distribution of likelihood ratios of a typical DINA model that assumes same attribute prevalence among countries versus the MG-DINA model that assumes different attribute prevalence among countries. We also applied Wald tests to test attribute differences via omnibus tests and pairwise comparisons using TIMSS.

Assessing DIF in Computerized Adaptive Testing Environment under Cognitive Diagnostic Models

LI Xiaomin & Wen-Chung Wang
The Hong Kong Institute of Education

Item bank update is an important operational issue in computerized adaptive testing (CAT). New items have to be carefully assessed before putting into item banks. Assessment of differential item functioning (DIF) is a routine exercise. In CAT, old items are adaptively administered until a pre-specified stopping rule is reached. Then, new items are administered. After the CAT program, the parameters of the new items are calibrated and their DIF is assessed. Previous studies have developed DIF assessment methods in CAT environment, but under item response theory models (Lei, Chen, & Yu, 2006). In recent years, CAT algorithms under cognitive diagnostic models have been developed (Cheng, 2009). In this study, we developed DIF assessment methods in CAT environment under cognitive diagnostic models and conducted a series of simulations to evaluate their performance. The results of the simulations showed that the modified Mantel- Haenszel method and the concurrent CDM method performed satisfactorily in terms of Type I error and power. However, when both the guessing and slipping parameters had DIF favoring the same group, only the concurrent CDM method performed appropriately. Another advantage of the concurrent CDM method was its direct estimation of model parameters between reference and focal groups.

Detecting Differential Item Functioning Using DINA Model with Logistic Regression Method

Zhuoran Wang, *Beijing Normal University*

This paper tries to detect Differential Item Functioning(DIF) in Cognitive Diagnostic assessment with Logistic Regression method. Logistic Regression is a widely used DIF detecting method. Not only can it distinguish uniform DIF and nonuniform DIF, but it is also easy to be applied. In this paper, we used a simulation study to show the way DIF detection can be done with Logistic Regression method, and compared the traditional matching criteria in DIF procedure(e.g., total score) to new conditioning variable for DIF detection, that is attribute mastery pattern. Three variables were manipulated in this study: two sample sizes(500 and 1000 examinees in each group), five types of DIF (introduced by manipulating the item parameters in the DINA model), and two levels of DIF amount(moderate and large DIF). The result showed attribute mastery pattern matching is more effective than the traditional matching criteria, for it has a lower Type 1 error rates and a higher power rates, especially when the sample size and DIF amount are small.

Balkonzaal

G.4.-(CCC) Classification, Clustering and Correspondence Analysis -2

A measure and statistical test for regular minimality

Ali Ünlü, *TU München*
Matthias Trendtel, *Bifie*

The law of regular minimality (RM) for same-different judgments was introduced by Ehtibar Dzhafarov as a fundamental property of discrimination and a necessary condition for Dzhafarov and Colonius' theory of Fechnerian scaling. A matrix of discrimination probabilities satisfies RM if every row and every column of the matrix contains a single minimal entry, and an entry minimal in its row is minimal in its column. In this paper we propose a measure based on which the compliance of a matrix of discrimination probabilities with RM can be tested statistically. The measure for testing RM is demonstrated on real data.

Flexible Multiclass Support Vector Machines: An Approach using Iterative Majorization and Huber Hinge Errors

Gertjan van den Burg, *Erasmus University Rotterdam*

In psychology one often attempts to predict class membership based on predictor variables. Traditionally this is done through multinomial regression or discriminant analysis. In the Machine Learning literature the Support Vector Machine (SVM) is very popular for predicting class labels in binary classification problems. We propose a flexible multiclass SVM which can be used for classification problems where the number of classes $K \geq 2$. Traditional extensions of the binary SVM to multiclass problems such as the one-vs-all or one-vs-one approach suffer from unclassifiable regions. This problem is avoided in the proposed method by constructing the class boundaries in a $K – 1$ dimensional simplex space. Nonlinear classification boundaries can be constructed by using kernel functions or spline transformations. Similar to earlier work by Groenen et al. (2008), an Iterative Majorization algorithm is derived to minimize the constructed loss function. From comparisons with existing methods it is found that in most cases the proposed method has similar performance and in some cases shows a higher classification accuracy.

Functional Generalized Reduced Clustering

Michio Yamamoto, *Osaka University*

In social science and many other fields, there are a number of circumstances when the data are curves; many statistical methods to analyze the functional data have been developed. This work presents a new procedure for simultaneously finding the optimal cluster structure of multivariate functional objects and finding the subspace to represent the cluster structure. The proposed method is conducted by minimizing a distance between a functional object and its low-dimensional representation with clustering penalties, and it can be considered to be a generalized model including existing functional cluster analyses with dimension reduction. In addition, even if the data have a structure which is independent of the true cluster structure and affects the performance of clustering, the proposed method finds the optimal subspace to partition the objects by eliminating the effect of the disturbing structure. An efficient alternating least-squares algorithm, consisting of the gradient projection algorithm and the k-means algorithm, is described. Analyses of artificial and real data examples demonstrate that the proposed method can give correct results but existing methods cannot.

Classification of Person Score Profiles in Biplot via Cluster Analysis

Joe Grochowalski & Se-Kang Kim
Fordham University

Purpose/objective: The authors introduce a method that allows researchers to classify person score profiles as a limited number of groups with scores, and to interpret them via a biplot.
Method: Use PCA biplot analysis of the score data to assign component scores to each person. Apply a clustering algorithm to component scores to optimize group classification. Group membership is plotted in a biplot, groups are interpreted by projecting group centroid scores onto biplot variable vectors, and scores are calculated for each person on each group. With the classification and the score we can describe the shape of the person's score profile as well as the "amplitude" of the profile peaks and valleys. We demonstrate this method using longitudinal elementary school reading and math scores for 2,708 students.
Results: We identified five groups of profiles. An example score pattern has reading and math scores that start very high (relative to the mean) in early grades, but at the latest time point the scores are near average. The higher a person's score on this group, the more pronounced this pattern will be.
Significance: Plots are easy to interpret and summarize complex data in a few simple profiles. Person profiles can be classified and used as quantitative predictors in models. Clinical diagnoses can be made based on any two measures (assuming a high proportion of variance explained in the biplot). This is a simple method for grouping and interpreting person profiles and can be extended to multiple dimensional cases (i.e. dim 1 vs. dim 3, etc.).

Stadszaal

G.5.-Missing data

A Comparison of Maximum Likelihood and Multiple Imputation for Structural Equation Models With Missing Data

Ashley Lawrence & Taehun Lee
University of Oklahoma

It has been known in theory that two classes of missing data procedures, maximum likelihood (ML) and multiple imputation (MI) are equivalent. It has also been known empirically that ML and MI are not equivalent as practiced. However, there has been a surprising lack of empirical research into the relative performance of ML and MI in the context of structural equation modeling (SEM). A primary goal of this paper is to examine conditions in which ML and MI yield equivalent or divergent results via Monte Carlo simulation studies. The Monte Carlo studies were designed to examine the impact of various independent variables, including missing percentage, missing data mechanism and the number of imputations, on the outcome measures, including convergence rates of model estimation, bias and root-mean-squared-error for parameter and standard error estimates, and empirical Type I error rates of model-fit testing. The effects of the independent variables were examined using both correctly specified and misspecified analysis models. Preliminary results indicate that the two key independent variables have dramatic effects on the outcome measures. In particular, different types of MAR missing data mechanisms and the number of imputations have important implications on the relative performance and equivalence of MI and ML.

Comparison of methods for handling missing data in a multi-item instrument on item-level and scale-level

Iris Eekhout, *VU medical center*

Regardless of the proportion of missing values, complete-case analysis is most frequently applied, although advanced techniques such as multiple imputation are available. The objective of this study is to explore the performance of simple and more advanced methods for handling missing data in case some, many, or all item scores are missing in a multi-item instrument.
Real-life missing data situations were simulated in a multi-item scale used as a covariate in a linear regression model. Various missing data mechanisms were generated with an increasing percentage of missing data. Subsequently, several techniques to handle missing data were applied to decide on the most optimal technique for each scenario. Fitted regression coefficients were compared using the bias and coverage as performance parameters.
Mean imputation caused biased estimates in every missing data scenario when data are missing for more than 10% of the cases. Furthermore, when a large percentage of cases contained missing items (>25%), multiple imputation methods applied to the items outperformed methods applied to the scale score.
We recommend applying multiple imputation to the item scores in order to get the most accurate regression model estimates. Moreover, we advise not to use any form of mean imputation to handle missing data.

Comparison of Nested Models for Multiply Imputed Data

Yoonsun Jang, Laura Lu & Allan S Cohen
University of Georgia

Missing data are almost inevitable in social science research, especially when data are collected through surveys, tests, or questionnaires (e.g., Little 2012 & Rubin, 2002). A number of statistical approaches for dealing with missing data have been proposed. Multiple imputation in particular has become an almost indispensable method (Schafer, 1997). Model comparison and model selection with missing data, however, have not been fully developed (Lee & Chi, 2012). As hierarchical data are very common in social science research (Singer & Willett, 2003), in this paper we focus on comparison of hierarchically nested models by using differently weighted log-likelihood ratios for multiply imputed data. We compare four types of weights suggested by Kientoff (2011) for the likelihood ratio test. Simulation studies will be conducted to compare nested models with multiply imputed data. The data will be generated under different missingness conditions, sample sizes, and missing data rates. A real data example also will be presented to illustrate the application of different model selection criteria for the different weight conditions.

Hemelzaal

G.6.-Modeling issues

Applying the Axioms of Additive Conjoint Measurement to a Hierarchy of Latent Variable Models

David Torres Irribarra, *University of California*

Ronli Diakow, *University of California*
Ben Domingue, *University of Colorado Boulder*
Derek Briggs, *University of Colorado Boulder*

Various item response models could be presumed to underlie the generation of any dataset when an examinee interacts with an assessment instrument. Understanding the differences between various models and the score scales that can be created to support them is important since different models lead to different inferences about inter-individual differences. Domingue (In Press) evaluated whether data consistent with the Rasch model possessed sufficient structure, according to the axioms of Additive Conjoint Measurement (ACM; Luce & Tukey, 1964), to support a score scale with interval properties.

In this paper we focus on an expanded hierarchy of item response models that vary with respect to assumptions about the structure of the latent variable and monotonicity constraints on items and respondents (Torres Irribarra and Diakow, 2012). Using Domingue's methodology to check the cancellation axioms of ACM, we highlight two main findings. First, although models that imply either invariant person or item orderings are formally symmetric, there are substantial differences in the results due to the difference in the number of persons and items typically observed. Second, the methodology used to detect axiom violations appears insensitive to violations of the double cancellation axiom, opening theoretical and practical issues when testing for interval structure.

Robustness study of the polychoric correlation estimation: with the applications to the elliptical distribution family

Shaobo Jin & Fan Yang-Wallentin
Uppsala University

Asymptotic robustness against the misspecification of the underlying distributions for the estimation of the polychoric correlation is studied. The asymptotic normality of the quasi-maximum likelihood estimator is derived under the two-step estimation framework. Different underlying continuous distributions, both normal and non-normal distributions, are applied to the general form. The numerical results show that the underlying t distribution with fixed degrees of freedom (i.e. 3 or 4) leads to smaller biases and asymptotic variances than the standard normal distribution within the elliptical distribution family.

Modeling survey data with inflated zero and K responses

Ting Hsiang Lin, *National Taipei University*

Zero-inflated Poisson (ZIP) regression is a popular tool used to analyze data with inflated e zeros. Much work has been devoted to fit zero-inflated count, however, most models heavily depend on special features of the individual data.

When there is a sizable group of respondents who endorse the same answers other than zero, this makes the data have two peaks. Survey question such as ‘On average, how many days a week do you drink alcohol?’ we would expect many counts of zeros and sevens for nondrinkers and regular drinkers, respectively. Other situation involves questions about major life events, such as the number of marriages, in which case we expect that ‘zero’ and ‘one’ will dominate the responses. Also, most respondents tend to endorse some particular responses and provide answers with a rounded-off number or multiples of 5 or 10. In these situations, we expect to see some peaks in the data structure.

In this paper, we proposed a model with the flexibility to model excessive counts in addition to zeros, and the model is a mixture of multinomial logistic and Poisson regression, in which the multinomial logistic component models the occurrence of excessive counts, including zeros, K (where K is a positive integer) and all other values. The Poisson regression component models the counts that are assumed to follow a Poisson distribution. Two examples are provided to illustrate our models when

the data have counts containing many ones and sixes. As a result, the zero-inflated and K-inflated models exhibit a better fit than the zero-inflated Poisson and standard Poisson regressions.

15:00 - 15:20 ***Break***
Ravelijn

15:20 - 16:20 **Presidential Address**

Concertzaal **Psychometrics Behind Computerized Adaptive Testing**
Hua Hua Chang, *University of Illinois at Urbana-Champaign*

Over the past twenty years Computerized Adaptive Testing (CAT) has become an increasingly important testing mode in large scale educational assessment. This presentation reviews a 2 decade psychometric progress of supporting various CAT designs and implementations. It is anticipated that more researches are much needed to address issues emerged from practical challenges in a range of fields in addition to large scale high stake testing, such as k-12 accountability testing, quality of life measurement, patient reported outcome, and media and information literacy measurement. CAT is making a substantial influence on the functioning of society by affecting how people are selected, classified, and diagnosed. CAT research will lead to better assessment, and hence benefit society

16:25 - 17:45 **Parallel session H**

Concertzaal **H.1.-Reliability 1**

Efficient reliability: A standard for test reliability in group research
Jules L. Ellis, *Radboud University Nijmegen*

In the past century, considerable effort has been devoted to estimating reliability coefficients, but almost no theory exists about the acceptable values of reliability coefficients. Thus, if a student asks the obvious question of how high a test's reliability should minimally be, we cannot give a clear answer. Many authors adhere to the standard that reliabilities should be at least .70 or .80 in a group experiment where the test is being used as the dependent variable. This standard lacks rigorous justification. However, a rational standard can be derived on basis of the principle that the researcher wants to maximize the statistical power of a t-test or ANOVA, given the costs of the experiment, measured in time. The outcome of this maximization is surprisingly simple: The reliability should be equal to $1 - c/t$, where c is the time needed for administering the test, and t is the total experiment time. This value will be called the efficient reliability. It can be smaller than .70 or larger than .80, depending on the experiment and the test. In this derivation, several additional assumptions are needed, such as the Spearman-Brown formula. These will be discussed.

Introducing a Categorical Reliability that is Spearman-Brown-Consistent
Kappler, Gregor, *University of Vienna*

Researchers report observer agreement of categorical ratings by using agreement statistics (ASs), mostly Cohen's kappa. Numerous critical reviews pointed out weaknesses of kappa, and many alternative ASs have been proposed. Furthermore, kappa and alternative ASs lack a theoretical model (Kraemer, 1979) unlike reliability rho which is grounded in Classical Test Theory (CTT). This contribution defines reliability iota information-theoretically using entropies within a theoretical model (NPMM) in parallel to interval-scale reliability rho. To compare iota with other ASs, results of new simulation test clarify (a) that Spearman-Brown aggregation extends to categorical data because multiple observations per subject reduce unreliability, and (b) that emerging Spearman-Brown curves of Cohen's kappa strongly depend on the number of categories K while Spearman-Brown curves of rho and iota are independent of measurement variance and K respectively (Spearman-Brown Consistent). For example, kappa=.60 for K = 6 is equivalent to kappa=.71 for K = 2. For larger K kappa is increasingly underestimated. Simulation and examples demonstrate that iota is more sensitive than kappa to both agreement and disagreement. The information-theoretic construction parallel to rho and the simulation results strongly suggest to use iota for reporting observer agreement.

Split-half Reliability for Test Scale Scores

Rashid S Almehrzi, *Sultan Qaboos University*

One of the commonly used method to estimate reliability of test scores is the split-half reliability. It requires dividing the assessment items to two parallel halves and compute two half scores for each examinee. All of the existing methods for the split-half reliability estimate the internal consistency reliability for the total scores on the assessment. Until now, there is no existing way to estimate the split-half reliability for test scale scores resulting from any nonlinear transformation of raw scores (e.g., percentile ranks, rounding or truncated linear scale scores, normalized standard scores and others).

Recently, Almehrzi (in press) outlined a method to obtain coefficient Generalized Alpha for test scale scores. The method requires obtaining the conditional distribution of raw scores given the matrix of observed proportions of all possible item scores under the assumption of independent item responses and uncorrelated error scores. Using Guttman's (1945) formula to obtain split-half reliability which is equivalent to using coefficient Alpha for the two split-halves (parts), coefficient Generalized Alpha is outlined to estimate the corresponding split-half reliability for different test scale scores with different ways for splitting the two halves. In addition, the research examine whether the fact that coefficient Alpha is the mean for all possible split-halves reliability is applicable to the coefficient Generalized Alpha.

On Cronbach's alpha as the mean of all split-half reliabilities

Matthijs J. Warrens, *Leiden University*

Coefficient alpha (Cronbach, 1951) is the most commonly used statistic for estimating reliability of a test if there is only one test administration (Cortina, 1993; Raju, 1977; Sijtsma, 2009). A famous description of alpha is that it is the mean of all (Flanagan-Rulon) split-half reliabilities. This is an important result, since it provides a proper interpretation of alpha. The result is exact if the test is split into two halves that are equal in size.

In this talk we consider how alpha is related to the mean of all (Flanagan-Rulon) split-half reliabilities when the number of items is odd. Here, the split-half is made so that the group sizes are as similar as possible, that is, one half contains one more item. It turns out that the difference between alpha and the mean of all split-half reliabilities is less than 0.01 if the test consists of at least eleven items. We conclude that, given a moderate number of items alpha is approximately identical to the mean of all (Flanagan-Rulon) split-half reliabilities.

Parkzaal

H.2.-IRT modeling

Thurstonian item response theory and an application to attitude items

Edward H. Ip, *Wake Forest School of Medicine*

The assessment of attitude has a long history dating back at least to the work of Thurstone. The Thurstonian approach had its "golden days," but today it is seldom used, partly because judges are needed to assess the location of an item, and also because of the emergence of contemporary tools such as the IRT. The current work is motivated by a study that assesses medical students' attitudes toward obese patients. During the item development phase, the study team discovered that there were items on which the team members could not agree with regard to whether they represented positive or negative attitudes. Subsequently, a panel of n = 201 judges from the medical profession were recruited to rate the items, and responses to the items were collected from a sample of n = 103 third-year medical students. In the current work, a new methodology is proposed to extend the IRT for scoring student responses. An affine transformation maps the judges' scale onto the IRT scale. Methodologically, the approach blends two latent-variable models—a continuous-response-factor-analytic model for the judges' ratings and an IRT model for the dichotomous students' responses. The model also takes into account measurement errors in the judges' ratings.

Archimedean IRT: latent trait models or copula-based marginal models

Johan Braeken, *Tilburg University*

Archimedean item response models are characterized by a skewed and absolute positive latent trait and item response functions that take the form of power laws.
At the same time, Archimedean item response models can be seen as instances of copula-based marginal models for multivariate discrete data.
This connection provides additional insight in the link between copula functions and latent variable models, and the uniqueness of such models in the discrete case.

Identification in Models for Pairwise Preference Data: Relating the Multi-Unidimensional Pairwise Preference Model and the Thurstonian IRT Model

Vicente Ponsoda, *Universidad Autónoma de Madrid*
Iwin Leenen, *Universidad Nacional Autónoma de México*
Jimmy de la Torre, *Rutgers University*
Daniel Morillo, *Universidad Autónoma de Madrid*
Pedro Hontangas, *Universidad de Valencia*

By virtue of their potential to reduce different types of response biases, pairwise preference items are increasingly considered as an alternative over the popular Likert-type items for the measurement of personality traits, motivation, and attitudes. Recent developments in IRT modeling, which allow to extract normative information from pairwise preference data (contrary to traditional scoring which allows for intra-individual comparisons only), have played a key role in this trend, in particular, the Multi-Unidimensional Pairwise Preference (MUPP) model (Stark, Chernyshenko, Drasgow, 2005, Applied Psychological Measurement) and the Thurstonian IRT (TIRT) model (Brown & Maydeu-Olivares, 2011, Educational and Psychological Measurement). Our research group has recently shown that replacing the ideal point model for single stimuli in the MUPP framework by a dominance (viz., the 2PL) model, comes down to a reparametrization of the TIRT model. Based on the implied mathematical equivalence, this paper focuses on identification issues (i.e., indeterminacies) in both models. In particular, it is explored to what extent constraints applied in the framework of TIRT relate to MUPP models. Such a comparison may have implications for the extension of the MUPP framework to forced-choice items that contain more than two components/stimuli.

An option-based partial credit IRT model for multiple-choice tests

Yuanchao Bo, Charles Lewis & David V. Budescu
Fordham University

Multiple-choice (MC) tests have been criticized for allowing guessing and the failure to credit partial knowledge, and alternative scoring methods and response formats (Ben Simon, Budescu, Nevo, 1997) have been proposed to address this problem. Modern

test theory addresses these issues by using binary models (e.g., 3PL) with guessing parameters, or polytomous IRT models.
We propose an option-based partial credit IRT model and a new scoring rule based on a weighted Hamming distance between the option key and the option response vector. The test taker (TT)'s estimated ability is based on information from both correct options and distracters. These modifications reduce the TT's ability to guess and credit the TT's partial knowledge. The new model can be tailored to different formats, and some popular IRT models can be presented as special cases.
A simulation study shows that the weighted Hamming distance outperforms others scoring rules: It has the highest correlation with TTs' true ability and their distribution is also less skewed than those for the other scores. Markov Chain Monte Carlo (MCMC) analysis was used to recover the model parameters. Root mean squared errors for the model parameter estimates and the ability estimates are small.

Jubileumzaal **H.3.-Measurement of change and growth**

A Bayesian Multi Level Model for Change
Rivka de Vries, *University of Groningen*

In recent decades, the focus on evaluation of mental health services has risen enormously in several countries. This means that clinicians now need to collect data from their clients, keeping track of how clients change over time, often known as Routine Outcome Measurement. As the observed changes in scores typically consist of a mixture of true change and random measurement error, just reporting the observed changes in scores over time can be misleading. Hence a variety of statistical inferential techniques have been developed to disentangle true change in clients from random measurement error. However, none of them allows the formal evaluation of evidence in the data for or against the hypothesis that true change has occurred. In this paper we develop a Bayesian model that quantifies the strength of evidence for the amount of individual change, the amount of average group change, and the number of clients that truly changed, assuming that all clients come from the same population. In a simulation study we present some of the properties of the model. We believe the model will be of great use in the field of Routine Outcome Measurement, allowing the quantification of evidence for several mental health services.

A two-level hierarchical model for examining baseline x treatment interactions in randomized trials.
Roger E. Millsap & Jenn-Yun Tein
Arizona State University

In randomized field trials of mental health interventions, a common finding is that intervention effects depend on baseline status, or that effects vary depending on the severity of the condition being treated. As a result, in an ANCOVA with baseline status as the covariate, a treatment by baseline interaction will be found. The research question then becomes: when does the treatment become effective given baseline status? The answer to this question will help determine who should receive the treatment in practice. Within the ANCOVA, the Johnson-Neyman method provides one way of addressing this question using simultaneous confidence bounds for the treatment effect across the range of the covariate. This method does not always provide useful results however. We present an alternative approach using a two-level hierarchical model. The method is illustrated using real data from a randomized field trial of a preventive intervention. Limitations and extensions of the method will be discussed.

Multilevel Poisson Dynamic Modeling
Tanja Krone, Casper J. Albers & Marieke E. Timmerman
University of Groningen

The Bayesian Dynamic Generalised Linear Model (DGLM) is a flexible model capable of working with time series data with non-normal error distributions. We will define a multilevel time series Poisson DGLM, including covariates, and outline modelling and fitting strategies. Its usefulness will be illustrated with data from a randomized clinical trial to examine the rate of improvement during and across three treatments for Panic Disorder (Cognitive Behavioral Therapy (CBT), Selective Serotonin Reuptake Inhibitor (SSRI) or both combined (CBT+SSRI)). The dependent variable is the weekly panic attack frequency, which was monitored for one year. The multilevel structure of the model allows us to study both between-treatment differences as within-treatment differences. We will study whether within-treatment differences relate to clinically relevant covariates as the presence of axis 1 and 2 disorders, and agoraphobia.

Item response growth modeling: a new application of multilevel factor model for longitudinal item response data

Zhen Li, *University of California*
Ji Seung Yang, *University of California, Los Angeles*
Li Cai, *University of California, Los Angeles*

An item response growth model based on multilevel item factor analysis is proposed in this research. The model uses between (level-2) factors to parameterize individual differences in growth parameters, e.g. individual initial status and growth rate, and within (level-1) factors to capture item-level properties and residual variances. The level-2 factors mean vector and covariance matrix provides averages and associations among the individual growth parameters. All item parameters and structural parameters are estimated simultaneously, taking measurement error into account. A simulation study shows that key parameters can be recovered properly for both dichotomous and polytomous item response data. The precision of estimation improves as sample size increases. As an illustration, empirical data from Drug Abuse Treatment Outcomes Study (DATOS) were analyzed. The model can be efficiently estimated by full information maximum likelihood utilizing currently available IRT software that adopts dimension reduction techniques (Gibbons & Hedeker, 1992).

Balkonzaal **H.4.-Understanding the individual: modeling dynamics & development**

Detection of Differential Development

Matthieu Brinkhuis, *Cito Institute for Educational Measurement*

The amount of data available in the context of educational measurement vastly increased in recent years. Such data is often incomplete, involves tests administered at different time points and during the course of many years, and can therefore be quite challenging to model. In addition, intermediate results like grades or report cards being available to pupils, teachers, parents and policy makers likely have an influence on performance, which adds to the modeling difficulties. We propose the use of simple data filters to obtain a reduced set of relevant data, which allows for simple checks on the relative development of persons, items, or both.

Discovering the dynamics of mental disorders

Claudia van Borkulo, *University of Amsterdam*
Abe Hofman, *University of Amsterdam*
Noemi Schuurman, *University Utrecht*
Silvia Rietdijk, *University Utrecht*

Stationarity and equidistance of measurements are assumptions that are often violated when modeling individual dynamics of continuous-time processes like mental disorders. We constructed a model that takes unobserved events into account and allows for non-equidistant intervals between measurements. The model is based on viewing the dynamics of a mental disorder as the spread of a virus about the nodes of a network. In this contact process, the topology of the network plays an important role;

an infected symptom can only infect neighboring symptoms. We estimated the infection and recovery parameters of the partially observed contact process and investigated the quality of the estimators. Moreover, we derived a likelihood ratio test and a t-test. In this presentation, I will show that our results indicate that the quality of the estimators is reasonable to good. Moreover, it seems possible to differentiate between a patient and a control with the t-test. To conclude, this is a promising model that might prove to be useful in clinical practice and can be used to make predictions about the development of the disorder in the long run.

Standardizing multilevel multivariate autoregressive models to compare the relative strength of directed (lagged) associations

N.K. Schuurman, *Utrecht University*

By modeling variables over time it is possible to investigate directed reciprocal associations between these variables. An association between variables over time is a precondition for causality, and as such indicates a potential causal relationship between these variables. Comparing the strength of their directed associations can provide direction for studying this possible causal relationship more in depth. Multilevel multivariate autoregressive models are especially suited for investigating these associations, and can provide estimates for the directed associations for individual subjects, and for groups of subjects as a whole. The relative strength of the directed associations can be evaluated by comparing the standardized estimated coefficients that reflect the directed associations.
However, the hierarchical structure of multilevel models complicates the technical aspect of standardization, and the interpretation of resulting standardized coefficients (Nezlek, 2001). For instance, both subject-based statistics or group-based statistics can be used to standardize the coefficients, but each standardization results in potentially different conclusions about the relative strength of the directed associations. We will discuss several issues concerning the standardization of multilevel multivariate autoregressive models aimed at comparing directed associations, including the interpretation of the standardized coefficients using group- or subject-based statistics, illustrated with an applied example.

Multilevel Latent Markov Models for studying transitions over time. A general framework for application and interpretation.

Silvia Rietdijk & Ellen L. Hamaker
Utrecht University

In the social sciences, longitudinal data are valued for providing a way of assessing both the differences between individuals, and the variability within individuals over time. Unfortunately, many widespread modeling approaches are limited to continuous outcome variables or to data with a small number of measurement occasions. Thus, psychology could benefit from a flexible modeling approach for multilevel longitudinal data, that is especially suited for categorical data and can handle a large number of measurements. It is argued that Multilevel Latent Markov Models can serve to fill this gap.
Latent Markov Models and related techniques have been applied in several fields but are unfamiliar to many psychologists. In this presentation, we will look at the general structure and interpretation of the model, uniting the terminology and developments from different scientific fields into a general framework. Using illustrative examples from psychology, I will emphasize that LMMs can be used for different purposes, depending on the researcher's theoretical assumptions, and it is crucial that the interpretation of the model varies accordingly. The presentation will also address the practical issues of model estimation and model comparison, illustrated by applications to empirical psychological data.

Stadszaal **H.5.-SEM 3**

Causal Mediation Analysis in Behavioral Experiments: Addressing Omitted Variables and Measurement Error

Walter Herzog, *WHU-Otto Beisheim School of Management*
Anne Boomsma, *University of Groningen*

In many experimental studies, researchers manipulate X to estimate its effect on Y and to test whether the effect is mediated via M. Regression or structural equation models are frequently used for this purpose. Although randomization of X guarantees unbiasedness of the effects of X on both M and Y , the indirect effect of X on Y via M and the direct effect of X on Y are biased in the presence of (a) omitted mediators, (b) omitted confounders that affect both M and Y , (c) measurement error in M, and (d) omitted method factors (i.e., common method variance). In the present paper, we first argue that complications (a)-(d) are usually present in experimental studies and that the resulting biases are largely neglected in applied research. Second, we show how 2x2 between-subjects experiments can be combined with the idea of instrumental variables to create unbiased estimates of direct and indirect effects in the presence of complications (a)-(d). Third, we argue that the assumptions underlying our estimation strategy are likely to hold in practice. Finally, the results of simulation studies suggest that the proposed estimation strategy produces acceptable results for the sample and effect sizes usually encountered in experimental research.

Mediation Analysis for Ordinal Outcome Variables

Fang Luo, *Beijing Normal University*
Hongyun Liu, *Beijing Normal University*
Shanshan Zhang, *Capital university of economics and business*

Based on the methods of MacKinnon (1993, 2007), we focused on the sensible use of categorical data analysis rather than continuous data analysis for exploring mediation model when outcome variable is binary or ordinal. Sample size, true mediation level and the number of categories for outcome variable was taken into account for comparison of the two analysis method. The main results indicated as follows. For ordinal outcome mediation model, there was always low estimation precision, poor statistical test and SE underestimated when forced to use continuous data analysis. On the contrary, estimation results from logistic mediation model always performed better than the continuous data analysis methods among some comparative indicators. The method of product of coefficient always performed better than the method of difference of coefficient whether or not ordinal data analysis was used. As the number of categories increased, RMSE and estimated standard error of mediation effect derived from continuous data analysis showed decreases, the statistical power increased.

Using the pair-wise likelihood method to analyze discrete responses

M.T. Barendse, *University of Groningen*

Using factor analytic methods, discrete responses data can be analyzed by assuming that each observed variable is a manifestation of an underlying normally distributed variable. Because maximizing the likelihood of all observed response patterns is computationally very intensive, alternatives have been proposed, like weighted least square methods based on the polychoric correlations. Another alternative is based on the bivariate (i.e., pair-wise) distributions given in two-way contingency tables. When it comes to assessing model fit, there is very little known about this pair-wise method. In this presentation we introduce new fit criteria for the pair-wise method. We compare the performance of different estimation methods and several fit criteria in a simulation study. We provide recommendations for the use of fit criteria in practice.

Hemelzaal **H.6.-Computerized Adaptive Testing**

Multidimensional Computerized Adaptive Testing for classifying examinees

M.M. van Groen, *Cito / RCEC*
Theo Eggen, *Cito / Twente University*

Computerized adaptive tests (CATs) were developed for obtaining an efficient estimate of the examinee's ability but they can also be used for classifying the examinee into one of two levels (e.g. master/non-master). Several methods are available for making the classification decisions for constructs that can be modeled with an unidimensional item response theory model. These methods stop the test when enough confidence has been reached for making the decision. But if the construct is considered to be multidimensional, no classification method is available. A classification method based on Wald's Sequential Probability Ratio Test was developed for application to CAT in conjunction with a compensatory multidimensional item response theory model. This method tries to maximize the classification accuracy as well as minimizing the test length.
Simulation studies were used to investigate the average test length and the percentage of correct decisions by the classification method. Comparisons of the efficiency and the accuracy were made between different item selection methods and between different settings for the classification method. The impact of exposure control as well as content control on the average test length and the percentage of correct decisions was also investigated in the simulation studies.

Developing Online Calibration Methods for Multidimensional Computerized Adaptive Testing

Ping Chen & Tao Xin
Beijing Normal University

Multidimensional computerized adaptive testing (MCAT) features a combination of tailored testing and multi-trait estimation which shows great potential to support formative assessments (Wang & Chang, 2011). Just like cognitive diagnostic computerized adaptive testing (CD-CAT) and regular CAT, item replenishing is essential for item bank maintenance and management in MCAT. Calibration of new items is a technical difficulty in item replenishing, and the precision of calibration directly impacts the accuracy of the estimation of examinees' abilities. In CD-CAT and regular CAT, online calibration is commonly employed to calibrate new items (Ban, Hanson, Wang, Yi, & Harris, 2001; Chen, Xin, Wang, & Chang, 2011). However, until now no reference about online calibration is publicly available in the realm of MCAT. Thus, this study extends some current methods used in CAT to MCAT under the multidimensional 2-parameter logistic model (Reckase, 2009). Three representative methods, Method A (Stocking, 1988), MMLE with one EM cycle (OEM) method (Wainer & Mislevy, 1990) and MMLE with multiple EM cycles (MEM) method (Ban et al, 2001), are generalized to MCAT context and the new methods are referred to as M-Method A, M-OEM and M-MEM respectively. Two simulation studies were carried out to compare the performance of the three methods in terms of item parameter recovery under different sample size conditions. The simulation results showed that all three methods were able to recover the item parameters accurately, and M-Method A outperformed the other two methods in that it yielded the smallest estimation errors and the adaptive calibration design (new items were adaptively selected following the logic of MCAT) improved the item parameter recovery compared with the random calibration design (new items were randomly selected).

Computerized Adaptive Testing under a Multidimensional Unfolding IRT Model

Shiu-Lien Wu, *National Chung Cheng University*
Wen-Chung Wang, *The Hong Kong Institute of Education*

Unfolding IRT models have been developed in recent years, including the confirmatory multidimensional generalized graded unfolding model (CMGGUM; Wu & Wang, 2011). In this study, we implemented computerized adaptive testing algorithms based on the CMGGUM. The Fisher information was derived. Simulations were conducted to evaluate the performance of the algorithms, including the maximum a posteriori estimation for ability estimation, the maximum priority index (MPI) and D-, A-, T-, E-optimality criteria for item selection, and a fixed-precision stopping rule. Results showed that all the four criteria achieved the predetermined precision level of .25, the bias in ability estimation was between -0.017 and 0.019, the mean square error was between 0.064 and 0.114, and the correlations between estimated and true latent traits was between .94 and .96. The mean test length for each dimension was similar, indicating that the MPI would facilitate content balance. Among the four optimality criteria, the E-optimality criterion was not recommended due to its longer tests and consuming time. The other three optimality criteria performed similarly. In addition, the differences of the Fisher information between unfolding and dominance IRT models were described.

An Item-driven Adaptive Design for Selecting Pretest Items in Multistage Testing

Usama S. Ali & Longjuan Liang
Educational Testing Service

Large-scale adaptive testing programs require precalibrating many new items, and therefore, it is crucial to be able to calibrate test items in large quantities efficiently and economically. Few studies (e.g., Kingsbury, 2009; Makransky, 2009) have used adaptivity to improve item calibration in the context of item-level adaptive testing. To the extent of our knowledge, there is no published work that investigates the adaptive assembly and delivery of pretest item bundles (or modules) in MST; which constitutes the objective of our study. An adaptive pretest design is developed to fit the MST settings. In this study, we propose a suitability index that combines the information of examinees' ability, item difficulty, and the distance from a target sample distribution to select items to administer to examinees. The proposed method has potential benefit of requiring smaller sample size to reach the same standard error of item parameter estimates. A simulation study using a MST1-3 (i.e., two stages, with one section in the first stage and three sections in the second stage) is performed to evaluate the performance of the proposed design compared to a design where items are randomly assigned to examinees. Multiple-cycle EM procedure is used for estimating item parameters. Sample size and item parameters recovery through bias and RMSE are used as evaluation criteria. Results, discussion, and future research are reported.

Friday July 26

08:30 – 17:00 Registration and Information Desk Open

08:30 - 09:50 **Parallel session I**

Concertzaal **I.1.-Ordinal inference and latent variable models**

Using the Sum of Item Scores to Order Respondents on the Latent Trait
L. Andries van der Ark, *Tilburg University*

For the overwhelming majority of tests and questionnaires used in the social and behavioral sciences, researchers use the unweighed sum of item scores of a respondent to measure the construct of interest. It is important to understand the conditions in which this practice is reasonable. In the last three decades, Psychometrika has published a dozen papers that address the question whether the unweighed sum of item scores can be used to stochastically order respondents on the latent trait. Firstly, I will give an overview of the results that have been obtained for both dichotomous and polytomous items. Secondly, I discuss the implications of these results for practical testing and challenges for the future.

On the Use of Sum Scores for Ordinal Inferences from IRT Models
Robert Zwitser & Gunter Maris
Cito Institute for Educational Measurement

For the evaluation of ability with a simple sum or number correct score, common IRT models are not fully satisfactory. Parametric models are sometimes considered as too restrictive, while with nonparametric models not all available information about the parameter is included in the sum score. To arrive at a fair evaluation of ability with a simple number correct score, ordinal sufficiency is defined as a minimum condition for scoring. It is shown that for the Rasch Model, as well as for some 2PLMs, the sum score is an ordinal sufficient statistic. The monotone homogeneity model, together with the property of ordinal sufficiency of the sum score, is introduced as the nonparametric Rasch Model (npRM). A basic outline for testable hypotheses about ordinal sufficiency, as well as illustrations with real data, are provided.

An inequality for correlations in unidimensional monotone latent variable models for binary variables
Jules L. Ellis, *Radboud University Nijmegen*

A unidimensional monotone latent variable model for binary items implies a restriction on the relative sizes of item correlations: The negative logarithm of the correlations satisfies the triangle inequality. This inequality is not implied by the condition that the correlations are nonnegative, the criterion that coefficient H exceeds .30, or manifest monotonicity. The inequality implies both a lower bound and an upper bound for each correlation between two items, based on the correlations of those two items with every possible third item. It is discussed how this can be used in Mokken’s (1971) scale analysis.

An ordinal approach to measurement invariance
Rudy Ligtvoet, *University of Amsterdam*

Measurement invariance (MI) analysis consists of fitting one model to data from different groups and imposing equality restrictions on the model parameters across the groups. By considering a set of inequality restrictions for the data from different groups, a weak version of MI can be defined for a wide family of models. The set of inequality restrictions considered in this talk allow for an initial test of MI for models that include the partial credit model, the graded response model, and a variety of multi-dimensional factor models. In contrast to traditional MI analysis, this initial test does not require any of these models to be explicitly specified prior to the analysis.

Parkzaal **I.2.-SEM 4**

On the Importance of the Skewness Parameter in Modeling Latent Traits

Iris A. M. Smits, Marieke E. Timmerman & Alwin Stegeman
University of Groningen

Standard linear factor models assume normally distributed item scores. Deviations from normality due to non-normality of the expected item scores can be modeled through a skew-normally distributed factor, or through level dependent factor loadings (Molenaar, Dolan, & Verhelst, 2010). We show that these two models are equivalent to the third moment of the distributions of the expected item scores. This implies that they are empirically indistinguishable. We explain that this connotes the importance of including a skewness parameter in the latent trait for the establishment of measurement invariance across populations.

Limited Information Goodness-of-Fit Testing for Structural Models of Categorical Data

Scott Monroe & Li Cai
UCLA

Typically, popular software use multistage estimation procedures to fit structural models of categorical data. First, sample thresholds and polychoric correlations are estimated from the observed contingency table. Then, the structural parameters are estimated by minimizing some form of least squares. Thereafter, a test of the model to the polychoric correlations can be obtained by adjusting the minimum fit function value, as in Muthén (1993) or Satorra and Bentler (1994). Maydeu-Olivares (2006) developed an overall test of the model to the contingency table, as well as separate distributional and structural tests. In this research, the work of Maydeu-Olivares (2006) is extended by applying the M2 methodology of Maydeu- Olivares and Joe (2006). This results in a chi-square distributed test statistic of overall fit for structural models with categorical data. A simulation study is conducted to assess the performance of the M2-type statistic, and compare its performance to that of the standard moment-corrected tests. Finally, the new statistics are illustrated through empirical applications.

Measurement invariance with respect to unmeasured level 2 variables in multilevel SEM

Suzanne Jak, Frans Oort & Conor Dolan
University of Amsterdam

Cluster bias refers to measurement bias with respect to the clustering variable in multilevel data. Cluster bias can be investigated using two-level factor models, by constraining the factor loadings to be equal across levels, and testing the absence of residual variance at the cluster level (level 2). It is suggested that absence of cluster bias implies absence of bias with respect to any level 2 variable (Jak, Oort & Dolan, 2012a). Although this is true in the population, in practice it depends on the power of the test for cluster bias. In this study we will evaluate the tenability of the claim that absence of cluster bias implies absence of bias with respect to all level 2 variables using simulated data. In the simulation study, bias will be introduced by an observed level 2 variable, under varying conditions. Results of the test for cluster bias will be compared with results obtained by applying the RFA model (Oort, 1992) to detect the bias. If the power of the test for cluster bias is higher than the power of the RFA analysis, this indicates that if no cluster bias is found, no bias with respect to any level 2 variable will be found.

Testing measurement non-invariance in multilevel data with a within-level violator

Yao Wen, *University of Wisconsin Milwaukee*
Wen Luo, *University of Wisconsin Milwaukee*
Eun Sook Kim, *University of South Florida*
Oi-Man Kwok, *Texas A&M University*

Establishing measurement non-invariance is crucial when we want to compare trait-level scores among groups. However, the task is challenging when the data have multilevel structure and the potential violator is at the within-cluster level. Empirical researchers are often concerned about (1) the appropriate model to use (i.e., multilevel MIMIC models or multilevel mixture factor models), (2) the order of establishing the within-level and the between-level non-invariance, (3) the best indices to use for model selection (e.g., Satorra-Bentler likelihood ratio, AIC, BIC, and ssBIC), and (4) the correct interpretation of the results. The extant literature on this topic is scattered and inadequate. The majority of the methodological investigations only tackle one aspect of the problems, with findings that are not generalizable to more complicated situations in reality. In this study, we are going to review the existing methodological literature, propose an integrated 4-step procedure to examine the measurement invariance for multilevel data when the potential violator is at the within-cluster level, and demonstrate the proposed procedure using the data set from the Early Childhood Longitudinal Study. The significance of study is that it will provide empirical researchers with a big picture of testing measurement non-invariance in multilevel data, clear step-by-step procedures, and practical guidelines for interpreting the results.

Jubileumzaal **I.3.-Testing and measurement invariance**

The Multilevel First-Order Autoregressive Model: Testing Hypotheses Regarding Inter-Individual Variability.

Joran Jongerling, *Utrecht University*

This study discusses a multilevel expansion of the first order autoregressive (AR(1)) model, that includes a random mean, random AR-parameter, and random innovation variance. The aim is to determine the ability of different model selection methods to detect whether a model parameter is random or fixed. Specifically, we use an extensive simulation study to investigate two types of hypotheses regarding inter-individual variability, with two different model selection methods. With the DIC we test whether the amount of inter-individual variance in the model parameters is larger than zero. With Bayes Factors (BF) we test whether this amount of inter-individual variability is larger than a certain boundary value B. This second type of hypothesis is included because it is likely that, if the contribution of a parameter to the total variance of the timeseries is smaller than a certain cut-off value, the parameter can be modeled as fixed, even if its amount of inter-individual variance is not equal to 0. We therefore derive a decomposition of the total variance of a multilevel timeseries, and determine the contribution of the different model parameters to this total variance. We then present theoretical and statistical argumention for different cut-off values for the different parameter contributions.

Bayes Factor tests for measurement invariance

Josine Verhagen, *Universiteit van Amsterdam*
Jean Paul Fox, *University of Twente*

In the investigation of measurement invariance, the evidence in favor of the null hypothesis of invariance is of main interest. Using a Bayesian approach to hypothesis testing enables researchers to investigate exactly this, through the quantification of evidence in favor of invariance by a Bayes factor. The presented Bayes factors for nested models allow the investigation of invariance of item parameters across many

groups and for many items simultaneously, without the need for one or more invariant anchor items. Bayes factor tests for the variance of item parameters over groups or measurement occasions and for the difference of item parameters between specific groups will be compared to recent frequentist measurement invariance tests. Examples from longitudinal and cross-national research will illustrate the applicability in practice.

Balkonzaal **I.4.-Cognitive diagnostic assessment 1**

Probabilistic Reasoning: When Does Prediction With Baserates Outperform Prediction With a Diagnostic Test?
Ehsan Bokhari, *University of Illinois at Urbana-Champaign*

Some fifty years ago, Meehl and Rosen (1955) stated a condition under which a diagnostic test would be “clinically efficient.” As they defined it, clinical efficiency refers to a situation where prediction by a diagnostic test is better than prediction using only the raw baserates. Although cited extensively, the actual importance of the Meehl and Rosen condition for deciding on when to use a diagnostic instrument seems generally ignored in the literature. The presentation reviews the Meehl and Rosen condition and offers two equivalent others attributed to Dawes (1962) and Bokhari and Hubert. The relationships are detailed between those various equivalent conditions for clinical efficiency and several measures of association in a 2 × 2 contingency table (e.g., the Goodman-Kruskal lambda coefficient). The condition is demonstrated using the Classification of Violence Risk assessment tool.

Automatic model selection for three conjunctive diagnostic models
Lihong Song, Wenyi Wang, Haiqi Dai &Shuliang Ding
Jiangxi Normal University

Cognitive diagnostic assessment (CDA) is designed to measure specific knowledge structures and processing skills in students so as to provide information about their cognitive strengths and weaknesses. A large number of cognitive diagnostic models (CDMs) are developed and based on different cognitive assumptions about how skill interaction (disjunctive vs. conjunctive) influence item performance. Unfortunately, little is known about the relationship between attributes and item response in most testing situations. This challenges the researcher to make a conscious thought model selection before data analysis. In some circumstances, at least from a cognitive perspective, a conjunctive model appears to be more appropriate for the cognitive assumptions. This proposal introduced the reversible jump Markov Chain Monte Carlo (RJMCMC) method for the determination of three conjunctive CDMs that based on different assumptions. Firstly, three conjunctive CDMs were described briefly. Secondly, the consequence of the misspecification of models is examined through simulation studies. Thirdly, the algorithm of RJMCMC for automatic model selection was established. Finally, a simulation study and an analysis of real data were presented to verify the algorithm. The simulation and the real data analysis results demonstrated that the model selection algorithm of RJMCMC can work well among three models.

Balancing test efficiency with item bank usage efficiency: Development of item selection strategies for cognitive diagnosis computerized adaptive testing
Wenyi Wang, Lihong Song & Shuliang Ding
Jiangxi Normal University

This proposal proposes two new item selection methods for Cognitive diagnostic computerized adapting testing (CD–CAT): the randomization halving algorithm (R-HA) and Kullback–Leibler expected discrimination method (KL-ED). Two simulation studies are carried out, one using a simulated item bank, and the other based on items calibrated from real data. Compared to item selection based upon random, Kullback–

Leibler (KL) , Shannon entropy (SHE) (Xu, Chang, & Douglas, 2003), Posterior–Weighted KL (PWKL) , Hybrid KL algorithm (HKL) (Cheng, 2009), restrictive progressive method (RP–PWKL) and restrictive threshold method (RT–PWKL）(Wang, Chang, & Huebner, 2011), expected discrimination method (ED) (Shang & ding, 2011), halving algorithm (HA) (C. Tatsuoka, 2004) . The simulation results show that: (1) RHA, HA and RP–PWKL strike a better balance between test efficiency and item bank usage efficiency for simulated item bank, while RHA performs slighter better others under real data analysis; (2) KL-ED can improve considerably the efficiency of testing. The cognitive diagnostic model used here is the “Deterministic Input; Noisy ‘And’ Gate”(DINA) model. Though the results from the simulation study are encouraging, further studies of CD-CAT are proposed for the future investigations such as different cognitive diagnostic models.

Stadszaal **I.5.-Forced choice**

Item Response Modelling of Forced Choices: A Unified Framework

Anna Brown, *University of Kent*

To counter response distortions associated with the use of rating scales, test items may be presented in so-called ‘forced-choice’ formats, whereby respondents are asked to rank-order a number of items, or distribute a fixed number of points between several items. Until recently, basic classical scoring methods were applied to this kind of data, leading to scores relative to the person’s mean (ipsative scores). Recent advances in estimation methods enabled rapid development of IRT models for comparative data, including the Multi-Unidimensional Pairwise Preference Model (Stark, Chernyshenko & Drasgow, 2005), and Thurstonian IRT model (Brown & Maydeu-Olivares, 2011).

As these models rely on different traditions, it is hard for researchers to compare their properties. In the present talk, a unified framework for modelling forced-choice data is presented. Three distinct elements are differentiated, namely, 1) the design and structure of the forced-choice task, 2) the preference decision process leading to selection of items, and 3) the model describing relationships between items and constructs they measure (i.e. dominance or ideal-point models). It is shown that the proposed framework provides a common language for describing and comparing a variety of forced-choice models, and that the existing IRT approaches have more similarities than differences.

Computerized Adaptive Testing under the Rasch model for Ipsative Forced-Choice Items

Chia-Wen Chen & Wen-Chung Wang

Hong Kong Institute of Education

Forced-choice (pairwise comparison) items have been widely used in personality and attitude tests, such as the Edwards Personal Preference Schedule and Gordon Personal Profile-Inventory. In these tests, respondents are requested to select one statement from a pair of statements, which makes the tests become ipsative (self-comparison). The analysis of ipsative tests within the IRT framework has recently attached research attention. The Rasch model for ipsative forced-choice items is especially promising because of its good measurement properties (Wang & Chen, 2013). Ipsative tests often involve many latent traits and many items. Computerized adaptive testing (CAT) is thus especially useful for such tests. In this study, we developed CAT algorithms under this model in order to increase its feasibility. We developed several item selection procedures and conducted a series of simulations to evaluate their performance. The simulation results showed that proposed information-stratified method and the progressive method outperformed the random selection method and the maximum information method, in terms of the absolute bias and the correlation between the true and estimated ability. As expected, the random selection method had the best item exposure control.

Relationship between IRT models for Forced-Choice items

Daniel Morillo Cuadrado, *Universidad Autónoma de Madrid*
Vicente Ponsoda Gil, *Universidad Autónoma de Madrid*
Iwin Leenen, *Universidad Nacional Autónoma de México*
Francisco José Abad García, *Universidad Autónoma de Madrid*
Jimmy de la Torre, *Rutgers University*

Forced-choice questionnaires have been proposed as a way to control faking and social desirability biases in personality measurement, but the ipsative data structures they yield make classical scoring methods inappropriate. IRT methods have been claimed to give account of the latent trait structure of these instruments more accurately. Stark, Chernyshenko, & Drasgow (2005, Applied Psychological Measurement) proposed the MUPP model, an ideal point model, while Brown & Maydeu-Olivares (2011, Educational and Psychological Measurement) introduced the TIRT model, a dominance model based upon the Thurstonian Law of Comparative Judgment. Given certain restrictions, it can be shown that these models are equivalent when a dominance rather than an ideal point function is specified for the MUPP model. An MCMC Bayesian estimation procedure, suitable for joint estimation of person and item parameters in these (and possibly more complex) models is introduced and compared to the TIRT model estimation method proposed by Brown & Maydeu-Olivares (2011). Results are discussed in terms of the relationship between both estimation procedures, issues affecting the estimation quality, identification of the model, and optimization of the algorithmic efficiency.

Towards creating forced-choice personality assessments 'on the fly': do Thurstonian IRT assumptions hold empirically?

Yin Lin, Ilke Inceoglu & Dave Bartram
SHL

In pre-employment non-cognitive assessments such as personality tests, three measurement trends pervaded the last few decades: the rise of IRT models; the migration towards computer-based dynamic test construction; and the adoption of the forced-choice format above rating scales. The forced-choice response format reduces response biases often associated with rating scales, and is becoming less controversial thanks to recently developed IRT models suitable for the associated comparative judgement response process. With many companies now using assessments to inform hiring decisions, the benefits of these psychometric advances are multiplied. However, it is not until recent years that these three offerings were successfully combined into practical applications (NCAPS, GPI-A, TAPAS). This study contributes to the same line of research by conducting a test-retest study using a forced-choice personality questionnaire (OPQ32r, SHL, 2012) and an alternative form. A sample of over 3000 individuals completed both forms with counter-balanced orders, and the responses were modelled using the multidimensional dominance Thurstonian IRT model (Brown & Maydeu-Olivares, 2011). Robustness of model parameter recovery and score recovery across forms provides confidence in the assumptions required to move towards dynamic forced-choice assessments using Thurstonian IRT as the measurement framework.

Hemelzaal **I.6.-Reliability 2**

An empirical assessment of Guttman's L4 reliability coefficient

Tom Benton, *Cambridge Assessment*

Numerous alternative indices for test reliability have been proposed as being superior to Cronbach's alpha. One such alternative is Guttman's L4 (Guttman, 1945). This is calculated by dividing the items in a test into two halves such that the covariance between scores on the two halves is as high as possible. However, although simple to understand and intuitively appealing, the method can potentially be severely positively biased if the sample size is small or the number of items in the test is large (Ten Berge and Socan, 2004).

To begin with this paper compares a number of available algorithms for calculating L4. We then empirically evaluate the bias of L4 for 51 separate upper secondary school examinations taken in the UK in June 2012. These tests each contain between 10 and 37 items and were each taken by a minimum of 5000 candidates. For each of these tests we have evaluated the likely bias of L4 for a range of different sample sizes. Using the results of this analysis this paper provides recommendations as to the required sample size for L4 to be essentially unbiased depending upon the number of items in the test and the estimated level of reliability.

Reliability and validity of weighted sum scores by multiple measuring methods: In a confirmative factor analysis model for multitrait-multimethod data

Saori Kubo, *Waseda University*
Hideki Toyoda, *Waseda University*
Kosuke Fukunaka, *The SANNO Institute of Management*

When a trait is measured by a number of methods, in many cases, the composite score, or any linear combination of scores obtained by different measuring methods, with fixed weights, is used as each individual’s score of the trait. The purpose of this study is to develop the expressions for coefficients of reliability and the convergent and discriminant validity of weighted sum scores by
multiple measuring methods, in a confirmative factor analysis model for multitrait-multimethod (MTMM) data. The decomposition of variance components is used to derive these coefficients.

The coefficients were calculated in actual 360-degree feedback data, which was a typical example of MTMM data, and reliability and the convergent and discriminant validity of weighted sum scores were examined. In 360-degree feedback, difference of raters corresponds to
“methods” of MTMM. The results of the analysis showed that both supervisor and peer rating contributed to the improvement of reliability and the validity of sum score; however, adding the weight of self rating decreased the quality of measurement. The practicality of our approach in examining the appropriate weights, taking into account the reliability and validity of sum scores was demonstrated through this example of application.

Revisiting Internal Consistency Reliability: Evaluating the Performance of New and Old Estimation Techniques

Tyler Hunt, *University of Utah*

For the last seventy years the conceptualization of reliability has changed dramatically. Likely because of the large number of methods for calculating internal consistency reliability, few studies have made a comparison of the strengths and weaknesses of all or even the majority of the estimators in simulation studies. Recently there has been confusion and disagreements in the literature on the most appropriate estimator of internal consistency reliability in a variety of situations. To asses these concerns, a simulation was designed with three different factor structures, parallel, tau-equivalent, and congeneric, with 1, 3, or 5 factors, and 5 different sample sizes, 50, 100, 400, 1000, 2000. For each simulation 1000 samples were generated and yielded 1000 estimates of Guttman's Lambda coefficients (including coefficient alpha), McDonald's Omega, Kristoff's coefficient, raju's Coefficient, Angoff's coefficient, the Greatest Lower Bound, Maximized Lambda 4, Quantile Lambda 4, and Covariance Maximized Lambda 4. For each estimator the variance, mean square error and expected value with be estimated. The focus of this investigation is to determine situations where it would be appropriate to use these estimators to assess internal consistency reliability.

09:55 - 10:55 Panel discussion

Concertzaal **Everything you always wanted to ask to senior people in the field**

In this session the audience can ask questions to the panel members on a broad range of topics related to psychometrics. There is no specific format or topic, so questions can be on any aspects of the field. Potential discussion topics are developments and history of the field of psychometrics and the role of psychometrics in society. But for example also questions related to career development will fit into the format. Junior staff is encouraged to send in questions prior to the meeting. But all members of the audience will have the opportunity to participate during the session.

Panel Members:
Paul de Boeck, *Ohio State University*
Alina von Davier, *Educational Testing Service*
Cees Glas, *Twente University*
Michael Kane, *Educational Testing Service*
Jimmy de la Torre, *Rutgers University*

Chair :
Xiang Wang, *College Board*

10:55 - 11:15 ***Break***
Ravelijn

11:15 - 12:35 Parallel session J

Concertzaal **J.1.-Invited symposium: Network Psychometrics**

The network as an alternative psychometric framework
Denny Borsboom, *University of Amsterdam*

In psychometric applications, test items are typically considered to be a function of a latent variable. In many applications, however, the assumptions and implications of the latent variable model are implausible; for example because the covariance between items is more likely to result from local interactions between directly measures factors (e.g., pain->lack of mobility) than from a common latent variable (e.g., pain<-quality of life->lack of mobility). In such cases, network models offer a fruitful alternative way of modeling the association patterns that exist among a set of items. In the present talk, I will outline some network modeling techniques that we have developed or tailored to psychometric applications, and discuss cases in which these efforts have proven useful. In addition, I will address some limitations of current modeling techniques and discuss options for future research.

Ising meets Rasch: A network perspective on IRT, and vice versa
Gunter Maris, *Cito / University of Amsterdam*
Denny Borsboom, *University of Amsterdam*

In this presentation we establish the formal relation between the class of Ising network models, from statistical physics, and the class of multi-dimensional Rasch models. Specifically, we demonstrate that an Ising model can be represented as a Rasch model, in which the latent trait figures as a mathematical construct(ion). This makes the class of Ising-Rasch models theoretically attractive, as it offers a novel interpretation of latent variables. Ising-Rasch models have the important advantage over other latent variable models that their marginal likelihood function is available in closed form, which has advantages for both estimating the model parameters and for the evaluation of model fit. Both of which will be demonstrated using a real data example. A final advantage of Ising-Rasch models is that it opens up a huge literature on the emergence and dynamics of complex systems to the field of educational measurement. Vice versa, much of the computational/statistical tools that have been developed for latent trait models become available to researchers interested in complex systems.

Major depression as a complex system

Angélique O.J. Cramer, *University of Amsterdam*

In this talk, contrary to a latent variable perspective on psychopathology, I argue that major depression should be characterized as a complex dynamical system in which symptoms (e.g., insomnia and fatigue) causally interact with one another: insomnia --> fatigue --> psychomotor retardation. Next, we hypothesize that individual people can be characterized by their own network with unique architecture and resulting dynamics. With respect to architecture, we argue that individuals vulnerable to developing major depression are those with strong connections between symptoms: e.g., only one night of poor sleep suffices to make a particular person feel tired. Such vulnerable networks, when pushed by external forces such as stressful life events, are more likely to end up in a depressed state; whereas networks with less strong connections tend to remain in or return to a healthy state. We show this with a simulation in which we model the probability of a symptom becoming ‘active’ in a person as a logistic function of the activity of its neighboring symptoms. Additionally, we show that this model explains some well-known empirical phenomena (e.g., spontaneous recovery) and accommodates both continuous and taxonomic views on major depression. Finally, we elaborate on how therapeutic strategies (e.g., cognitive behavioral therapy) can be understood within this causal systems perspective. To our knowledge, we offer the first intra-individual, symptom-based, process model with the potential to explain the empirical reality of major depression

Estimating large networks from ESM data

Francis Tuerlinckx, *University of Leuven*
Laura Bringmann, *University of Leuven*
Geert Verbeke, *University of Leuven*
Denny Borsboom, *University of Amsterdam*

With Experience Sampling Methodology (ESM) researchers aim at “capturing life as it is lived” (Bolger, et al., 2003). Typically in ESM studies, a sample of persons produces relatively short high-dimensional time series data (e.g., behaviors, symptoms or emotions). Such ESM data can be used to construct dynamic networks that shed light on the complex interplay between the measured variables. Recently, Bringmann et al. (2013) have used a multilevel vector autoregressive (ML-VAR) model to extract a network for a limited set of variables. The ML-VAR model results in a population network and in person-specific networks. However, the ML-VAR can only be applied to data with at most 10 variables and therefore, the task of inferring dynamical networks from truly high dimensional ESM data (e.g., 20 variables or more) remains an unsolved problem. In this talk we will discuss and compare several ways of estimating the ML-VAR model so that large dynamical networks can be constructed. A first method, based on the pseudo-likelihood (PL) theory, estimates repeatedly a ML-VAR model with a limited set of random effects and pools the estimates afterwards. Second, we will propose a method that applies a transformation to the data such that the model for the transformed data has less parameters and a less intricate structure. Third, we discuss a regularization approach in which a penalty is applied that drives small coefficients effectively to zero.

Parkzaal **J.2.-Cognitive diagnostic assessment 2**

A Diagnostic Model for Estimating Ability and Misconceptions as Discrete Latent Traits

Laine Bradshaw, *The University of Georgia*

Recent advances in psychometrics have focused on measuring multiple dimensions of ability to provide more detailed feedback for students, teachers, and other stakeholders. Diagnostic classification models (DCM) provide multidimensional feedback by using categorical latent variables that represent distinct skills underlying a test that students may or may not have mastered. The Scaling Individuals and Classifying Misconceptions (SICM) model is a nominal response psychometric model that was developed as a combination of a unidimensional IRT model and a DCM where the categorical latent variables represent misconceptions instead of skills. The present study alters the SICM model and assumes that not only misconceptions, but also the overall ability is a categorical latent variable. Through an empirical data analysis, this study will show how discrete feedback about examinee ability and misconceptions can be used (1) by researchers to examine the relationship of ability and misconception presence, and (2) by stakeholders to tailor instruction for students' needs.

Attributes, Model Equivalencies, Hierarchies, and Labels

Matthias von Davier, *ETS*

Cognitive diagnosis models are often used when it is hypothesized that multiple mastery type skills are involved in solving items on a test. The application of these models require a Q-matrix describing which observed variables depend on which skills, and how skills interact. Skill interactions are described in terms of compensatory, disjunctive, conjunctive, etc. functioning of skills. The matrix is either provide by experts or algorithms 'learn' the entries of the matrix. Recent research showed that there are issues with evaluating a model for cognitive diagnosis based on using a single Q-matrix. DeCarlo (2011) showed that for certain Q-matrices, not all attribute patterns can be identified. Von Davier (2011) showed by means of transformations of skill space that a conjunctive model (DINA) with a distinct interpretation (all required skills are needed to solve an item) can be re-expressed as a compensatory model (GDM) .

This talk addresses another set of issues, namely, which problems are introduced if variables are believed to stand in a hierarchical relationship. More specifically, it will be investigated what happens mathematically if skills or items are prerequisite to other skills or items.

How to design polytomous cognitive diagnostic test

Shuliang Ding, Wenyi Wang & Fen Luo
Jiangxi Normal University

It is well known that polytomous items provide more diagnostic information than dichotomous items. In cognitive diagnostic assessment, How to design cognitive diagnostic test is very important for diagnostic information feedback. If the test Q matrix can establish a bijection (i.e., one-to-one and onto mapping) from the set of knowledge states to the set of expected response patterns, the bijective property is benefit to improve the cognitive diagnostic accuracy.

For dichotomous items, if the cognitive attributes are not compensatory and conjunctive ,and if the reachability matrix is a sub-matrix of the test blueprint Q matrix whose rows represent the attributes, then the Q matrix has the bijective property.

For the polytomous items, if the expected response score equals to the production of the knowledge state by the test Q matrix, and if the test Q matrix contains the reachability matrix, the bijective property can also be established. However, this requirement is a sufficient condition only, there is a counter example to show it is not a necessary condition, yet this sufficient condition could improve the diagnostic accuracy. From the linear algebra point of view, we proved one essential statement that the bijective property can be established if the rank of Q matrix is equal to the number of attributes.

Facilitating standard setting with cognitive diagnostic analysis

Tao Xin & Jiahui Zhang
Beijing Normal University

Standard setting methods and cognitive diagnostic analysis both seek to classify examinees into two or more categories. This study proposes a method to facilitate standard setting with cognitive diagnostic analysis where cognitive diagnostic models with a Q matrix are used. Cutoff points are set in two stages. First, examinee responses are analyzed with cognitive diagnostic models and standards are related to certain attribute mastery patterns (AMPs). In the second stage, item parameters obtained in the first stage are used to simulate response data with known AMPs; then the cutoff points that optimize classification accuracy can be identified. The proposed method fully makes use of information from the test itself and the test takers. Unlike the widely used Bookmark methods or Angoff methods, experts are only required to identify the Q matrix instead of predicting the examinee performances, which can be a huge burden and add to the total error.

Jubileumzaal **J.3.-Differential Item Functioning -3**

Detecting differential item and differential step functioning by means of model-based recursive partitioning

Carolin Strobl, *Universität Zürich*
Basil Abou El-Komboz, *Ludwig-Maximilians-Universität München*
Julia Kopf, *Ludwig-Maximilians-Universität München*
Achim Zeileis, *Universität Innsbruck*

Based on a flexible model-based recursive partitioning framework, Strobl, Kopf and Zeileis recently suggested a procedure for detecting differential item functioning (DIF) in the Rasch model. The main advantage of this approach is that it allows to detect groups of subjects exhibiting DIF that are not pre-specified, but are detected automatically from (combinations of) covariates. The statistical methodology behind this approach is outlined and its properties are illustrated by means of simulation studies and real data examples. Moreover, two extensions for polytomous responses based on the rating scale model and the partial credit model are presented, that are applicable for detecting both differential item and differential step functioning (DSF).

Anchor methods for DIF detection: A comparison of the iterative forward, backward, constant and all-other anchor class

Julia Kopf, *Ludwig-Maximilians-Universität München*
Carolin Strobl, *Universität Zürich*
Achim Zeileis, *Universität Innsbruck*

In the analysis of differential item functioning (DIF) using item response theory (IRT), a common metric is necessary to compare item parameters between groups of test-takers. In the Rasch model, the same restriction is placed on the item parameters in each group in order to define a common metric. However, the question how the items in the restriction - termed anchor items - are selected appropriately is still a major challenge. A conceptual framework for categorizing anchor methods is proposed: The anchor class to describe characteristics of the anchor methods and the anchor selection strategy to guide how the anchor items are determined. Furthermore, a new anchor class termed the iterative forward anchor class is proposed. Several anchor classes are implemented with two different anchor selection strategies (the all-other and the single-anchor selection strategy) and are compared in an extensive simulation study. The results show that the newly proposed anchor class combined with the single-anchor selection strategy is superior in situations where no prior knowledge about the direction of DIF is available.

A Comparison of Multilevel Modeling, Structural Equation Modeling, and the Likelihood Ratio Test for Detecting Differential Item Functioning (DIF)

Mei Ling Ong, Zhenqiu (Laura) Lu, Allan Cohen &Sunbok Lee
University of Georgia

Multilevel models are used to address the natural hierarchical structure that is present in most educational data. More complex forms of nesting are not only possible but typical, such as when teachers teach more than one level of a course within a subject area. Although teachers are usually nested within schools, it is also possible that some teachers may teach the same subject but do it in more than one school in the district. In this context, using standard methods based on non-IRT or IRT based methods, for detection of differential item functioning (DIF), runs the risk of ignoring this hierarchical structure, thus likely biasing the results. A number of recent articles have employed the multiple-indicators, multiple-causes approach (Acar, 2012; Woods, 2009) or a multilevel modeling approach (Finch, 2005; French & Finch, 2010) to detect DIF. As yet, however, no results have been reported directly comparing MIMIC, HLM, and IRT-likelihood ratio test methods. Thus, the purpose of this study is to compare these three methods for detecting DIF in hierarchical data. We do so using the Rasch model. Type I error rates will be examined using simulated hierarchical data. A real-data example will be used to motivate the study.

An Empirical Comparison of IRT and CTT based Differential Item Functioning Indices

Muhammad Naveed Khalid, *Cambridge English Language Assessment*
Faisal bin Abdullah Al-Mishari Al Saud, *National Center for Assessment in Higher Education*
Farah Shafiq, *University of Glasgow*

The analysis of differential item functioning (DIF) examines whether item responses differ according to characteristics such as language and ethnicity, when people with matching ability levels respond differently to the items. DIF analysis can be performed by a range of statistics. In present study, we will explore IRT based index critical ratio test and CTT based indices Mantel-Haenszel (MH) & Standardized Proportion Difference (SPD). This study will empirically examine the behaviors of the DIF statistics derived from these two measurement frameworks using simulation studies and real data. Study will focus on number of aspects such as computational methodology, significance level, identification of DIF items, effect size and scale purification.

Balkonzaal **J.4.-IRT estimation 2**

A logistic function of a monotonic polynomial for estimating item response functions

Carl F. Falk & Li Cai
University of California

In this talk, we present a semi-parametric approach to estimating item response functions (IRF) for dichotomous items when the IRF does not follow a smooth 2PL function. Our current approach models individual items using a logistic function of a monotonic polynomial (L-MP) and is implemented using Bock-Aitkin EM MML estimation. The L-MP approach ensures monotonically increasing IRFs while allowing one or more additional “bends” to occur in a logistic IRF with higher degree polynomials allowing more flexibility. Addition of a lower asymptote to the L-MP model allows estimation of items with a non-zero probability of “guessing” simultaneously with a more flexible IRF. At the lowest order polynomial, these models reduce to the familiar 2PL and 3PL models. We present simulation results comparing the L-MP model to these more standard approaches and discuss cases where each approach has better recovery of the true IRF. Finally, we briefly discuss the possible advantages/disadvantages of the L-MP model versus other non-parametric approaches to estimating IRFs.

Comparison of maximum likelihood with conditional pairwise likelihood estimation of person parameters in the rasch model

Clemens Draxler, Gerhard Tutz & Katharina Zink
Ludwig-Maximilians-Universität München

This paper is concerned with person parameter estimation in the Rasch Model. In particular, an approach recently proposed by Andrich is investigated which is a special case of the so-called pseudo, quasi or composite likelihood methods. By means of a Monte Carlo study it is compared to two well-established maximum likelihood approaches, one of which being the well-known weighted likelihood procedure. From the theory of maximum likelihood it is well-known that composite likelihood estimators are consistent but inefficient. Therefore, the magnitude of the loss of efficiency of Andrich’s approach compared to maximum likelihood procedures is of interest. The results show that the observed values of the root mean squared error are practically equivalent for the compared estimators in the case of a sufficiently large number of items. Andrich’s pseudo likelihood estimator yields a considerably larger bias and variance for small item numbers only.

The bifactor latent regression model: efficient parameter estimation and an application to PISA 2006

Frank Rijmen, *ETS*

In this presentation, we will combine two methods for reducing the computational burden for the latent regression bifactor model and for related models. As is the case for the regular bifactor model (without a regression of the latent variables on background variables), full-information maximum likelihood estimation can be carried out efficiently by exploiting the conditional independence relations implied by the model. Specifically, it involves function evaluations in two-dimensional spaces only. The computational burden will be reduced further by relying on adaptive quadrature when integrating over the latent variables in these two-dimensional subspaces. The combination of dimension reduction and adaptive quadrature has not been described before in the literature. The approach will be illustrated with a simulation study and applied to NAEP Science.

Discrete MCMC for the 2 Parameter Logistic Model

Tammy Trierweiler, *Educational Records Bureau*
Robert L. Smith, *American Institutes for Research*
Yuanchao (Emily) Bo, *Fordham University*
Charles Lewis, *Fordham University*

Markov chain Monte Carlo methods often take considerable time to converge. Although drawing randomly sampled values from conditional posterior densities is straightforward when the distributions follow standard forms, often the form is not standard and requires another approach. Here, a method is proposed where approximation of the conditional posterior density function is achieved through the use of discrete probability distributions evaluated on sets of “quadrature points.” This method is applied to the estimation of Rasch model parameters. Suppose we have a joint posterior distribution for person and item parameters $p(\theta_1, \ldots, \theta_N; \beta_1, \ldots, \beta_n \mid \mathbf{Y})$.. We define across a set of Q quadrature points *c*. For a fixed set of $(\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots,$

$\theta_N; \beta_1, \ldots, \beta_n$), we define $v_{i(q)} = p(\theta_1, \ldots, \theta_{i-1}, \theta_{(q)}, \theta_{i+1}, \ldots, \theta_N; \beta_1, \ldots, \beta_n \mid \mathbf{Y})$ for q=1,…, Q and let $c_{i(q)} = \sum_{r=1}^{q} v_{i(r)} \big/ \sum_{r=1}^{Q} v_{i(r)}$. To sample θ_i from its conditional posterior distribution, we start with u, a random variable sampled from a uniform distribution over (0, 1). Let q_{min} be the smallest value with $u < c_{i(q)}$ and let the sampled value of θ_i equal $\theta_{(qmin)}$¬. This is repeated for each θ_i, and values are updated as they are sampled. The same process is applied to each item difficulty β_j. Results suggest that this method produces parameter estimates efficiently with almost the same level of precision as more time-intensive methods.

Stadszaal **J.5.-GLM**

Estimation of cross-level interactions in a three-level nonlinear latent variable model with a Metropolis-Hastings Robbins-Monro algorithm

Ji Seung Yang & Li Cai
UCLA

Nonlinear multilevel latent variable modeling has been suggested as an alternative approach to traditional multilevel modeling for the situations in which latent predictors are measured by categorical manifest variables. While nonlinear multilevel latent variable modeling allows simultaneous estimation of measurement and structural parameters (taking measurement error and sampling error into account), obtaining full information maximum likelihood estimates (FIML) of this model is computationally intensive. The computational burden gets even worse when a structural model is expanded from two to three levels since the number of latent variables increases. The purpose of this research is to efficiently estimate cross-level interactions in a three-level nonlinear latent variable model using a Metropolis-Hastings Robbins-Monro algorithm (MH-RM; Cai, 2008, 2010a, 2010b). The objectives include to build an efficient Metropolis-Hastings sampler and to identify optimal conditions for the algorithm (e.g. tuning constants). The MH-RM algorithm is implemented in R (R Core Team, 2012), and parameter recovery and computational efficiency are compared with an EM algorithm that uses adaptive Gauss-Hermite quadrature for numerical integration (e.g. Mplus; Muthén & Muthén, 1998-2011). Preliminary results for a two-level nonlinear latent variable model indicate that the MH-RM algorithm can obtain FIML estimates much faster than the EM algorithm with numerical integration.

Monte Carlo Local Likelihood for Estimating Generalized Linear Mixed Models

Minjeong Jeon, Cari Kaufman & Sophia Rabe-Hesketh
University of California

We propose the Monte Carlo local likelihood (MCLL) method for estimating generalized linear mixed models (GLMMs). MCLL initially treats model parameters as random variables, sampling them from the posterior distribution in a Bayesian model. The likelihood function is then approximated up to a constant by fitting a density to the posterior samples and dividing it by the prior. In the MCLL algorithm, the posterior density is approximated using local likelihood density estimation (Loader, 1996) in which the log-density is locally approximated by a polynomial function. In his Monte Carlo kernel likelihood (MCKL) method, De Valpine, (2004) proposed a similar approach using kernel density estimation instead of local likelihood density estimation. We also develop a novel method to compute standard errors and the Bayes factor. Using empirical and simulation examples, we show that our proposed method outperforms existing methods such as the Laplace approximation and Monte Carlo maximum likelihood estimation (MCMLE).

Estimation methods for categorical marginal models: comparing MAEL, GEE, and GSK.

Renske E. Kuijpers, *Tilburg University*
Wicher P. Bergsma, *London School of Economics*
L. Andries van der Ark, *Tilburg University*
Marcel A. Croon, *Tilburg University*

Categorical marginal models can be used for modeling clustered or dependent data. For example, marginal models are used to construct hypotheses tests and standard errors for certain coefficients, such as Cronbach's alpha and scalability coefficients (e.g., Kuijpers, Van der Ark & Croon, 2013). The most used estimation method for marginal models is maximum likelihood (ML) estimation. However, for larger sets of items, marginal modeling needs much computation time, and problems with memory capacity occur. These problems can be avoided by using maximum augmented empirical likelihood (MAEL) as an estimation method (Van der Ark, Bergsma & Croon, 2013). MAEL estimation uses all nonzero cells in a contingency table, plus a number of well-chosen zero cells. MAEL is a rather new method, and further investigation is needed. More common estimation methods for marginal models are generalized estimating equations (GEE), and GSK. GEEs (Liang & Zeger, 1986) represent an extension of the generalized linear model (GLM), and can be used for the analysis of clustered or correlated data. The GSK method (Grizzle, Starmer & Koch, 1969) is based on Weighted Least Squares (WLS) procedures. In this paper, the new estimation method MAEL is compared to GEE and GSK, using simulation studies as well as a real-data example.

Receiver operating characteristic analysis with nonlinear mixed effects models – A model architecture

Rüdiger Mutz, *ETH Zurich*

Receiver operating characteristic (ROC) analysis has been developed to an independent statistical tool in psychological and in particular medical diagnostic beyond logistic regression. In spite of the huge literature to different aspects of ROC, a methodological architecture is missing which incorporates the different aspects in order to push research, as structural equation models do for all kind of latent variable models. Starting with the binormal model as a basic model for ROC, the nonlinear mixed effects model as it is implemented in SAS "proc nlmixed" (Gönen, 2007) will be presented as such a model architecture to cover the different aspects of ROC analysis as (1) estimation of ROC curve parameters, the area under the curve (AUC), and cut points (2) considering homogeneity of measurements within two or three levels (mixed effects), (3) mixture models, (4) models with measurement errors, (5) continuous or ordinal scaled diagnostic tests, (6) causality, and (7) parametric versus nonparametric conditions. Ex-ante peer evaluations of proposals and funding decisions of the Austrian Science Fund, Austria`s leading funding organization for basic research, provide for the data to illustrate some aspects of the framework.

Hemelzaal **J.6.-Factor analysis 2**

Examining the results of different models used in multitrait-multimethod analysis

Elif Bengi Ünsal Özberk & Nuri Dogan
Hacettepe University

In this reseach, method related and trait related specifications are compared in multitrait-multimethod confirmatory factor analysis . Based on persons' problem solving and crticital thinking abilities, a model related to creative thinking abilities was introduced. The model datas collected with self-report, questionnaire and friend report methods in 500 students. Reliability and validity was interpreted in a multitrait-multimethod matrix by measuring the different traits using different methods. Later,

correlated method, uncorrelated method and “minus-1” method in multitrait-multimethod confirmatory factor analysis are compared to decide which method works better to evaluate the models.
The reliability diagonal values confirmed to be highest values of the matrices whlie collecting validity evidence in creative thinking ability. The validity coefficients are also different from zero. In validity diagonal, heterotrait-heteromethod triangle and heterotrait-monomethod triangle values is lower than validity diagonal values.
Total scores, which was calculated by three diffrent method on problem solving and creative thinking abilities set as a structural model on creative thinking ability. Fit indices of creative thinking model were found acceptable. In multitrait-multimethod matrices, best model to assess creative thinking ability is the “minus-1” model.

A comparison of confirmatory factor analysis of binary data on the basis of tetrachoric correlations and of probability-based covariances: a simulation study

Karl Schweizer, *Goethe University Frankfurt*

The investigation of the structure of binary data by means of confirmatory factor analysis can be conducted on the basis of tetrachoric correlations that include the estimation of latent thresholds. An alternative approach to investigating the structure of binary data that is independent of latent thresholds is confirmatory factor analysis on the basis of probability-based covariances. This approach additionally includes the consideration of a link function according to the generalized linear model for bridging distributional differences between the binomial and normal distributions. Simulated data were generated according to a uniform 9x9 population pattern with n=200, 400 and 1000. The nine originally continuous variables were dichotomized by nine different splits one of which was the median split. The fit results obtained in investigating 100 datasets indicated that the n should be rather larger in the case that tetrachoric correlations provide the basis. Another important finding of the study was that only versions of the tau-equivalent model in combination either with tetrachoric correlations or probability-based covariances enabled a quite precise recovery of the factor loadings according to the uniform population pattern. In the case of probability-based covariances disattenuation is additionally necessary.

Evaluating Bayesian, Robust Maximum Likelihood, and Robust Least Squares Approaches in Ordinal Confirmatory Factor Analysis for Small Samples

Prathiba Natesan, *University of North Texas*

Although several software and estimation solutions exist for fitting ordinal CFA models and SEMs, possible non-convergence and lack of reliable statistics are still some challenges in overall model estimation. This problem is exacerbated in small samples and nonnormal data. Bayesian estimation depends less on asymptotic theory and therefore can be particularly useful with nonnormality in small samples. What needs to be known is a minimum requirement of sample size when planning a study with ordinal CFA using Bayesian estimation.
The present simulation study compares MCMC estimation of ordinal CFA models for small sample cases with RML, WLS, RULS, and RDWLS estimations. A simple two factor CFA model with five items per factor, four sample sizes (n=2df, 3df, 4df, and 5df), and three Pearson correlation coefficient values (low = 0.2, medium = 0.5, and high = 0.8) were studied. MCMC produced lower RMSEs especially at small samples and better credibility intervals. In MCMC, 94.8% of the 95% HDIs for 1200 datasets (all conditions) contained the true parameter while only 65% of the confidence intervals captured the true values even in the best case of least squares estimates. Non-convergence for small samples continues to be an issue in least squares and ML.

The Ratio Information Maximum likelihood Estimator
Steffen Grønneberg & Ulf Henning Olsson
BI Norwegian School of Management

When estimating Structural Equation models with full information procedures like the MLE, all model parameters can be influenced by model misspecification. The measurement mechanism of the latent phenomena one is trying to measure may then become severely skewed, and one faces the risk of analyzing latent variables void of meaning. We propose a new estimation method that interpolates between full information methods and multi-stage estimators. It can provide a user-specified degree of robustification against this problem, and we provide a set of diagnostic tools based on our estimation scheme. General consistency results are reached and we and we provide empirical examples for the technique's usefulness.

12:40 - 13:55 Lunch break + Poster session II

27. **Use of Crude Prior Information for Item Parameter Estimation in the Item Response Theory**
Kentaro Kato, *Benesse Corporation*

28. **On constrained estimation of factor analysis by the EM algorithm and Bayesian approaches**
Kentaro Hayashi & Lu Liang
University of Hawaii at Manoa

29. **The Latent Class Analysis of Bullies, Victims and Bully-Victims in Adolescence: Relations with Social and School Adjustments**
Aihui Shao, *Beijing Normal University*

30. **Towards a ICF-based measure of functioning in spinal cord injury**
Carolina Ballert, *Swiss Paraplegic Research*

31. **Using the DINA Model with Between-Group Heterogeneity**
Sohee Kim, *Sungshin Women's University*
Hyejin Shim, *Sungshin Women's University*
Chanho Park, *KICE*

32. **Performance of Model Selection Criteria in Structural Equation Modeling**
Li-Chung Lin & Li-Jen Wenig
National Taiwan University

33. **Incorporating response time to diminish the impact of careless responses in questionnaire survey**
Rung-Ching Tsai & Ke-Chian Lin
National Taiwan Normal University

34. **Power Analysis for Moderation Effect using Multiple-Group Structural Equation Modeling**
Yi-Chun Lin, *National Cheng Kung University*

35. **Effectiveness and robustness of Bollen-Stine bootstrapping in testing nested models**
Ching Lin, *National Cheng Kung University*

36. **An effect of local dependence for conventional item statistics**
Naoya Todo, *The University of Tokyo*

37. **Analyzing Number Sense Performances of Students on the Broken Calculator Assessment with Cognitive Diagnostic Model**
Shu-Chuan Shih, Shu-Juan Lee & Bor-Chen Kuo
National Taichung University of Education, Taiwan

38. **Latent Transition Analysis with distal outcomes: Stability in Victimization and its Association with Emotional Problem in Adolescence**
Lichan Liang & Aihui Shao
Beijing Normal University

39. **Model Specification of the Nonlinear Effects between First-order Latent Variable**
Shu-Ping Chen & Chung-Ping Cheng
National Cheng Kung University

40. **Detecting Differential Attribute Functioning (DAF) for gender using the multi-group DINA (MGDINA) model for TIMSS Grade 8 Mathematics Assessments**

Ruchi Sachdeva, Mathew Johnson, Young-Sun Lee, Jianzhou Zhang & JungYeon Park
Columbia University

41. **A Cognitively Diagnostic Modeling of Attribute Mastery in Different Content of Grade Fifth Math Using the Basic Competence Test**
Chien-Ming Cheng, *NAER*

42. **A paradoxical fall of r and reliability by increasing the number of rating categories**
Kenpei Shiina, *Waseda University*
Saori Kubo, *Waseda University*
Yoshihiro Ouchi, *Josai International University*
Takashi Ueda, *Waseda University*

43. **Development of the basic Literacy test for university students**
Po-Hsi Chen, Chun-Yu Hsu, Po-Wei Li, Yu-Shin Chen, Chuan-Yi Yang, Tai-Ting Yeh & Hsin-Ying Huang
National Taiwan Normal University

44. **Comparison of Partial Least Squares Path Modeling to Covariance-Based SEM**
Justin Neil Young, *University of Houston*

45. **The effect of different item parameter estimation methods on vertical tests equating**
Yuan-Chieh Yang, *National Taiwan Normal University*

46. **Nested Measures and Transitive Variance**
Ralph Carlson, Hilda Medran & Gabriela Rosa
The University of Texas-Pan American

47. **The power of fit indices of Rasch model to detect items with non-zero asymptote parameters**
Jen-Hua Hsueh & Po-Hsi Chen,
National Taiwan Normal University

48. **Generalized reliability used for comparison of reliability estimates in tests with binary items**
Patricia Martinkova, *Academy of Sciences of the Czech Republic*

49. **The Accuracy of Mislevy's Methodology to Estimate the Latent Proficiency Distribution**
Hyung Jin Kim, *The University of Iowa*

14:00 - 15:20 Parallel session K

Concertzaal **K.1.-Cognitive diagnostic assessment 3**

Longitudinal Analysis of TIMSS Grade 8 Mathematics Assessments Using CDM

Jianzhou Zhang, Young-Sun Lee, Matthew Johnson, Jung Yeon Park & Ruchi Sachdeva
Columbia University

Cognitive diagnosis models (CDMs) have become increasingly popular along with the increasing availability of large-scale assessment data for international comparison of educational achievement. In addition to its ability to diagnose the cognitive attributes of students, it could also be used to compare the knowledge skills of students across countries. “Multi-group DINA” (MG-DINA) model was developed to fit 2007 TIMSS grade 8 mathematics assessments. We analyze responses to 88 released items by students in 11 participating countries, the diagnostic information could be used in various instructional practices by identifying the presence or absence of the specific attributes. In addition to the cross-sectional analyses on TIMSS 2007 for international comparisons, we also developed a Q-matrix for the items that were repeatedly used through 2nd and 4th cycles (1999, 2003 and 2007 TIMSS) for a longitudinal study in order to detect any changes in attributes over time. Among the 88 released items in TIMSS 2007 8th mathematics assessment, 21 items were tested in both 1999 and 2003. These items represent all four content domains and three cognitive domains specified in the TIMSS 2007 framework and they were used to detect the differences of attributes among the three administrations.

Extending the DINA Model to Incorporate Continuous and Discrete Covariates using Linking Functions: Model Development, Parameter Estimation, and Applications

Marcus Waldman, Matthew Johnson & Young-Sun Lee
Columbia University

Cognitive diagnostic models provide highly detailed information about student learning on a predefined set of skills, making them ideal for cross-country comparisons using international assessments, such as TIMSS. One specific type of cognitive diagnostic model, the multi-group deterministic, inputs, noisy, “and” gate (MGDINA) model was recently developed to distinguish the strengths and weaknesses of individual countries or groups. Yet, the MGDINA model has been unable to incorporate well-known continuous covariates that predict student learning, such as socioeconomic status. Moreover, the MGDINA model suffers from an exponential growth in the number of parameters needed to model the latent classes defined by the set of skills specified in the Q-matrix. In response, the DINA model is modified by embedding linking functions to predict the probability of skillset prevalence. This modification both reduces the parameter space and seamlessly allows for the incorporation of both discrete and continuous covariates. Parameter point estimates, model fit, and methods for analyzing the effect of socioeconomic status on student learning using the TIMSS 2007 eighth grade mathematics assessment are then compared and contrasted between this developed DINA model and the MGDINA model.

How to judge a cognitive test as an efficient test

Shuliang Ding, Fen Luo & Wenyi Wang

Jiangxi Normal University

A cognitive diagnostic test is defined as an efficient test if it could completely represent the cognitive model within the domain that is concerned. The existing cognitive diagnostic tests are usually an incomplete representation of hypothetical or theoretical cognitive models, resulting in less accurate assessment of students' knowledge states than desired. The question is how to assess the representation of a cognitive diagnostic test. With the Attribute Hierarchy Model, a hypothetical cognitive model of task performance reflects a attribute hierarchy within a domain, and a formal representation of the construct measured by the test is the test Q matrix . If the hierarchy drawn from the test Q matrix coincides with the hierarchy that is drawn from the theoretical analysis, the efficiency of the Q matrix is perfect, otherwise, defective. The efficiency equals to the ratio of the number of unequal item types (NOUIT) derived from the test Q matrix to the NOUIT derived from the theoretical hierarchy using the augment algorithm. And how to improve the efficiency of a diagnostic test is discussed under the non-compensatory and conjunctive condition. To draw hierarchy from test Q matrix an algorithm that is different from Tatsuoka's method is proposed.

DINA model and knowledge state estimation: a multiple imputation–based approach

Xin tao, *Beijing Normal University*

Traditional estimators of cognitive diagnostic model knowledge state ignore uncertainty carried over from the item calibration process. About this topic, researchers have compared a variety of approaches in item response theory (IRT) modeling. Among approaches suggested over the past few decades, a multiple imputation–based approach is particularly flexible and useful because it can be easily applied to tests with mixed item types and multiple underlying dimensions. In this paper, the authors will use the multiple imputation–based approach in DINA model to correct the estimates of individual knowledge states.

In the simulation study, Monte-Carlo experiment was used to test the classification accuracy rate of a multiple imputation–based approach under 4(attribute hierarchy)×3(sample sizes)×2(magnitudes of item parameters) conditions. We used MATLAB to run L=1000 replications for each experimental condition. Pattern correct classification rate and attribute correct classification rate were computed to compare the estimated MAP and EAP with traditional estimators.

The results showed that the impact of item parameter uncertainty is generally quite small, though the uncertainty carried over from item calibration contributes substantially to the estimators of knowledge states when the item parameters are relatively large.

Parkzaal **K.2.-Missing data in IRT**

Treatment of missing data in covariates in large-scale educational assessments.

Cees Glas, *University of Twente*

In large-scale educational surveys (PISA, TIMSS, PIRLS, NEAP, PIAAC) the cognitive level of the participants is assessed using a block-rotated test administration design, and the results are subsequently analysed using an IRT model. Further, a host of background information is collected simultaneously. This information can be either manifest (age, gender, region, etc.), or derived from scales measuring latent variables (SES, reading attitude, etc.). Students are nested in schools and also school characteristics are collected. The ensemble of the data is then analysed using a multilevel IRT model (see, for instance, Fox & Glas, 2001, 2002).

A new development in the design of large-scale educational surveys is the wish to also use a block-rotated design to collect the background information. The problem is how to incorporate missing data in covariates into a multilevel IRT model. The problem is tackled in a fully Bayesian framework using simultaneous regression equations for estimation of the multilevel IRT model and the imputation of the missing information. The model is both evaluated using a data-augmented Gibbs-sampler and using the WinBugs software. The precision of various possible block-rotation designs is evaluation is carried out using simulated data and data from the PISA databases.

Modeling missing-data processes: A tree-based IRT approach.

Dries Debeer, *University of Leuven*
Rianne Janssen, *University of Leuven*
Paul De Boeck, *Ohio State University*

In large-scale educational assessments missing data are not uncommon. In fact, two missing data processes can be discerned. Firstly, non-response can be caused by the inability of test takers to complete the entire test, which leads to missingness at the end of the assessment (“not-reached items”). Secondly, test takers may omit certain responses well before reaching their last answered item (“skipped items”). Both missing-data processes may be related to the proficiency of the test taker and, hence, cause non-ignorable missingness (Rubin, 1976; Little & Rubin, 1987). One strategy to reduce possible bias is to model these processes. Current psychometric solutions focus on only one of these two missing-data processes (Holman & Glas, 2005; Glas & Pimentel, 2008). Both methods combine an IRT model for the observed data (i.e. accuracy process) with an IRT model for the missing data, as proposed by O’Muircheartaigh and Moustaki (1999). In this paper a new approach will be presented to deal with both non-ignorable missing data processes in binary item-response data, based on the IRTree framework of De Boeck and Partchev (2012). After discussing the modeling approach, the method is illustrated with data from the PISA 2009 reading assessment.

Can IRT solve the missing data problem?

Maria Bolsinova, *Utrecht University / Cito*
Gunter Maris, *University of Amsterdam / Cito*

Test equating is considered as a missing data problem: the unobserved responses of the reference population to the new test must be imputed to specify a pass/fail criterion in such a way that the proportion of students from a reference population failing at the new exam is equal to the proportion of students failing the reference exam. This is possible only if we assume a parametric ability distribution in the population, which is a not fully testable assumption and can also lead to bias in equating when the distribution is mis-specified. In this study we investigate whether IRT allows to make inferences about the distribution of the missing responses from the data only. The Extended marginal Rasch model is used. Although the parameters of the model are not fully identified, the uncertainty about the score distribution on the new test, caused by the non-identifiability is very small and can be ignored in practical applications. This is illustrated using simulated examples. To show the practical advantage of the distribution-free approach, which does not have bias caused by mis-specification of the ability distribution, the Extended Rash model for test equating is applied to the data of the Dutch Central Examinations.

How to obtain estimates of latent trait scores for ordinal data when some participants' responses are missing

Joost R. van Ginkel, *Leiden University*
Urbano Lorenzo-Seva, *Universitat Rovira i Virgili*

Researchers frequently have to analyze scales in which some participants have failed to respond to some items. In this presentation we focus on the exploratory factor analysis of multidimensional scales (i.e., scales that consist of a number of subscales) where each subscale is made up of a number of Likert-type items, and the aim of the analysis is to estimate participants' scores on the corresponding latent traits. Our approach is based on the following steps: (1) multiple imputation is used to create a few copies of the data, in which the missing values are arbitrarily imputed; (2) each copy of the data is subject to independent factor analysis, and the same number of factors is extracted from all copies; (3) all factor solutions are simultaneously orthogonally (or obliquely) rotated so that they are both (a) factorially simple, and (b) as congruent with one another as possible; (4) latent trait scores are estimated for ordinal data in each copy; and (5) participants' scores on the latent traits are estimated as the average of the estimates of the latent traits obtained in the copies.

Jubileumzaal **K.3.-Differential Item Functioning -4**

Modeling the influence of interactions among subgroups on differential item functioning detection using the likelihood ratio test and the item response theory approach

Hui-Fang Chen, Kuan-Yu Jin & Wen-Chung Wang
Hong Kong Institute of Education

Differential item functioning (DIF) analysis is usually conducted with an interested variable one by one. It means that a common procedure only concerns a main effect of a demographic variable on items and ignores complex interactions among them. The study examined the impact of ignoring an interaction among group memberships using a simple scenario: Two 2-level demographic variables (e.g., gender and country) forming four subgroups were involved in DIF detection. Different patterns of DIF among subgroups were simulated and examined by three DIF detection methods, including analyzing either one variable only, both variables at the same time, and an interaction. Factors such as IRT model, test length, sample size, and DIF size were manipulated. Simulations were carried out by using WinBUGS. Findings suggested that the DIF item can be precisely detected if the critical variable(s) or interaction was considered. In addition, disregarding the interaction resulted in a lower true positive rate, but did not increase the false positive rate.

Using Multilevel Logistic Mixture Model to Detect Multilevel Manifest DIF and Latent DIF

Jungkyu Park & Hsiu-Ting Yu
McGill University

Demographic variables such as gender, age and race are often used as grouping variables in traditional Differential Item Functioning (DIF) analyses. However, several DIF studies pointed out that such manifest groups cannot effectively detect the presence of DIF in data (e.g., Kang & Cohen, 2003; Samuelsen, 2005). Sometimes, the unobserved nuisance latent attributes can cause DIF more than the characteristics of manifest groups. The latent class approach has been suggested as an alternative method to overcome such heterogeneity of subjects within same manifest groups. Taking the latent class approach to detect DIF has been mostly applied to individual level. However, DIF can exist at individual and group levels simultaneously under a nested data structure. In this study, an approach of utilizing Multilevel Latent Class model (Vermunt, 2003) to detect DIF at both individual and group-levels simultaneously is proposed. An empirical example is used to illustrate the proposed method. The advantages of the proposed method are highlighted by comparing the results of four-step procedure suggested by Samuelsen (2005) and the Mantel-Haenszel technique. This study concludes with recommendations and remarks on detecting DIF simultaneously at different levels under the multilevel data structure.

A different view on DIF

Timo M. Bechger & Gunter Maris
Cito Institute for Educational Measurement

IRT can be said to provide maps with items and persons located in ability space; e.g., a line in uni-dimensional models and a plane in 2-dimensional models. From this “cartographer’s view” on IRT the investigation of measurement invariance involves a comparison of maps obtained from different samples of respondents. When the maps are not the same, we wish to determine in what way the maps are different: This is the purpose of differential item functioning (DIF). The study of DIF (i.e., the comparison between maps) is complicated by the fact that there are properties of the maps that we cannot determine such as the location of the origin and/or the size of the unit. This restricts the statements that we can do about the differences between the maps. For example, even in the simplest IRT models we can only determine that relative locations of items are different in two maps. With more complex IRT models, the comparison becomes more complex and the main purpose of the talk is to sketch how.

Bayesian Nonlinear Mixed Model for DIF Analysis

Zairul Nor Deana Md Desa, Carol M. Woods, Kevin J. Grimm &Andrés Sandoval-Hernández
IEA Data Processing and Research Center

In this study, we describe a Bayesian approach for the study of uniform and non-uniform differential item functioning (DIF) using a nonlinear mixed model, a model equivalent to a MIMIC model with a latent interaction. A Monte Carlo study was carried out to test the performance of the proposed model in detecting DIF. For the simulation study, true item and examinee parameters were used to generate response data for different sample sizes and test lengths. For computing intensity, a parallel Markov Chain Monte Carlo (MCMC) with the Gibbs sampling simulation method was applied to estimate the parameters in the model using R2OpenBUGS. Type I error rate and power were evaluated and discussed in the performance of the proposed DIF detection method. This study also examined test items from TIMSS 2011, the fifth in the IEA’s series of international assessments of student achievement. Multiple-choice questions with correct and incorrect responses from all five content domains on the eighth grade mathematics assessment (i.e., Number, Algebra, Geometry and Data Chance) were examined for both uniform and non-uniform DIF.

Balkonzaal **K.4.-Factor analysis 3**

An empirical Kaiser criterion for factor retention decisions.

Marcel van Assen & Johan Braeken
Tilburg University

In any application of factor analysis, a crucial step is deciding how many factors to retain. Parallel analysis (PA) is currently considered to be the most accurate factor retention method in exploratory factor analysis. The method is based on a comparison of the eigenvalues of the correlation matrix of the observed data to eigenvalues of randomly generated correlation matrices under the assumption of independence. Despite its generally good performance, PA is also known to underestimate the number of factors when factors in the population are correlated. We propose a new factor retention method that (i) performs equally good compared to PA when factors are uncorrelated, (ii) performs better than PA when factors are correlated, and (iii) has a straightforward implementation that does not require simulation.
The performance of both PA and the proposed new method is examined in practical situations where researchers aim to construct scales with sufficient reliability and sufficiently strong items. Our examination reveals that the new method works well in almost all practically relevant situations, whereas PA indeed fails in practically relevant situations with short correlated scales.

*Factor Analysis with Weighted L1 Regularization:_x000D_*

A Procedure for More Efficient Estimators

Po-Hsien Huang & Li-Jen Weng
National Taiwan University

Factor analysis is one of the most widely used statistical methods in psychological research. In the present study, factor analysis with weighted L1 regularization (FA-WL1) was proposed for estimation with small samples. Regularization is one way to improve estimation quality through controlling model complexity. FA-WL1 utilizes weighted L1 regularized likelihood as the estimation criterion. A core feature of the L1 regularization lies in its ability to produce sparse estimates. Hence, the regularized estimation reduces the effective number of factor loadings to be estimated and yields a more efficient estimator. An ECM algorithm (Meng & Rubin, 1993) was implemented to optimize the regularized estimation criterion. A simulation study was conducted to evaluate the empirical performance of FA-WL1 under small and moderate sample sizes.

Which method detects the right number of dimensions? comparison of EFA and CA

Stella Bollmann, Moritz Heene, Helmut Küchenhoff & Markus Bühner
Ludwig-Maximillians University Munich

This study investigated the performance of exploratory factor analysis (EFA) and cluster analysis (CA) in dimensionality assessment. Additionally to simulated data a real, large data set was analysed to overcome limitations of classical simulation studies. To achieve a criterion for number of real dimensions the population model of this data set was used. Then we drew samples of different sizes out of this data set and examined how often they yielded the same number of dimensions. Finally, these results were compared to the results from simulated conditions in which cross loadings, factor correlations and sample sizes were manipulated. Results indicate that parallel analyses with principal axis factoring (PA-PAF) as well as the MAP-test reach satisfactory results for samples sizes starting at around 150. Their performance was reached only by one CA method, the average linkage - correlation of correlation (AL-RCC). The BIC was highly dependent on sample size. All EFA methods performed worst in the high cross loadings condition, while PA for principal component analysis (PA-PCA) was most and PA-PAF least affected. The MAP test only was affected by high cross loadings, not by varying cross loadings.

Dimensionality assessment using parametric bootstrap hypothesis testing

Vincent Kieftenbeld, *Southern Illinois University Edwardsville*
Ratna Nandakumar, *University of Delaware*

Unidimensionality is a fundamental assumption underlying many common item response models. Because parameter estimates may otherwise be inaccurate, confirming the unidimensionality of the response data is an essential step in item calibration. One approach to dimensionality assessment is the DIMTEST procedure, based on Stout’s nonparametric test for essential unidimensionality. Stout's test statistic has an asymptotical standard normal distribution. However, this may not be an accurate approximation to the sampling distribution with smaller samples or shorter tests. Type I error rates are known to be inflated in these situations. In this simulation study, we replaced the dependency on asymptotic normal theory with a Monte Carlo hypothesis test based on the parametric bootstrap. This procedure controls for type I error rates in most situations. Moreover, the procedure is very flexible, and can be used with more complex tests (for example, those containing a mix of dichotomous and polytomous items) or other test statistics.

Stadszaal **K.5.-(CCC) Classification, Clustering and Correspondence Analysis**

Using Ordinal Latent Class Analysis to Study Learning Performances

David Torres Irribarra, *University of California*

This paper presents the ordered linear logistic test model (O-LLTM), which combines the strengths of the linear logistic test model (LLTM; Fischer, 1973) and the ordered latent class analysis (OLCA; Croon, 1990). The O-LLTM can also be considered as an ordered extension of the mixture LLTM model presented by Mislevy & Verhelst (1990). The combination of these two models will allow researchers and practitioners to model student proficiency according to explanatory (Wilson & De Boeck, 2004) models expressed through the LLTM part of the model, while providing simple and interpretable results in terms of ordered performance groups.
The O-LLTM offers a simple, high level, interpretation of the respondent classes according to overall proficiency, as well as an explanatory interpretation in terms of the specific item features. The former interpretation lends itself for use in contexts where summative assessments are needed and the latter is more appropriate when diagnostic information is required.
I will illustrate the advantages of the O-LLTM using two empirical examples, comparing its performance to a traditional Rasch model (Rasch, 1960), LLTM and OLCA. I will examine the different substantive interpretations that can be obtained with the different models, their usefulness for summative and diagnostic purposes.

Applying multilevel latent class analysis to large-scale educational assessment data: predicting students' mathematical strategy choices from teachers' instructional practice

Marije F. Fagginger Auer, *Leiden University*

The study demonstrates the usefulness of multilevel latent class analysis (LCA) for educational data, by applying this technique to data from the 2011 large-scale assessment of Dutch primary schools' mathematics. The relation between the instructional practice reported by 107 teachers and the mathematical strategy choices of 1619 students was investigated. Multilevel LCA allowed modeling of the often ignored classroom effects, and one of its so far sparsely exploited features - the possibility of including predictors at different hierarchical levels - enabled modeling of the joint influence of teacher and student characteristics on learning outcomes. Four latent strategy choice classes of students were found, and teachers had a strong effect on students' probability of being in these classes. Effects were found of student characteristics and of teachers' strategy instruction, instruction formats and instruction differentiation. It is concluded that multilevel (teacher) effects should not be ignored in strategy research, and that multilevel latent class analysis is especially suited for application in educational research.

The bias-adjusted three-step approach to latent class modeling with external variables

Zsuzsa Bakk, Daniel Oberski & Jeroen K Vermunt
Tilburg University

A popular way to connect latent class membership to external variables is to relate the external variables to the estimated scores on class membership; this approach is called three step latent class analyses (LCA). While the three step LCA is a popular approach, until recently it had the disadvantage that the parameters describing the association of latent class membership and auxiliary variables were underestimated (Bolck, Croon, Hagenaars, 2004). In the current paper we present how unbiased parameter estimates of this association can be obtained, by using the known classification error probabilities as fixed value parameters in the third step analysis (Vermunt, 2010, Bakk, Tekle and Vermunt in press). Next to correct parameter estimates we also show how correct standard error estimates can be obtained. The correction for parameter bias we propose is already implemented in Mplus 7 (Asparauhov, Muthen 2012), while the correction for standard errors is yet to be implemented in commercial software. We show the results of a simulation study where we test the performance of the parameter bias correction, and the SE bias correction methods.

Intraindividual Attentional Coherence Across Scales of Time and Methods of Measurement

Stephen Aichele, *University of California*

The dynamics of psychological processes are fundamental to the study of behavioral, emotional, and cognitive development. The incorporation of multiple methods of assessment and observational time scales in research can greatly enhance knowledge of these dynamics. However, as researchers examine increasingly complex facets of psychological change, there is greater need for awareness of potential bias introduced by data aggregation during analysis. For example, "optimizing" levels of data aggregation in predictor-criterion variables can dramatically increase the strength of an estimated linkage, but pooling information merely to maximize predictive accuracy can compromise validity. These considerations inform analyses of data from the Shamatha Project, a longitudinal study of the psycho-physiological effects of intensive meditation practice. Sixty participants recorded daily changes in attentional stability and vividness across a three month training retreat. They also completed a laboratory measure of sustained attention at pre-, mid-, and post-retreat. Intraindividual coherence in the self-report measures of attention across the 3-month training period is evaluated as predictive of within-task coherence in attentional performance and reaction time variability at mid- and post-retreat. The analytical framework incorporates moving window correlations and characteristic-based time series clustering methods to minimize information loss due to data aggregation across persons and time.

Hemelzaal **K.6.-Problems with the use of NHST**

Towards reducing statistical reporting errors in psychology: co-piloting in scientific practice

Coosje Veldkamp & Jelte M. Wicherts
Tilburg University

Despite peer-review, an alarmingly high number of statistical reporting errors manages to enter even the most prestigious of journals (Bakker & Wicherts, 2011; Wicherts, Bakker, & Molenaar, 2011). In the present study, we aimed to examine whether a system of collaboration on a statistical analysis helps reduce such statistical errors. In this so-called copilot model, co-authors or colleagues double-check each others' statistical results .We investigated whether co-piloting practices among co-authors are associated with reduced error rates in the reporting of statistical results. We surveyed all authors and co-authors of all empirical articles published in six high-ranked psychology journals in 2012 about the extent to which they double-checked the analyses and results of the first (or only) study reported in their article. We then related their responses to the number of errors that we detected in the corresponding section of their article. We discuss the most common errors.

Outlier Removal and the Inflation of the Type I Error Rate

Marjan Bakker, *University of Amsterdam*
Jelte M. Wicherts, *Tilburg University*

No clear guidelines exist about what a researcher should do with outliers. Our review of the psychological literature shows that outliers are typically identified by using Z score thresholds and removed before running an analysis. Furthermore, John, Loewenstein and Prelec (2012) found that 38% of psychological researchers admitted to having decided to exclude data after looking at the impact of doing so on the results. However, outlier removal is not recommended as this will results in smaller estimates of the standard error and therefore a higher Type I error rate. We simulated sum scores based on dichotomous and polytomous items, for tests that fit the ability of the test-taker and for difficult tests, and show the impact of (subjective) removal of outliers on the Type I error rate of the t test. Result show that removing outliers with a threshold-value of Z of 2 in a short and difficult test, increases Type I error rates up to .222. We recommend non-parametric and robust methods as an alternative to outlier removal.

The Replication Paradox: Replication studies decrease bias of estimated effect sizes only if they are powerful

Michèle B. Nuijten & Marcel A. L. M. van Assen
Tilburg University

Studies with significant results are overrepresented in the literature because of publication bias. Due to this bias, most published research findings may be false (Ioannidis, 2005), and population effects are overestimated, particularly in small studies. An often proposed solution for this problem is to replicate studies more. Replication studies should uncover false positives and strengthen belief in true positives. In this study we examine the effect of replication on bias of the estimated population effect size as a function of publication bias and the studies' power. We show analytically that incorporating the results of a published replication study to estimate effect size will increase bias if the power of the replication study is smaller than that of the original study. We therefore conclude that mere replication will not solve the problems of many false published research findings and of overestimation of effect sizes. The most obvious solution is discarding and not publishing small studies with low power (Kraemer et al., 1998). Another less feasible solution is implementing practices that completely eliminate publication bias.

Researchers' perceptions of the weight of evidence from NHST outcomes and CIs

Rink Hoekstra & Richard Morey
University of Groningen

In science it is critical that claims be backed by some form of evidence. Within the social sciences, evidence is often statistical in nature. Statistical outcomes like p-values are frequently interpreted as measures indicating the strength of statistical evidence, and are often described using terms like “strong evidence”, “weak evidence”, and even “no evidence”. This implies that the strength of evidence is quantifiable for scientists.
It is not clear, however, whether scientists are actually convinced that quantifying evidence can be done in a valid way, and if so, whether they do this similarly. We present the outcomes of a questionnaire on scientists' evaluations of scenarios with fictitious study outcomes presented by means of commonly used inferential statistics. Notably, we found that many scientists do not believe that the strength of evidence is quantifiable given these statistics.
Furthermore, scientists indicating that the strength of evidence was quantifiable were enormously variable in their estimates quantification of the strength of evidence. Although significance tests and confidence intervals are often argued to be objective, our results indicate that in practice they are rather subjective because researchers do not appear to be constrained by any principle in their assessments of evidence.

15:25 - 16:05 Invited speakers

Concertzaal

Issues in IRT Modeling with Personality Test Data

Paul De Boeck, *Ohio State University*

Three issues in the domain of personality and personality testing will be focused on. Each of these implies a challenge for IRT modeling, with consequences for personality measurement and for the substantive study of personality phenomena.
1. Is ordinal ordinal? The dimension that is involved in different points on a response scale can change depending on the definition and location of the points. As a consequence, ordinal response scales (e.g., Likert scales) are not necessarily ordinal with respect to the latent trait(s) one intends to measure. IRT models can help to investigate whether response scales are ordinal or not.
2. Traits or types? The response consistency of respondents can vary as a function of the position on the latent trait. As a consequence the three-parameter or even the four-parameter IRT model may be needed without guessing being involved. It means that the degree of traitedness varies as a function of the position on the latent trait. Having a trait can then be understood as being located within a high consistency range on the trait continuum. This is one of the ways to reconcile the notions of traits and types.
3. Quantitative or qualitative? Personality disorders can be seen either as extreme positions on common personality dimensions (“quantitative difference”) or they can be seen as qualitative shifts to another kind of dimension (“qualitative difference”). The issue has been studied thus far with two designs: (a) two groups of subjects (extreme and less extreme) and one type of items, (b) one group of subjects (extreme) and two types of items (extreme and less extreme). A combined design is possible and IRT is a powerful tool to differentiate between the “quantitative” and “qualitative” hypotheses.

Parkzaal

Why Bayesian Psychologists Should Change the Way they Use the Bayes Factor

Herbert Hoijtink, *Utrecht University*

Bayesian psychologists had an important contribution (see, for example, Rouder and Morey, 2011, and Wagenmakers, Wetzels, Borsboom, and van der Maas, 2011) to the discussion following Bem’s (2011) research into psi. In this discussion, among other things, they promoted the use of the Bayes factor over the p-value for hypotheses evaluation. After reading their work and re-reading my own work on the Bayes factor (Hoijtink, 2012), I came to the conclusion that the Bayes factor, as it is currently used for the evaluation of point-null-hypotheses like a mean is zero, or the differences between two means is zero, may not be as good an improvement over the use of p-values as we all hoped for. In this presentation I will highlight the problems that we face and sketch avenues for further research that will render a Bayes factor for the evaluation of point-null-hypotheses that is an improvement over the use of p-values. Two problems with the Bayes factor for the evaluation of point-null-hypotheses as it is currently used will be highlighted:

- The prior distributions that are used for the computation of the Bayes factor are not well-calibrated, that is, prior distributions are chosen such that either the null or the alternative hypothesis is favored.
- Rules for the interpretation of the size of the Bayes factor are used that date back to Jeffreys (1961). As will be argued in this presentation, these rules render inappropriate quantifications of the strength of evidence in favor of either the null or alternative hypothesis.

Furthermore, two solutions to these problems will be proposed:

- The use of frequency arguments to calibrate prior distributions, that is, choose prior distributions such that the probability of a correct decision between null and alternative hypothesis is maximized. When doing this, strong features of both the Bayesian and frequentist philosophy of science are combined.

- The use of truly subjective prior distributions both for the parameters of the model under the null-hypothesis and the parameters of the model under the alternative hypothesis. This renders a fully subjective Bayesian approach.

16:10 - 17:00 Keynote session

Concertzaal **Truman Lee Kelley (1884--1961)**

Lawrence Hubert, *University of Illinois, Urbana-Champaign*

This talk is devoted to psychometric history in the first half of the 20th century, particularly as it involved Truman Lee Kelley. Kelley was one of the most prominent psychometricians (or, for that matter, statisticians) of this period. He was the fourth Psychometric Society President (1938-9; after Thurstone, Thorndike, and Guilford, but before Holzinger). I will emphasize Kelley's contributions that are less than well-recognized:

1) James-Stein (shrunken) estimators well before James and Stein (from his 1923 book, *Statistical Method*);
2) Principal component analysis at the same time as Hotelling (from his 1935 text, *Essential Traits of Mental Life*);
3) Canonical correlation and canonical variates at the same time as Hotelling (from his 1940 book, *Talents and Tasks*);
4) Unbiased estimation of the Correlation Ratio (1935, *PNAS*);
5) Asymptotic variance formulas for tetrad difference using the delta method (from his 1928 text, *Crossroads in the Mind of Man*);
6) Pentad conditions for factor analysis (also from the 1928 text*, Crossroads in the Mind of Man*)

(Spearman): Four variables may be thought of as due to one general factor plus four specific factors when $r_{12}r_{34} = r_{13}r_{24} = r_{14}r_{23}$ Or, when we have three tetrad differences being equal to zero.

(Kelley): Five variables may be thought of as due to two general factors plus five specific factors when a pentad criterion is zero. For terms such as $r_{12}\, r_{13}\, r_{24}\, r_{35}\, r_{45}$, six are added together and six subtracted.

17:00 - 17:10 Closing ceremony + Best junior presenter & Best poster award

Concertzaal

Sponsors

DPC
Dittrich & Partner Consulting GmbH
www.dpc-software.nl
stata

eMetric

CNTEST
海 云 天 教 育 测 评

海云天科技
SEA SKY LAND

GMAC
GRADUATE MANAGEMENT
ADMISSION COUNCIL

CollegeBoard
inspiring minds

LSAC

University of Amsterdam

Mc Graw Hill Education | CTB

measured progress

CITO now you know

IMPS 2013

The 78th Annual Meeting of
the Psychometric Society

Photo cover: Willem Maatman