IMPS 2015

International Meeting of the Psychometric Society

Beijing Normal University, China

Abstract Book: Talks

See Also:

- Abstract Book: Posters
- Program Book (printout included in conference package)
- Schedules for Monday, Tuesday, Wednesday (printouts included in conference package)
- Conference Website

Updated on July 9, 2015

Thanks to Anders Skrondal, Daniel Adams, Yanyan Fu, Wen-Chiang Lim, and Saed Qunbar for copy-editing

Contents

Monday, July 13

Keynote Speaker: Houcan Zhang

SH -1 Psychometrics in China: History and its development

Houcan Zhang, *Beijing Normal University, China*

The idea of individual difference and its measurement has a long tradition in China. About 3000 years ago, Confucius noticed individual differences among people and Mencius proposed that these differences can be measured. In the 7th century, China established the Civil Service Examination for personnel selection, which continued for more than 1000 years till early 20th century. The method of testing had also been spread to many other countries like Europe. China is thus called the birthplace of test.

Psychometrics, as a branch of psychological science, was introduced into China at the early 20th century. It was mainly disseminated in the field of education and went through a rather prosperous stage. Unfortunately, psychology was criticized and testing was forbidden because its belief in individual differences was in conflict with contemporary ideology in the 1950s. And it was totally denied as a discipline of science during the Cultural Revolution (1965-1976).

After the Cultural Revolution, psychology was revived. As for psychometrics, we started the first nationwide workshop on testing and psychological statistics in 1980 to train teachers of most universities. Then psychometrics regained its reputation through its successful application in the National College Entrance Examination reform, which promoted the development of psychometrics in education. Further, greater demand for mental tests emerged in the field of clinics and counseling.

As with the development of society, personnel selection reform also took place. Psychometrics was viewed as an invaluable tool for informing decision-making about employee and related organizations. It also contributed a lot to the field of counseling. In the late 1990s, psychology, especially psychometrics, became well-recognized and welcomed by the public. At the same time, we psychologists were aware of cultural difference and the urgent need for development of native tests.

In order to monitor the basic education quality and students' mental health to further inform education policy making, the Ministry of Education developed a comprehensive scale for educational assessment. Besides, the National Education Examinations Authority cooperated with the OECD and took a part in the PISA Program since 2006, and found many of its methods can be applied in China. The Shanghai Institute of Education led the same PISA program and gained excellent results.

To follow recent advanced development in psychometrics, besides its application, many studies in the field of Cognitive Diagnosis Theory, Computerized Adaptive Testing and advanced statistics can be found as well.

Parallel Sessions, Monday AM

YH Detecting Cheating on Tests [Invited Symposium] Organizer & Chair: Wim J. van der Linden

YH-1 Incorporating suspicious answer changes in the detection of aberrant response patterns

Arianto Wibowo, *Measurement Incorporated, USA*
Leonardo S. Sotaridona, *Measurement Incorporated, USA*

The l_z statistics (Drasgow et al., 1985) is considered one of the most powerful statistics to detect aberrant item response patterns. We propose a variant of l_z incorporating information about suspicious answer changes in the computation of the statistics; that is, when a test administrator is suspected of changing incorrect answers to correct. Given an observed item response vector $w = (w_1, w_2, \ldots, w_{nI})$, where $w_i \in \{0,1\}$, $P_i = P(w_i = 1)$, $Q_i = P(w_i = 0) = 1 - P_i$, $i \in \{1, 2, \ldots, nI\}$ and a response vector $v_i \in v_1, v_2, \ldots, v_{nI}$, $v_i \in \{0,1\}$ containing the responses from the examinee's initial attempt to answer the questions, the (new) statistics l_{zc} can be expressed as:

$$l_{zc} = \frac{l_{0c} - E(l_{0c})}{\sqrt{Var(l_{0c})}}$$

where

$$l_{0c} = \{\sum_{i=1}^{nI} w_i \ln P_i + (1 - w_i)\ln Q_i\}$$
$$+ \{\sum_{i=1}^{nI} v_i \ln P_{w_i,i} + (1 - v_i)\ln Q_{w_i,i}\}$$

The probabilities of answer changes, $P_{w_i,i}$, $i \in \{1, 2, , nI\}$, are conditional on the final response w_i, $P_{0,i} = P(v_i = 1|w_i = 0)$, $P_{1,i} = P(v_i = 1|w_i = 1)$, $Q_{0,i} = 1 - P_{0,i}$, $Q_{1,i} = 1 - P_{1,i}$. Expressions for the expectation and variance of l_{0c} can be derived. Notice that l_{0c} is comprised of two terms; the first term is the familiar (non-standardized) l_0 statistics (Levin & Rubin, 1979); the second term includes the information about the answer changes. The statistic l_{zc} is asymptotically standard normal; an alternative critical value is proposed for the case the number of items is too small expect normality. While the l_z statistic was designed to detect any type of aberrant response behavior, we are interested in determining whether or not the inclusion of the information about answer changes would improve its performance. The Type I error and detection rates of the proposed statistic for detecting aberrant responses due to cheating will be investigated in several simulated scenarios.

YH-2 Detection of item preknowledge in a high-stakes test

Dmitry I. Belov, *Law School Admission Council, USA*

Item preknowledge occurs when some examinees had access to a subset of compromised items from a prior administration. As a result, these aberrant examinees might perform better on compromised items as compared to uncompromised items. In general, item preknowledge is difficult to detect due to three unknowns: (i) unknown subgroups of aberrant examinees at (ii) unknown test centers who (iii) had access to unknown subsets of compromised items prior to taking the test. This research combines statistics, combinatorial optimization, and the structure of the Law School Admission Test (LSAT) in order to detect all three unknowns. Our algorithm works through three different corresponding stages. First, using pretest item parameters, it generates large random samples of item subsets from operational sections of the LSAT and computes cumulative person-fit statistics to detect affected test centers. A byproduct of this stage is the identification of subgroups of examinees at each affected test center with extreme values for the statistic. Second, each identified subgroup is used to compute an objective function for a combinatorial search of compromised items at the affected test centers. Third, for each affected test center and their compromised items, the aberrant examinees are detected. Advantages and limitations of the algorithm will be demonstrated using both simulated and real data.

YH-3 Exact null and posterior distributions for the detection of cheating

Wim J. van der Linden, *McGraw-Hill Education CTB, USA*

Methods for the detection of cheating typically calculate the probabilities of random suspicious responses on the test items and then use a statistical criterion to separate observed large numbers of suspicious responses from random instances. The goal of this paper is to show how both the null distributions and the posterior odds for the hypotheses of no cheating can be calculated through simple modifications of an extremely fast recursive algorithm in use for the calculation of n-fold discrete convolutions. The procedure will be illustrated using a few examples. If time permits, we will also address the use of (continuous) response times on the items for the detection of cheating, which generally involves the calculation of intractable convolutions but appears to allow for close approximation for the case of the loglinear response model proposed by the author.

YH-4 Detecting candidate preknowledge of items using a predictive checking method

Xi Wang, *University of Massachusetts Amherst, USA*
Yang Liu, *University of North Carolina at Chapel Hill, USA*
Ronald K. Hambleton, *University of Massachusetts Amherst, USA*

In continuous high-stakes computer-based testing programs such as GRE and TOEFL, many items are frequently used across test administrations. Item exposure provides an unfortunate opportunity for examinees to have knowledge of particular items in advance of their administration. This poses a threat to test security and could result in invalid test scores. A predictive checking method is proposed in this study to detect examinee preknowledge on exposed items. We consider a scenario where a test can be divided into two subsets of items: one consisting of secure items with very low exposure rates and the other consisting of possibly compromised items which have been exposed for a while. A Bayesian posterior distribution of an examinee's proficiency is first obtained from secure items, and then the corresponding predictive distribution for the examinee's test score on the unsecure items is constructed. The extent to which the unsecure items are compromised is determined by comparing the observed score on the unsecure items with the predictive distribution. This approach is also applied to flag a single problematic item: the statistic of interest is the difference between the proficiency estimate obtained from all items and that obtained from all but the item being examined. Five factors are manipulated through simulation: the choice of prior distribution, sample size of both sets of items, proportion of compromised items and item difficulty range. This study has implications for test quality control both after and during test administration, so as to enhance test security and ensure testing fairness.

Y3 New Developments in Network Psychometrics [Symposium] Organizer: Riet van Bork; Chair: Denny Borsboom

Y3-1 The factor-analytic disease: Etiology, diagnosis, and treatment

Denny Borsboom, *University of Amsterdam, The Netherlands*

In this talk, I will present a syndromal condition characterized by persistent psychometric delusions that give rise to patterns of dysfunctional scientific activity. The condition, which I argue has a non-negligible prevalence in psychometrics and is pervasively present in psychology, is characterized by the central conviction that the purpose of psychometrics is to subject all conceivable questionnaire data to latent variable models. This delusion gives rise to the compulsory execution of factor analysis, irrespective of its psychometric and substantive plausibility in a given situation. In addition to compulsive behavior, the factor analytic disease is typically associated with a set of more peripheral symptoms which can cause considerable scientific suffering: (a) the patient unwittingly shifts between mutually incompatible philosophical positions on the status of latent variables, (b) the patient is unable to say what the factors in the model refer to, but does not recognize this as a scientific problem, and (c) the patient is preoccupied with the question of whether there should be one, two, or three factors in the psychometric model, even in cases where there are clearly none. I will illustrate the condition by focusing on recent work in psychopathology research, which has concluded that there exists a higher-order "p-factor" analogous to the g-factor in intelligence research. Also, I will present preliminary evidence to show that even short exposure to products of alternative statistical modeling techniques, such as network graphs, can produce considerable improvement in the condition.

Y3-2 A new method for constructing networks from binary data

Claudia van Borkulo, *University of Amsterdam, The Netherlands*
Denny Borsboom, *University of Amsterdam, The Netherlands*
Sacha Epskamp, *University of Amsterdam, The Netherlands*
Tessa Blanken, *University of Amsterdam, The Netherlands*
Lynn Boschloo, *University Medical Center Groningen, The Netherlands*
Robert A. Schoevers, *University Medical Center Groningen, The Netherlands*
Lourens J. Waldorp, *University of Amsterdam, The Netherlands*

Network analysis is entering fields where network structures are unknown, such as psychology and the educational sciences. A crucial step in the application of network models lies in the assessment of network structure. Current methods either have serious drawbacks or are only suitable for Gaussian data. In the present paper, we present a method for assessing network structures from binary data. Although models for binary data are infamous for their computational intractability, we present a computationally efficient model for estimating network structures. The approach, which is based on Ising models as used in physics, combines logistic regression with model selection based on a goodness-of-fit measure to identify relevant relationships between variables that define connections in a network. A validation study shows that this method succeeds in revealing the most relevant features of a network for realistic sample sizes. We apply our proposed method to estimate the network of depression and anxiety symptoms from symptom scores of 1108 subjects. Possible extensions of the model are discussed.

Y3-3 Latent variable and network model implications for (partial) correlation structures.

Riet van Bork, *University of Amsterdam, The Netherlands*
Mijke T. Rhemtulla, *University of Amsterdam, The Netherlands*
Denny Borsboom, *University of Amsterdam, The Netherlands*

Network analysis is increasingly used to study psychological constructs that were previously studied using Latent Variable (LV) analysis. Whereas the latent variable approach introduces an unobserved common cause to explain correlations between the observed variables, the network approach portrays these correlations as direct causal relations between observed variables. These different approaches lead to contradicting views on how to understand the psychological constructs of interest. So far,

there is no method to decide whether a LV model or a network model is most likely to underlie the data. The present paper contributes to the development of such a method by exposing the implications for the correlation and partial correlation structure generated by these models. If a latent variable model underlies the data, the observed variables should constitute a fully connected network, and no partial correlation should equal zero. We visualize the correlation and partial correlation structure of several common LV models. Three promising findings emerge: (1) some LV models can be easily recognized from the correlational structure of the data; (2) other LV models (e.g., certain hierarchical models and correlated factor models) result in partial correlations very close to zero for which it is difficult to retrieve the common cause underlying these nodes; (3) for these models, simultaneous examination of both the correlation and partial correlation structure exposes the true underlying model.

Y3-4 Residual interaction modeling: Unifying SEM and network modeling

Sacha Epskamp, *University of Amsterdam, The Netherlands*
Mijke T. Rhemtulla, *University of Amsterdam, The Netherlands*
Denny Borsboom, *University of Amsterdam, The Netherlands*

In Structural Equation Modeling (SEM), latent factors are modeled as common causes to indicators that are assumed locally independent. In recent literature, This assumption of local independence has lead to critique of the factor model and its usage in psychology; local independence is seemingly violated often due to causal interaction between indicators (Borsboom, 2008). When local independence is violated, latent traits might not be needed to explain the correlational structure. Recently, network modeling has emerged in psychometrics in which the joint distribution of observed variables is modeled solely through a network of pairwise interactions (Epskamp, Maris, Waldorp, & Borsboom, in press). We propose a general framework that includes both SEM and network modeling, by modeling residual correlations in SEM as partial correlation networks. We term this framework Residual Interaction Modeling (RIM), which offers two main benefits. First, it offers a way to estimate a measurement model while taking into account that local independence is violated. As will be shown, the RIM model even allows all residuals to be correlated, while still being an estimable model with positive degrees of freedom. Second, the RIM model allows network analysts to estimate a partial correlation network while taking into account that some nodes are unobserved, possibly serving as common causes. RIM has been implemented in the "rim" package (https://github.com/SachaEpskamp/rim) for R, which includes both confirmatory and exploratory estimation methods as well as goodness of fit indices and model comparison tests.

H6 Equating in IRT

H6-1 An observed-score equating framework for latent variable models with polytomous data

Björn Andersson, *Uppsala University, Sweden*

Equating ensures that test scores from separate administrations of a particular standardized test are comparable. This presentation introduces a general framework for observed-score equating using latent variable models with polytomous data. It is shown how observed-score equating can be conducted using polytomous item response theory models and also using factor analysis. The asymptotic standard errors of equating using various polytomous item response theory models and factor analysis models are derived and examples of observed-score equating with the generalized partial credit model and the graded response model are provided.

H6-2 Composition of common items for equating with mixed-format tests

Stella Kim, *University of Iowa, USA*
Won-Chan Lee, *University of Iowa, USA*

Mixed-format tests including both Multiple-Choice (MC) items and Free-Response (FR) items have become more popular in real testing situations because the mixed format tests takes advantage of merits of both item types. However, in the measurement context, there are some challenges to including FR items in equating procedures due to the issues of a few number of FR items to be chosen as common items, item security, and rater leniency (Hagge, 2010; He, 2011; Muraki, Hombo, & Lee, 2000). Thus, it is common to consider MC items only in a common-item set even though the total test consists of both MC and FR items. This may raise the issue of representativeness of common items, which may introduce error into equating results. Limited research has been conducted to investigate the feasibility of using only MC items as common items in mixed-format test equating (Kim, Walker, & McHale, 2008; Lee, He, Hagge, Wang, & Kolen, 2012; Tan, Kim, Paek, & Xiang, 2009; Wang, 2013). However, none of these studies explore the acceptable level of common item proportion when using only MC items on mixed-format test to achieve the adequate equating results. Thus, in this study, several simulation conditions will be considered including the proportion of the common items, weights for MC and FR sections, and common item composition. The main purpose of this paper is to investigate effects of common item composition on equating the mixed-format tests.

H6-3 Selection of equating samples when the test taking population changes

Jinghua Liu, *Secondary School Admission Test Board, USA*
Albert Low, *Secondary School Admission Test Board, USA*
Keith Wright, *Secondary School Admission Test Board, USA*

When conducting score equating using an anchor test design, it is generally preferred to use similar ability groups regarding the construct to be measured between new- and old-form test taking populations, which yields robust equating (assuming other requirements for equating are met). For testing programs that have a stable and homogeneous population (e.g., the U.S. domestic population), composition of similar test taking populations can be achieved by carefully constructing an equating plan. For testing programs that have been evolving and becoming more heterogeneous, selecting equating samples can be a challenge. For example, a test is offered in a region (e.g., a foreign country) where no tests had been offered previously, which results in the new-form population substantially different from the old-form population. The purpose of this study was to explore the selection of equating samples when test taking populations change significantly between the new- and old-form: what should be done with respect to inclusion or exclusion of certain subgroup(s)? For illustration purposes, data was collected using the Secondary School Admission Test (SSAT), a standardized test designed for students who apply to independent schools in the US and Canada. Score linking was conducted on three different pairs of samples: total, domestic and international test takers. Score equity assessment (Dorans, 2004) was used to evaluate whether the linking relationship was subgroup sensitive. Conversions derived from different samples would be compared to each other and the historical trend would be analyzed.

H6-4 An empirical investigation of item response time stability among equating items

Huijuan Meng, *Graduate Management Admission Council, USA*
Xiaowen Zhu, *Xi'an Jiaotong University, China*

Common-item equating is a widely used method to place test scores from different test forms onto the same scale. Due to the statistical assumptions required for this approach (Kolen & Brennan, 2014), practitioners generally compare common-item statistics between test forms and remove any item from the anchor set if it functions differently from one test form to another. Several indices have been developed to screen common items for differences in difficulty and/or discrimination between test forms, such as the Mantel-Haenszel statistic (Michaelides, 2008), Rasch displacement measure, and Robust Z statistic (Huynh & Meyer, 2010). However, the stability of common-item response-time across years has rarely been examined. Although current equating approaches do not take item response-time into consideration, a significant response-time change may suggest that further item review is necessary. Therefore, the purpose of this study is to empirically examine fluctuations in equating-item response-time between any two adjacent years (from 2011 to 2014) for a somewhat speeded, aptitude-based testing program. Item-time parameters will be estimated with the hierarchical model (van der Linden, 2007) using the R-package cirt (Fox, Klein Entink & van der Linden, 2007). The pattern of response-time change can help item writers gain insights about the construct being measured – for example, a consistent response-time drop for items in a domain may indicate candidates' increasing familiarity with content covered by that domain. The findings may also help practitioners better understand how item response-time can contribute to the evaluation of equating-item functionality.

H6-5 Developing a multidimensional vertical scale of reading for understanding

Jonathan Weeks, *Educational Testing Service, USA*

Reading comprehension is a complex construct that involves the coordination of a number of theoretically integrated processes. The foundational skills in reading for understanding are not strictly hierarchical, nor do they develop in isolation. In order to examine changes in student ability on these skills over time there may be value in developing a multidimensional vertical scale. This type of scale is rarely, if ever, developed in practice given the large number of items needed to establish reliable factor scores and link the modeled dimensions across tests. This study uses data from a test developed to explicitly measure six reading component skills. The measure includes four forms with 204 items per form (22 to 50 items per component) an a 50% overlap in items between form pairs. The data include responses for 25,763 unique examinees in grades 6 through 9, collected over five administration cycles (Spring, Fall, Winter, Spring, Fall). Three multidimensional vertical scales and unidimensional vertical scale were created. The multidimensional scales correspond to three factor structures: six-factor simple structure, two-factor simple structure, and a bifactor structure with two specific factors. The models for these three structures fit better than the unidimensional model and have highly reliable factor scores. Changes in dimension-specific student abilities were examined across cycles and compared across the various multidimensional structures.

H1 Large-Scale Assessment

H1-1 Statistical projection through multidimensional latent regression: Linking between large scale educational survey assessments

Yue (Helena) Jia, *Educational Testing Service, USA*
Xueli Xu, *Educational Testing Service, USA*

Linking and scale alignment are common practices in educational programs. A number of statistical approaches have been applied in linking within educational programs and between different programs. However, little is known about what a general approach would be in linking between different educational survey assessments due to their uniqueness in inferring proficiency estimates. This study will focus on a latent regression approach and its application in linking two large scale survey assessments, which has the advantage of directly taking into account the relationship between the two assessments. Large scale educational survey assessments, such as the National Assessment of Educational Progress (NAEP) and the Trends in International Mathematics and Science Study (TIMSS), commonly use a combination of Item Response Theory (IRT) models and latent regression population models to estimate distributions of underlying proficiency for student groups of interest. To link between different survey assessments of such, a regression type of procedure was developed to model the relationship between proficiencies yielded from these two assessments via a bi-variate latent regression approach. This model was then used to project scores from one assessment onto the scale of the other assessment. The procedure was applied in linking NAEP to TIMSS in their 2011 administration.

H1-2 Using sample weights in item response data analysis under complex sample designs

Xiaying Zheng, *University of Maryland, College Park, USA*
Ji Seung Yang, *University of Maryland, College Park, USA*

Large-scale assessments (e.g., Programme for International Student Assessment (PISA) or Trends in International Mathematics and Science Study (TIMSS)) are often conducted using complex sampling designs that include the stratification of a target population and multistage cluster sampling. To address the nested structure of item response data under the complex sample designs, a number of previous studies proposed multilevel/multidimensional item response models (e.g., Fox, 2010). However, incorporating sample weights into the item response models has been less explored. The purpose of this study is to assess the performance of three different methods to analyze item response data that are collected under complex sample designs: 1) the design-based (aggregate) method, 2) the model-based (disaggregate) method, and 3) the hybrid method that addresses both the multilevel structure and sampling weights. A Monte Carlo simulation study is carried out to see whether the hybrid method can yield the least biased item/person parameter and level-2 variance estimates under different conditions (e.g., the size of clusters and intraclass correlation, and item response models). Item response data are generated using the complex sample design that is adopted by PISA, and bias in estimates and adequacy of standard errors are evaluated after applying three methods. As an empirical data illustration, a subset of PISA is analyzed and the results are compared. The results highlight the importance of using sample weights in item analysis when a complex sample design is used.

H1-3 Impact of item cluster position on parameter estimation using an international assessment

Rongchun Zhu, *ACT, USA*
Xiaohong Gao, *ACT, USA*

This study aims to evaluate the impact of item cluster position on item calibration and examinee proficiency estimation based on the PISA 2012 program. International educational assessments that use a standardized testing approach for international comparisons provide a valuable opportunity to evaluate important measurement issues across varied national samples. Furthermore, sophisticated design of test material samplings is rigorously implemented to achieve a randomly equivalent groups design, which makes it possible to directly connect measurement design element and outcome. This study focuses on the impact of item cluster position. When identical items appear in different positions on a test, examinee performance may differ. Item calibration using a full sample that includes item responses from different positions generally balances out any item position effect, but the estimation of item parameters may be biased in extreme cases. In paper and pencil testing of PISA 2012, each student was randomly assigned to complete one of 13 booklets. Each booklet includes four out of 13 item clusters (including seven mathematical clusters) in a fixed sequence where each cluster is placed in different positions across the booklets. Five mathematical clusters will be analyzed based on the groups who took them either as the first cluster or the last cluster in the corresponding booklets. Item parameter and examinee proficiency estimates will be compared using different item response theory models, including Rasch, 2-PL, and 3-PL models. The implications of item positions and parameter estimation for test programs are discussed in the end.

H5 Longitudinal Models

H5-1 Fitting multi-timescale differential equation models of adaptive equilibrium regulation

Steven M. Boker, *University of Virginia, USA*

Adaptive Equilibrium Regulation (AER) separates regulation processes into two timescales: a faster regulation that automatically balances forces and a slower adaptation process that reconfigures the fast regulation so as to move the system towards its preferred equilibrium when environmental forces persist. AER models can be used to account for interactions between different timescales of behavior and development, practice and learning, or regulation and adaptation. We present a method for multiple timescale time delay embedding and through simulation show how multiple timescale processes can be fit with latent differential equations structural models.

H5-2 Dynamic process for ordinal data: Latent differential equation modeling with threshold

Yueqin Hu, *Texas State University, USA*
Steven M. Boker, *University of Virginia, USA*

Ordinal data are widely used in psychology research. There are statistical models to deal with ordinal data, such as the probit model and the logit model. However, for data with complex structures, for example longitudinal data, ordinal variables are frequently treated as continuous. This study aimed to identify methods for modeling longitudinal ordinal data. The threshold probit model was combined with the latent differential equation model as one solution to analyze longitudinal ordinal data. The bias caused by fitting differential equation models to ordinal data was evaluated. Simulation results suggested that the naive model, which blindly treats ordinal data as continuous data, led to bias, especially when the ordinal data have very few levels and the levels are divided by unequal intervals. In comparison, the threshold model was unbiased under binary data condition and had less bias in other simulation conditions. Moreover, the naive model suffered from alpha inflation in all simulation conditions, whereas the threshold model did not. Therefore, for ordinal data with very few levels, we recommend to use latent differential equation models with threshold.

H5-3 Measurement error, reliability and stability in multilevel autoregressive modeling

Noémi Schuurman, *Utrecht University, The Netherlands*

More and more researchers in psychology are collecting intensive longitudinal data, such as daily diary and ESM data, in order to study psychological processes on an intraindividual level. An increasingly popular way to analyze these data is autoregressive time series modeling; either by modeling the repeated measures for a single individual using classic $n = 1$ autoregressive models, or by using multilevel extensions of these models, with the dynamics for each individual modeled at level 1 and interindividual differences in these dynamics modeled at level 2. However, while it is widely accepted in psychology that psychological measurements usually contain a large amount of measurement error, the issue of measurement error is largely neglected in psychological (autoregressive) time series modeling. The regular autoregressive model incorporates innovations, or ‘dynamic errors’, but not what can be considered measurement errors. We will discuss the distinction between innovations and measurement errors, and the consequences of disregarding measurement error in the autoregressive model. Further, we will present multilevel autoregressive models that can account for measurement errors and allow us to investigate individual differences in reliability. We illustrate these issues and models using an empirical application.

H5-4 Inferring longitudinal relationships between variables: Model selection between the latent change score and autoregressive cross-lagged factor models

Satoshi Usami, *University of Tsukuba, Japan*
Timothy Hayes, *University of Southern California, USA*
Jack McArdle, *University of Southern California, USA*

The present research focuses on the model selection problem, comparing the Latent Change Score (LCS) model and the Auto-Regressive Cross-Lagged (ARCL) model to infer the longitudinal relationship between variables. Usami et al. (2015) have shown that the LCS model reduces to an ARCL model when the constant change factor scores are assumed to be invariant across people. In the present study, a large-scale simulation is conducted in order to: (a) investigate the conditions under which these models return statistically (and substantively) discrepant results concerning the bivariate longitudinal relationships, and (b) ascertain the relative performance of an array of model selection procedures when such discrepant results arise. The results show that primary sources of differences in parameter estimates are model parameters related to the constant change factor scores in the LCS model (specifically, the correlation between the intercept and the constant change factor scores) as well as the size of the data (specifically, the number of time points and sample size). Among several model selection procedures, correct selection rates were higher when using model fit indices (i.e., CFI, RMSEA) than when using a likelihood-ratio test or any of several information criteria (i.e., AIC, BIC, CAIC and ssBIC). In closing, we show two actual examples of model selection using weight and height data from gerontological and psychological research.

H5-5 Piecewise latent growth modeling with random change-point

Chuyi Yu, *Beijing Normal University, China*
Xuelian Bao, *Beijing Normal University, China*
Xiaobo Wang, *Beijing Normal University, China*
Shumei Zhang, *Beijing Normal University, China*
Hongyun Liu, *Beijing Normal University, China*

Latent Growth Models (LGMs) can be used to describe the growth trajectory in longitudinal data analysis. The Piecewise LGM (PLGM), an extension of the LGM, allows the specification of each growth phase to conform to a particular functional form of the overall change process. An interesting feature of the PLGM is the time point at which the response function transitions from one phase to another, known as the change point. The change point can be known a priori or can be estimated, but the change point is always considered to be a constant for all individuals. This assumption is not practical in situations where different persons transition at different time points, and the change points for different individuals is a variable which follows a random distribution (e.g., normal distribution). Considering this point, a new Piecewise Latent Growth Model with Random Change-Point (RCP-PLGM) is proposed, which can be viewed as a generalized PLGM. Specifically, the change-point is assumed to follow a normal distribution in the RCP-PLGM, and the EM algorithm is used to estimate the growth parameters and distribution parameters of the change-point in the RCP-PLGM. Simulation studies for the RCP-PLGM are conducted and the estimation accuracy of the parameters is obtained. In addition, the RCP-PLGM is illustrated by analyzing a longitudinal study of early childhood reading ability. The results of the empirical data analysis provide the key information of the change-point for childhood reading ability.

XH Cognitive Diagnosis Models I

XH-1 A cognitive diagnosis model using multiple category scoring for constructed response items

Bor-Chen Kuo, *National Taichung University of Education, Taiwan*
Chun-Hua Chen, *National Taichung University of Education, Taiwan*
Mo Ching Magdalena Mok, *The Hong Kong Institute of Education, Hong Kong*

The aim of this study is to develop a cognitive diagnostic model to analyze tests with Constructed Response (CR) items. The problem solving process of CR item transferred to categorical data can provide more information than dichotomized data. However, most cognitive diagnostic models modeling categorical or dichotomized data are proposed for tests with Multiple Choice (MC) items. In this study, an extended DINA model using multiple category scoring is proposed for modeling categorical data obtained from CR items. Each response category of categorical data is a combination of multiple skills measured by the item. The Expectation-Maximization (EM) algorithm is applied to estimate the parameters in the proposed model. Finally, results of simulation studies and analyses of real data are presented and discussed.

XH-2 Bayesian model checking methods for cognitive diagnosis models

Jung Yeon Park, *Columbia University, USA*
Matthew S. Johnson, *Columbia University, USA*
Young-Sun Lee, *Columbia University, USA*

Cognitive Diagnosis Models (CDMs) are a type of multidimensional model that assume each item in an assessment measures a small number of discrete cognitive skills.

These models aim to diagnose and categorize each examinee using fine-grained skill patterns and cover a wide range of sub-models from conjunctive models such as the deterministic, inputs, noisy "and" gate model (DINA; Junker & Sijtsma, 2001), to compensatory models such as the compensatory General Diagnostic Model (GDM; von Davier, 2005). Because of the multidimensionality of the latent skill space, several structural models have also proposed (e.g., higher-order DINA; de la Torre & Douglas, 2004). Despite novel approaches of the aforementioned models, there is limited research regarding model-data misfit. For this reason we are motivated to investigate various circumstances where such model-data misfit occurs. Models considered in the study include DINA models with different structures (e.g., saturated, independent, and higher-ordered), and compensatory GDMs with different mapping methods that determine relationships between Q-matrix and skill patterns. To evaluate model-data fit, we employ a Bayesian model checking method, namely Posterior Predictive Model Checks (PPMC; Gelman, Meng, & Stern, 1996; Sinharay, Johnson, & Stern, 2006). An important issue with the application of PPMCs is the choice of discrepancy measures. In this study, we propose several discrepancy measures – item vs. item associations, correlation between attributes and log-likelihood. A simulation study with varying test lengths and sample sizes and a real data application demonstrate the effectiveness of the suggested model checking criteria.

XH-3 Detecting aberrant examinees in cognitive diagnosis models

Kevin Carl Santos, *University of the Philippines Diliman, The Philippines*
Jimmy de la Torre, *Rutgers University, USA*
Erniel Barrios, *University of the Philippines Diliman, The Philippines*

The mastery or non-mastery of required attributes in answering a test may not necessarily be reflected in the test scores due to examinees' aberrant response behavior such as guessing and cheating. Such misfit of response patterns to a psychometric model has an effect on classification decisions, item parameter estimates, and goodness-of-fit statistics. In this study, a forward search algorithm is being proposed in detecting aberrant response patterns of examinees assuming a DINA (deterministic inputs, noisy "and" gate) model. A nonparametric procedure based on ideal response patterns is being proposed in choosing the initial set. To progress with the forward search, the likelihood contribution of the examinee serves as the criterion. Forward plots of goodness-of-fit statistics, guess and slip parameter estimates, and Cook's distance are used to monitor the forward search. Simulated datasets were generated to illustrate the viability of the proposed method.

XH-4 A general nonparametric classification method for cognitive diagnosis

Yan Sun, *Rutgers University, USA*
Chia-Yi Chiu, *Rutgers University, USA*

The NonParametric Classification (NPC) method for cognitive diagnosis (Chiu & Douglas, 2013) estimates attribute patterns by minimizing the distance between observed and all possible ideal item responses. The NPC estimator is proven to be statistically consistent when data conform to the DINA, DINO, NIDA model or the Reduced RUM (Wang & Douglas, 2013). However, when the data conform to the NIDA model or Reduced RUM, extra assumptions are needed to guarantee the consistency. It is thus in demand to develop a general version of the NPC method that can be used for more complex models without any extra assumptions to maximize its applicability. The General NonParametric Classification (GNPC) method for cognitive diagnosis adopts a weighted ideal response, which is a linear combination of the conjunctive and disjunctive ideal responses. Examinees' attribute patterns and the weights are estimated simultaneously by minimizing the difference between the observed responses and the weighted ideal response. In the study, data generated from the DINA model or the Reduced RUM are classified using the GNPC method and MMLE method via the EM algorithm. The performance of the GNPC method is examined by comparing the pattern-wise and attribute-wise classification rates with those obtained by using the EM algorithm.

XH-5 An extended multiple-strategy DINA model for strategy estimation

Chih-Wei Yang, *National Taichung University of Education, Taiwan*
Bor-Chen Kuo, *National Taichung University of Education, Taiwan*
Chun-Yen Cheng, *National Taichung University of Education, Taiwan*

The aim of this study is to develop a cognitive diagnostic model to analyze not only cognitive attributes but also problem solving strategies. A cognitive diagnostic model, Multiple-Strategy DINA (MS-DINA), which allows examinees can use various strategies, has been proposed. Nevertheless, MS-DINA only concerns and reports the mastery of skills. But, in remedial instruction, considering both strategy and cognitive attributes would be more appropriate than only cognitive attributes. This study proposes an extended MS-DINA model to estimate the mastery information of strategies and skills simultaneously. An Expectation-Maximization (EM) algorithm is applied to estimate the parameters in the proposed model. Finally, results of simulation studies and analyses of real data are presented and discussed.

X3 Model Fit/Selection in IRT

X3-1 Towards effect size measures for local item dependence

Johan Braeken, *University of Oslo, Norway*

Although educational tests are commonly aimed to assess a theoretically unidimensional ability construct, it is not uncommon to have test blueprints that systematically build in task interdependence by design. Some argue that this is even necessary to create authentic assessments (e.g., problem solving) or a conceptual part of the actual construct (e.g., reading passages in the assessment of reading comprehension). This viewpoint clashes with the strict statistical definition of unidimensionality in a psychometric instrument that requires local stochastic independence between the items. Intentional (or nonintentional) task-interdependence can give rise to empirical Local Item Dependence (LID) that is unaccounted for by the psychometric model, which in turn may complicate measurement applications by distorting and biasing model parameters, uncertainty estimates, and classification decisions. The current study explores the use of copula models (Fréchet-Hoeffding mixtures & Gaussian copulas) to test and detect local item dependence. Special interest goes out to investigating to what extent minor violations of LSI are tolerated and whether it is possible to effectively quantify when LID is expected to have no major impact on model inferences. Monte Carlo simulation results are provided to verify the power and false discovery rate of the tests, but also to provide a range of expected values for the proposed effect size measures under the unidimensional reference model.

X3-2 Latent variable model selection for binary response data with application in educational and psychiatric data

Yunxiao Chen, *Columbia University, USA*

In this paper, we propose a family of models that is flexible and jointly models the latent and observed variables in a parsimonious way. This modeling framework combines latent factor models and graphical models (Pearl, 1988; Dawid & Lauritzen, 1993) to capture a dependence structure not attributable to the latent variables. The rationale is that the latent variables globally drive the dependency among all the observed variables and a sparse graphical model locally characterizes the dependency between the observed variables unexplained by the latent variables. In addition, model estimation is obtained through maximizing a regularized pseudo-likelihood function via a convex optimization algorithm. Using this algorithm, we are able to handle large scale data sets that may contain millions of subjects and hundreds of items. Methods are developed to visualize the estimated sparse graphical model, which helps people understand the dependence structure unexplained by the latent variables. Finally, the model is applied to several educational testing and psychiatric assessment data sets.

X3-3 Limited information goodness-of-fit testing of IRT models in the presence of planned missing data

Seungwon Chung, *University of California, Los Angeles, USA*
Li Cai, *University of California, Los Angeles, USA*

Questionnaires are used widely in social science research, but due to various reasons (efficiency, testing time, cost, respondent burden, etc.) each respondent may only be exposed to a subset of all available items, chosen randomly. An example is planned missing data designs (e.g., achieved with randomized incomplete block designs) frequently encountered in large-scale educational surveys. With full-information maximum likelihood item parameter estimation, IRT-based item analysis is generally not difficult in the presence of planned missing data, but the incomplete data matrix makes ascertaining the degree of model fit somewhat more complex because commonly used goodness of fit statistics (e.g., Maydeu-Olivares & Joe's M_2) have been studied only under complete data contexts. This study investigates the properties of overall goodness-of-fit statistics such as M_2 and Cai and Hansen's (2012) M_2^* in testing IRT model fit for planned missing data designs. A simulation study will be conducted to assess whether the M_2-type statistics remain applicable and calibrated in planned missing data designs and whether they can detect model misspecification. The planned missing data designs being considered in this study are the three-form design (Graham et al., 1996), varied by covariance coverage, and the matrix sampling design, where there is no guarantee of covariance coverage. Model misspecification is introduced using the Tucker, Koopman, & Linn (1969) procedure. This study will provide information to help address a practical issue that frequently arises in applications of IRT analysis.

X3-4 Summed score likelihood based statistics for detecting latent variable distribution fit in full-information item bifactor models

Zhen Li, *University of California, Los Angeles, USA*
Li Cai, *University of California, Los Angeles, USA*

Latent variables in a full-information item bifactor model are often assumed to follow a multivariate normal distribution. When the assumption is violated, the estimation of item parameters, latent variable means and variances can be biased. To detect latent variable distributional assumption violations, Li & Cai (2012) developed a series of summed score likelihood based indices. The statistics enjoyed high statistical power in unidimensional IRT models and were not sensitive (correctly) to other forms of model misspecification (Li & Cai, 2012). The statistical indices, however, were technically not exactly chi-square distributed. In this study, we extend Li & Cai's (2012) summed score likelihood based index $\overline{X}^2$ for detecting the latent variable distribution fit in item bifactor models. Additionally, a Satorra-Bentler type moment adjustment (Satorra & Bentler, 1994) is applied, so that the statistic more closely follows a chi-square distribution. The corrected statistic is referred to as $\overline{X}^2_C$. A simulation study and an empirical data analysis will be carried out to examine the performance of $\overline{X}^2_C$. It is conjectured that $\overline{X}^2_C$ can be better approximated by its purported chi-square distribution than $\overline{X}^2$ in null conditions. When the generating general factor of item bifactor models is non-normally distributed, $\overline{X}^2_C$ has larger power to detect the model misfit than $\overline{X}^2$ and the limited-information overall fit statistic M_2 (Maydeu-Olivares & Joe, 2005).

X3-5 Model fit evaluation in multilevel-multidimensional item response models: Sensitivity to model misspecification

Dandan Liao, *University of Maryland, College Park, USA*
Ji Seung Yang, *University of Maryland, College Park, USA*

Multilevel-multidimensional Item Response Theory (IRT) models have been developed to address the nested structure of item response data from the complex sampling or repeated measures designs (e.g., Fox, 2007). As a specific case of multilevel structural equation models with categorical manifest variables, a multilevel-multidimensional IRT model is composed of the measurement model that describes the relationship between abilities and observed responses to items, and the structural model that describes the relations among the latent abilities. While the high-dimensional models have been rapidly developed along with efficient computation algorithms (e.g., Metropolis-Hastings Robbins-Monroe (MH-RM) algorithm; Cai, 2009), the model fit indices that provide information about the overall goodness of data-model fit have been less explored within the framework of multilevel-multidimensional IRT models. The purpose of this study is to evaluate the performance of currently available relative model fit indices (e.g., log-likelihood, AIC, and BIC) and to suggest an alternative approach in the framework of multilevel-multidimensional IRT models. A simulation study is conducted to evaluate model fit indices under different misspecification scenarios where the misspecification is imposed on measurement and/or structural models. Full information maximum likelihood estimation with two algorithms (MH-RM and Expectation-Maximization (EM)) is used to estimate the parameters of interest. The simulation study results show that the current fit indices do not provide sufficient information about where the improper model fit actually comes from. More specific model fit indices are needed to serve the need of identifying level-specific misspecification.

X7 Statistical Inference

X7-1 Approximate Bayes factor for informative hypotheses

Xin Gu, *Utrecht University, The Netherlands*
Herbert Hoijtink, *Utrecht University, The Netherlands*
Joris Mulder, *Tilburg University, The Netherlands*

Psychological researchers often have expectations of the structure among model parameters. Their expectations can be formulated using informative hypotheses, which contain equality constraints ($\theta = 0$), inequality constraints ($\theta > 0$), about equality constraints ($\theta \approx 0$) and range constraints ($.5 < \theta < 1$) among parameters. Bayesian hypothesis testing evaluates informative hypotheses using Bayes factor, which requires the prior specification. This presentation will introduce two default methods to specify proper priors that are data-based. Therefore, this results in an objective Bayesian procedure of hypothesis testing. Furthermore, we approximate the posterior distribution by multivariate normal distribution such that the Bayesian approach for selecting the best of a set of informative hypotheses can be applied in general statistical models. We provide a software package for psychological researchers to evaluate informative hypotheses using Bayes factor. An example is given to illustrate how our Bayesian procedure works and how to use the software.

X7-2 Some thoughts concerning the influence of nuisance parameters on conditional inference

Clemens Draxler, *The Health and Life Sciences University, Austria*

Statistical theory provides a number of approaches to deal with nuisance parameters. One of them is the procedure of Conditional Maximum Likelihood (CML). The nuisance parameters are eliminated from the likelihood function by conditioning on their sufficient statistics. While the CML estimates of the parameters of interest are not affected by the nuisance parameters, this talk points to the dependence of the asymptotic covariance (of the parameters of interest) on the probability distributions of the sufficient statistics (for the nuisance parameters) and thus on the nuisance parameters themselves. The practical implications of this issue are discussed in the context of sample size planning and power computations of statistical tests derived from the CML estimator with focus on the family of Rasch models.

X7-3 The use of deviance plots for non-nested model selection in loglinear models, structural equations, and three-mode analysis

Pieter M. Kroonenberg, *Leiden University, The Netherlands*

Mallows defined the C_p statistic with an associated C_p plot to be used in model selection in regression analysis. The deviance plot is an extension of this idea, where the loss, expressed in residual sum of square, or the Chi-squared statistic is graphed against the degrees-of-freedom, thus allowing for comparing deviance/df ratios between models. It is shown that both the RMSEA (Root Mean Squared Error of Approximation) and the AIC (Akaike's Information Criterion) are lines in the Chi-squared deviance plot so that these criteria can be used together. Moreover work has been done for automatic selection by finding the point where the convex hull of the optimal solution shows the largest curvature. Brief examples for different methods will be provided.

X7-4 Generalized fiducial inference for graded response models

Yang Liu, *University of North Carolina at Chapel Hill, USA*
Jan Hannig, *University of North Carolina at Chapel Hill, USA*

Generalized Fiducial Inference (GFI) has been proposed as an alternative inferential framework in the statistical literature. Inferences of various sorts, such as confidence regions for (possibly transformed) model parameters, making prediction about future observations, and goodness of fit evaluation, can be constructed from a fiducial distribution defined on the parameter space in a fashion similar to those used with a Bayesian posterior. However, no prior distribution needs to be specified. In this work, the general recipe of GFI is applied to the graded response models, which are widely used in psychological and educational studies for analyzing ordered categorical survey questionnaire data. Asymptotic optimality of GFI is established, and a Markov chain Monte Carlo algorithm is developed for sampling from the resulting fiducial distribution. The comparative performance of GFI, maximum likelihood and Bayesian approaches is evaluated via Monte Carlo simulations. The use of GFI as a convenient and powerful tool to aid quantification of sampling variability in various inferential procedures is illustrated by an empirical data analysis using the patient-reported emotional distress data.

Parallel Sessions, Monday PM

YH Publication Trade Secrets: Straight from the Editors' Mouths [Invited Session] Organizer & Chair: Jimmy de la Torre

Panel Discussion and Q&A with Journal Editors:

Applied Measurement in Education
Kurt Geisinger, *University of Nebraska - Lincoln, USA*

Applied Psychological Measurement
Hua-Hua Chang, *University of Illinois at Urbana-Champaign, USA*

British Journal of Mathematical and Statistical Psychology
Matthias von Davier, *Educational Testing Service, USA*

Journal of Classification
Willem Heiser, *Leiden University, The Netherlands*

Journal of Educational and Behavioral Statistics
Li Cai, *University of California, Los Angeles, USA*

Journal of Educational Measurement
Jimmy de la Torre, *Rutgers University, USA*

Measurement: Interdisciplinary Research and Perspectives
Paul De Boeck, *The Ohio State University, USA*

Psychometrika
Irini Moustaki, *London School of Economics, UK*

Y3 Validity, Generalizability & Scoring

Y3-1 Degrees of test validity: Some new simulation results

Keith Markus, *The City University of New York, USA*

Many theories of test validity assume that validity comes in degrees but fail to explain the basis for such degrees. The Deductive Strength (DS) and Belief Centrality (BC) accounts provide contrasting accounts of degrees of validity. DS focuses on minimally acceptably supported claims in the construct theory and ranks construct theories in terms of the range of inferences that can be drawn from empirically supported elements of them. BC focuses on the centrality of various elements of the construct theory to the overall network of beliefs and ranks construct theories in terms of resilience to contradictory information. Simulation results focus on operationalizing BC in terms of the reinforcement of item validity claims by other types of evidence for the same item and similar evidence for related items. BC drops beliefs that do not cohere with other beliefs in these ways. DS and BC succeeded comparably in choosing the better of two tests. BC leads to more accurate beliefs overall than does DS. When support is associated across types of evidence within items, BC outperforms DS with respect to sorting items by the minimally supported form of evidence. However, both perform similarly when the association is absent, or items are sorted by the number of supported types of evidence.

Y3-2 Component universe score profile analysis: A new method of profile analysis

Joseph H. Grochowalski, *Fordham University, USA*
Se-Kang Kim, *Fordham University, USA*

Component Universe Score Profile Analysis (CUSP) is introduced as a new method of subscore profile analysis. CUSP provides useful information about subscores that is otherwise discarded in composite (total) score analysis and is a way for subscores to potentially have value even when they have low reliability. CUSP extracts typical multidimensional subscore patterns from subtest data. One pattern extracted is the mean (flat or level) profile, which is often preferred because it has higher reliability and predictive utility than individual subscores. Other profiles are pattern profiles, which also have higher reliability than individual subscores, but contain information orthogonal to (and different from) the overall mean score. Thus CUSP recovers some useful information from subscores that is not available in a total score or mean score analysis. CUSP is based on generalizability theory and singular value decomposition, both of which can be unstable with small samples, so the stability and bias of CUSP estimates were assessed in a simulation study. The results show that CUSP was overall very stable and accurate at all sample sizes, with the exception of conditions where subscores had a compound symmetric covariance structure. Reliable CUSP profile scores were also estimated for subscores that had low reliability. CUSP analysis may be a method that helps bridge the debate on subscore versus total score interpretation.

Y3-3 Optimal measurement procedures in generalizability theory

Yon Soo Suh, *Yonsei University, South Korea*
Dasom Hwang, *Yonsei University, South Korea*
Quan Meiling, *Yonsei University, South Korea*
Guemin Lee, *Yonsei University, South Korea*

The purpose of this study is to investigate optimal measurement procedures for producing the maximum reliability coefficient under several constraints of resources, specifically limited cost, items, and raters. Designing a specific measurement procedure incurs several different kinds of cost considerations, and test developers must stay within realistic conditions when designing a test and consider practical limitations such as restricted budget, number of items, and number of possible raters. Optimization procedures make it possible to pinpoint the most feasible number of observations of a measurement design within constrains of pre-specified resources. Allocating these resources in different ways will result in disparate levels of reliability. In this study, we apply nonlinear constraint optimization with multiple constraints of specified resources by the branch and a bound algorithm to obtain maximum reliability. We aim to locate the optimal numbers of observations per measurement facet that produces the highest reliability coefficient under specific cost, item and rater constraints and to compare the results to examine which provides the most preferable results for the same amount of resources.

Y3-4 Clustered factor score identification in matrix-factorization factor analysis

Kohei Uno, *Osaka University, Japan*
Kohei Adachi, *Osaka University, Japan*

Factor analysis – which is a time-honored dimension-reduction method – has two types of indeterminacies. The one which is well-known is rotational indeterminacy of the factor loadings, while the other type is indeterminacy of the factor scores: they cannot be uniquely identified. The first indeterminacy is exploited to obtain interpretable loadings in rotation methods. In this paper, we consider exploiting the indeterminacy of factor scores to find interpretable scores in the recently proposed framework of Matrix-factorization Factor Analysis (MFA). In MFA, the matrix of common and unique factor scores is parameterized as a block-matrix. This matrix can be partitioned into a uniquely determined the matrix and an undetermined matrix. We propose a method for choosing the undetermined matrices so that common factor scores are well classified into a small number of clusters. The usefulness of the proposed method is illustrated with a real data example.

Y3-5 Random or mixed-effect models should be used in generalizability theory

Jinming Zhang, *University of Illinois at Urbana-Champaign, USA*
Chih-Kai (Cary) Lin, *Center for Applied Linguistics, USA*

In Generalizability theory (G theory), facets are typically assumed to be random because G theory intends to generalize results towards the underlying universe. In an essay-rater study, for example, the raters are considered exchangeable with other raters in the universe of all qualified raters and the rater facet is generally treated as random, although raters are not, in practice, randomly selected from the universe. When a facet is treated as random, one actually assumes that there is no subject-by-facet (or specifically, essay-by-rater, in the example) interaction and consequently, the model is additive, as demonstrated in this study. In this study, we show that when subject-by-facet interactions exist, but regular G theory formulas based on random-effect models are used, some Variance Components (VCs) are underestimated and consequently generalizability coefficients are also underestimated. Instead, we demonstrate that mixed-effects non-additive models should be used in such a case, and derive formulas for the estimators of VCs for non-additive models. Thus, depending on the existence of interactions, an appropriate model, either additive or non-additive, should be used in applications of G theory. The non-additive G theory developed in this study generalizes the current G theory, and uses data at hand to determine when additive or non-additive models should be used to appropriately estimate VCs and generalizability coefficients. A simulation study is conducted to confirm the theoretical results. Finally, the implications of the findings are discussed in light of an analysis of real data.

H6 Response Time Models

H6-1 Modeling conditional dependence between response time and accuracy

Maria Bolsinova, *Utrecht University, The Netherlands*
Paul De Boeck, *The Ohio State University, USA*
Jesper Tijmstra, *Tilburg University, The Netherlands*

While the assumption of conditional independence is crucial for hierarchical modeling of response times and accuracy, this assumption may be violated in some cases. That is, it may be that the relationship between the response time and the response accuracy of the same item cannot be fully explained by the correlation between the overall speed and the ability. We propose to explicitly model the residual dependence between time and accuracy by incorporating the effects of the residual response time on the parameters of the IRT model for response accuracy. We present an empirical example of a violation of conditional independence from a low-stakes educational test and show that our new model reveals interesting phenomena about the dependence of the item properties on whether the response is relatively fast or slow. For more difficult items responding more slowly increased the probability of a correct response, whereas for the easier items responding more slowly decreased the probability of a correct response. Moreover, for many of the items slower responses were less informative for the ability because their discrimination parameters decrease with residual response time.

H6-2 Online calibration for a joint model of responses and response times in CAT

Hyeon-Ah Kang, *University of Illinois at Urbana-Champaign, USA*

Online calibration has been drawing great attention among many testing organizations and psychometric research as a technique to efficiently replenish an item bank during operational CAT administrations. This study develops online calibration strategies that incorporate collateral information in response times to optimally calibrate pretest items in CAT. Varying levels of exploitation of collateral information in response times are proposed. Parameters of pretest items are calibrated on-the-fly through marginal maximum likelihood estimation method via multiple expectation-maximization algorithm. Simulation studies are conducted to examine the feasibility of response time-aided online calibration as well as to investigate statistical gain of employing the auxiliary variable in item calibration. Findings of this study provide implications for online calibration strategies in the presence of data sparseness such that both the precision and bias of parameter estimates can be improved by using the collateral information in response times.

H6-3 Bayesian analysis of joint modeling of response times with dynamic latent ability

Abhisek Saha, *University of Connecticut, USA*

In measurement testing, inferences about latent ability of test takers have been mainly based on their responses to test items while the time taken to complete an item has been often ignored. With the advent of computerized testing, it becomes much easier to collect the response time of each item without additional cost. The separate analysis of response accuracy and response time in a test might be misleading. To better infer latent ability, a new class of state space models, conjointly modeling response time with time series of dichotomous responses, is put forward. The proposed models can entertain longitudinal observations at individually-varying and irregularly-spaced time points and can accommodate changes in ability and other complications, such as local dependence and randomized item difficulty. The simulation of our models illustrates that by jointly modeling response time with item responses for a series of tests, the precision and reduction of bias for the estimates of individual latent ability can be largely improved. In applying the models to a large collection of reading test data from the MetaMetrics company, we further investigated two competitive relationship in modeling response times with the distance of ability and item difficulty (i.e., monotone or inverted U-shape relationship). The empirical results of model comparison support that an inverted U-shape relationship is more suitable to exemplify student's behaviors and psychology in the exam.

H6-4 D*M estimation: Convolving choice RT data and decision model distributions to factor out non-decision time

Stijn Verdonck, *KU Leuven, Belgium*

Choice RT experiments are an invaluable tool in the fields of psychology and neuroscience. A common assumption is that the total choice response time is the sum of a decision and a non-decision part (time spent on perceptual and motor processes). While the decision part is typically modeled very carefully (diffusion models, neural networks), a simple and adhoc distribution (mostly uniform) is assumed for the non-decision component. Nevertheless, it has been shown that the misspecification of the non-decision time can severely distort the decision model parameter estimates. In this talk, I discuss a new estimation method for choice RT models that elegantly bypasses the specification of the non-decision time distribution by means of an unconventional convolution of data and decision model distributions. Once the decision model parameters are estimated, a non-parametric estimate of the non-decision time distribution can be computed. This step is optional and does not computationally burden the decision model estimation procedure. The method is tested on simulated data and is shown to systematically remove traditional estimation bias related to misspecified non-decision time, even for a relatively small number of observations. For all of the simulations, the shape of the non-decision time distribution that was added could be recovered. Next, the method is applied to a selection of existing diffusion model application papers. For one of them, the new estimates are in direct conflict with the outstanding conclusions of that paper, showing the necessity of our approach.

H6-5 The general linear ballistic accumulator model

Ingmar Visser, *University of Amsterdam, The Netherlands*

Rens Poesse, *University of Amsterdam, The Netherlands*

The Linear Ballistic Accumulator (LBA) model (Brown & Heathcote, 2008) has proven successful in modeling response times from experimental data. This paper presents an extension of this model in which the LBA parameters can be modeled with linear effects to accommodate explanatory variables. An R-package has been developed to fit these models using maximum likelihood estimation. The usefulness of this model is illustrated with various examples. We also present results from a parameter recovery study indicating some potential problems with parameter bias and the generation of starting values for model parameters.

H1 Ordinal & Mixed Responses in IRT

H1-1 Comparison of exposure controls, item pool characteristics, and stopping rules for CAT using the generalized partial credit model

Jiao Can, *Shenzhen University, China*
Huang Yuena, *Shenzhen University, China*
Yan Ding, *Shenzhen University, China*

Three variables were manipulated: four item pool characteristic, three exposure control procedures, and eight stopping rules. A number of measurement precision indexes such as bias, root mean squared error, exposure rates, item usage, and the number of item selected were compute to assess the impact of different item pools and stopping rules on the accuracy of ability estimation and the performance of exposure control procedures for CAT. The results showed that: (1) The standard error stopping rules were found to optimally protect test security while not significantly degrading measurement precision. (2) In fixed length CATs, the MI exposure control procedure was more efficient than the other procedures, while the items were worse evenly selected, and the item usage rate is lower. In variable length CATs, the MI exposure control procedures performed worse. (3) In fixed length CATs, the PR-SE exposure control procedure was more efficient than the other procedures. As far as the single PR-SE is concerned, under the conditions that item step parameters and discrimination follow uniform distributions, and the fixed item number is forty, the PR-SE procedure performed best. (4) In variable length CATs, the RA exposure control procedure was more efficient than the other procedures. As far as the single RA is concerned, when the standard error stopping rule was 0.25 and 0.2, it performed best. (5) The constructs of the item parameters had effect on the results of CAT item exposure control procedures.

H1-2 An investigative study of LR and DFIT methods using GGUM model: An analysis of differential item function assessment of polytomous items

Ebru Demircioglu, *Çankırı Karatekin University, Turkey*
Hulya Kelecioglu, *Hacettepe University, Turkey*

Differential Item Functioning (DIF) is one of the most important methods to identify item bias, which threatens the validity of measurement. There are numerous popular methods to determine DIF in order to find the item functions, however some of these methods do not satisfy the equations used for parameter estimation prior to the determination of the function. A few studies have been conducted on this topic and different DIF detection methods have inconsistent parameter estimations compared to the dominant methods. This study will examine power and Type I error of DIF detection methods with Generalized Graded Unfolding Model (GGUM), a new method that provides system of equations for parameter estimation. Two DIF determination methods will be used in this article: Likelihood Ratio (LR) and Differential Functioning of Items and Tests (DFIT). This paper will also examine methods mentioned above with a simulation study, which will cover a variety of conditions: DIF type (Null DIF, Uniform DIF and Nonuniform DIF), number of categories (3, 4, 5 and 7), percentage of DIF items (0, 20 and 40), impact of DIF (no impact, and N (0.50, 1)), scale length (10, 20 and 30), and sample size per comparison group (250, 500 and 1000). In addition, for each possible combination of conditions, the iterative Item Characteristic Curve (ICC), iterative Test Characteristic Curve (TCC), and iterative Operating Characteristic Curve (OCC) will be included.

H1-3 Incorporating different response formats of competence tests in an IRT model

Claus H. Carstensen, *University of Bamberg, Germany*
Kerstin Haberkorn, *University of Bamberg, Germany*
Steffi Pohl, *Free University of Berlin, Germany*

Competence tests within large-scale assessments usually contain various task formats to adequately measure the participants' knowledge, skills, and abilities. In this paper, partial credit models mixing two response formats frequently used, namely simple Multiple Choice (MC) items and Complex Multiple Choice (CMC) items with several subtasks, are investigated. For evaluating a scaling model for competence tests assuming one paramer models with mixed item formats two relevant issues have to be addressed: how many dimensions are necessary to model the different response formats appropriately and how shall the different response formats be weighted? Based on data of the National Educational Panel Study (NEPS), we analyzed science competence tests embedding MC and CMC items in order to examine the dimensionality of the competence tests and in order to explore which item format weighting best models the response functions. Further, we cross-validate the results with an ICT literacy test of the NEPS and a scientific literacy test of the Programme for International Students Assessment (PISA). Results suggest that the different response formats form a unidimensional structure independent of domain and study. Moreover, item fit indices gave evidence that weighting each subtask of CMC items with half points rather than one point as in an MC item yields an

acceptable model fit. Implications of the findings for the development of scaling models in other competence assessments using different response formats are discussed.

H1-4 Evaluating dimensionality in mixed-format tests using a multidimensional IRT approach

Weiwei Cui, *College Board, USA*
Amy Hendrickson, *College Board, USA*

In recent years, mixed-format tests have been widely used in many large scale testing programs. A mixed-format test contains both multiple choice and free-response items within one single test form to combine the strengths of different item formats. However, a mixture of different item formats may result in a more complex dimensional structure compared to single format tests. For example, different item formats may measure different types of skills even within one single content area and therefore introduce an extra unexpected dimension in addition to the intended test construct. Many psychometric methods for calibration, scaling and equating, especially item response theory-based methods, rely on knowledge of or assumptions about the dimensionality. Unexpected dimensions potentially pose challenges to the evaluation of the dimensionality, as well as to the selection of psychometric methods for calibration, scaling and equating. This study evaluates the dimensionality in a large scale mixed-format test using a multidimensional IRT approach. The number of dimensions and the dimensional structure is assessed by applying a between-item multidimensional IRT model and a bi-factor model to empirical data.

H1-5 Comparison of unidimensional IRT calibration and equating methods for mixed-format tests

Yujin Kang, *The University of Iowa, USA*
Won-Chan Lee, *The University of Iowa, USA*

This study intends to compare the robustness of Unidimensional Item Response Theory (UIRT) calibration and equating methods, when item format effects are ignored for mixed-format tests. In this study, a Common Item Non-Equivalent Groups (CINEG) design is considered for equating, and three UIRT calibration methods and two equating methods are studied. Specifically, separate calibration with linking, concurrent calibration, and fixed item parameter calibration are used as calibration methods, and Item Response Theory (IRT) true score equating and IRT observed score equating are used as equating methods. This study consists of two parts: an intact test forms and groups study, and a pseudo test forms and groups study. In the intact test forms and groups study, three real mixed-format tests are used, which are selected based on the disattenuated correlations between Multiple-Choice (MC) and Free-Response (FR) scores. The factors of investigation are calibration methods, equating methods, and item format effects. In the pseudo test forms and groups study, the factors of investigation are the proportions of common items, common item compositions, and group differences, in addition to the factors of the intact forms and groups study. It is expected that this study could answer the question about the possibility of using UIRT calibration and equating methods when there are item format effects in mixed-format tests, and it will contribute to the understanding of the relationship between UIRT calibration methods using flexMIRT (Cai, 2013). Moreover, this study would give the information about the interaction between UIRT calibration and equating methods.

H5 Classification & Latent Distributions

H5-1 Fast approximate mixed membership inference with rank data

Elena A. Erosheva, *University of Washington, USA*
Y. Samuel Wang, *University of Washington, USA*

We consider the analysis of multivariate ranking data. For example, surveyed individuals may be asked to rank priorities of the society in addressing various problems, given respective lists of potential actions for each problem. We assume that multivariate rankings by each individual follow a Plackett-Luce mixture model over K sub-populations, specified by the individual membership proportions in these sub-populations. In addition, we assume that only a small number of rankings can be distinguished for each variable. Using variational inference, we estimate the subpopulation parameters as well as mixed membership proportions for each individual, and identify optimal K and the number of distinguishable rankings that fits the data best. Variational approaches allow for faster computation and better scalability to large data sets than do Markov Chain Monte Carlo methods. We illustrate our method on the analysis of latent dimensions of public's priorities on addressing alcohol, drug, and AIDS problems from the Eurobarometer survey. The analysis is carried out using the mixedMem package in R. The package employs variational inference for mixed membership analysis of multivariate discrete data and allows for combinations of binary, mutinomial, or rank variables.

H5-2 A critical review of taxometrics: Why taxometrics is unsuited for psychological data

Robert Hillen, *Tilburg University, The Netherlands*

Whether psychological attributes can be best represented by dimensions or categories has been a longstanding debate in psychology and is referred to as the classification problem (Acton & Zodda, 2005; Kendell, 1975; Meehl, 1995). To solve the classification problem, Meehl introduced the statistical framework known as taxometrics (Meehl, 1965, 1968, 1973). Taxometrics has increased in popularity in the last two decades, particularly in the fields of psychopathology and personality psychology. The performance of taxometrics in detecting latent categories has been frequently studied under highly idealized data conditions (Ruscio et al., 2010; Walters et al., 2010; Walters & Ruscio, 2009, 2010). Our goal was to study how well taxometrics performed under data with measurement properties typical of clinical scales. We investigated whether the taxometric curves for the MAXCOV and MAMBAC procedures were biased for this type of data. Furthermore, we added a simulation study to explore the degree to which the Comparison Curve Fit Index (CCFI, Ruscio et al., 2007), a popular taxometric fit statistic, can distinguish categories from dimensions in data that are characteristic of scales in clinical psychology. Our results indicated that the taxometric curves often did not indicate a categorical latent structure or the base rate was underestimated. The simulation study indicated that the CCFI was an inaccurate method for detecting latent categories and that it is biased to a dimensional inference for typical clinical data. We discuss the properties of the measurement model and the properties of the population model associated with this underperformance and bias.

H5-3 Taxon separation using the bootstrap

Satoshi Watanabe, *Akita Prefectural University, Japan*

The purpose of this study was to provide a new way for using taxometric analysis. Taxometric analysis is a statistical method which has the ability to discriminate between a taxon (a group which has a certain characteristic) and a complement (a group which doesn't belong to the taxon), based on some variables. When a taxon and a complement can be separated, we call the situation 'taxonic', and the non-taxonic situation is called 'dimensional'. The performance of taxometric analysis is ideal when the taxon and the complement can clearly be separated. However, taxometric analysis gives little information when the taxon and the complement are not clearly separated. Taxometric analysis usually shows dimensional situations in most cases and taxonic situations in very few cases. In this case, we find the variable dimensional and discard the information about the taxon. We therefore provide ways for using the information about the taxon. When taxometric analysis shows a taxonic situation in a few cases, we consider each of taxon and complement to be samples from the population and we resample the bootstrap samples from each of taxon and complement. We calculate some statistics of the bootstrap taxon samples and the bootstrap complement samples and compare the differences between them. We call this method 'Taxon Separation Using the Bootstrap'. We tested the validity of the method on the data from the survey conducted on the vulnerability of special fraud.

H5-4 A method of estimating test score distributions by item sampling

Young Koung Kim, *College Board, USA*
Tim Moses, *College Board, USA*

Empirical investigations showing that test score distributions can be accurately estimated from a small sample of test items (Lord, 1961) support important psychometric processes used with large-scale assessments, including justification for estimating proficiency distributions from samples of test exercises (Mislevy, Beaton, Kaplan & Sheehan, 1992, p. 135), and using common items to link the scores of tests after administering these tests to nonequivalent groups (Kolen & Brennan, 2004, pp. 19, 22). This study compares methods for estimating a test score distribution from item samples and assumptions about these item samples. The hypergeometric model originally used in Lord's study (1961) is considered as well as the more complex version based on the four-parameter compound binomial model. Another modeling approach proposed in this study is an application of log-linear modeling based on discriminant information (Haberman, 1984). These log-linear models can be used to produce test score distributions that reflect the moments of test score distributions expected from item sampling assumptions (Lord, 1958), and do so more directly than the hypergeometric and compound binomial models. Data from recently administered large-scale assessments are used to illustrate the use of the hypergeometric, compound binomial and log-linear models for estimating test score distributions from smaller sets of items. The models will be compared across a range of item and examinee sample sizes. The representativeness of scores on the sets of items to the test will also be varied. The accuracies of test score distributions estimated using the three models of interest will be summarized.

H5-5 Standard errors for norm statistics

Hannah Oosterhuis, *Tilburg University, The Netherlands*

Norm statistics allow for the interpretation of scores on psychological and educational tests, because they relate the test score of an individual test taker to the test scores of individuals with, for example, the same gender, age, or education level. Norm statistics are estimated from a sample and therefore are subject to sampling fluctuation. Although this variability across samples should be reported, for example, by providing the standard errors, this is usually not done when norm statistics are presented by test constructors. There are two reasons why these standard errors are rarely reported: (1) For many norm statistics, the standard errors are unknown or hard to derive. (2) For other norm statistics, standard errors are known only if certain unrealistic assumptions about the data are made. In the current paper, we first explain the general framework used for deriving standard errors of norm statistics. The only assumption about the data that we made to obtain the standard errors is an underlying multinomial distribution. The multinomial distribution is less restrictive than the normal distribution and without loss of generality it can be assumed that test scores are multinomially distributed. Second, using this general framework, we derive the standard errors for the mean, standard deviation, variance, standardized test scores, percentiles, and stanines. Third, in a small simulation study we investigate the bias of the derived standard errors. Fourth, we present SPSS syntax to obtain the standard errors of norm statistics. Finally, we provide a brief discussion of the results.

Invited Speakers: Francis Tuerlinckx, Elizabeth Stuart, Xiangdong Yang

YH-1 Bridging psychology and psychometrics: Three case studies

Francis Tuerlinckx, *KU Leuven, Belgium*

In this talk, I examine through three case studies how mathematical and statistical models from psychometrics (and related fields) may help to solve substantive research problems in psychology. The discussed case studies range from semiparametric models to unravel the structure of affect, to dynamical systems models for understanding aspects of psychopathology, to stochastic process models shedding light on elementary decision making. For each of these case studies, I explain the main findings, the underlying models that were used and the challenges and problems that remain to be solved.

Y3-1 Propensity score methods in the context of covariate measurement error

Elizabeth A. Stuart, *Johns Hopkins University, USA*

Propensity score methods are commonly used to estimate causal effects in non-experimental studies. Existing propensity score methods assume that covariates are measured without error but covariate measurement error is likely common. This talk will discuss the implications of measurement error in the covariates on the estimation of causal effects using propensity score methods and investigates Multiple Imputation using External Calibration (MIEC) to account for covariate measurement error in propensity score estimation. MIEC uses a main study sample and a calibration dataset that includes observations of the true covariate (X) as well as the version measured with error (W). MIEC creates multiple imputations of X in the main study sample, using information on the joint distribution of X, W, other covariates, and the outcome of interest, from both the calibration and the main data. In simulation studies we found that MIEC estimates the treatment effect almost as well as if the true covariate X were available. We also found that the outcome must be used in the imputation process, a finding related to the idea of congeniality in the multiple imputation literature. We illustrate MIEC using an example estimating the effect of neighborhood disadvantage on the mental health of adolescents, where the method accounts for measurement error in the adolescents' report of their mothers' age when they (the adolescents) were born.

H6-1 Toward an explanatory scale of cognitively designed algebra story problems

Xiangdong Yang, *East China Normal University, Shanghai, China*

Cognitive item design entails a theory-driven approach to measurement (Bouwmeester & Sijtsma, 2004; Embretson, 1998, 2000; Lohman & Ippel, 1993). A theory of the to-be-measured construct is more than a simple definition or some abstract principles of item design. It formulates an elaborated description of the substantive nature of the construct, its structural representation in the mechanism of item solution for a given type of tasks, as well as operational models that link cognitive variables to task features and examinee responses. In addition, this approach sheds new light on fundamental issues of measurement, such as construct validity, and provides substantive bases for automatic item design, explanatory measurement scale construction, and cognitive diagnosis. In this talk, I will present results from several studies about principled task analyses and cognitive feature extraction, based upon a cognitive theory of algebra story problem solving, and show that how such results can lead to the construction of an explanatory measurement scale spanning from 2nd to 8th grade and possible schema of generating algebra story problems in a principled fashion. Implications of such results on diagnostic modeling will also be discussed.

Poster Session and Welcome Reception

See Abstract Book: Posters.

Tuesday, July 14

Parallel Sessions, Tuesday AM

YH Psychometric Issues in Large Scale Assessment in the International Context [Symposium] Organizers: Hongyun Liu & Xiaohong Gae; Chair: Tao Xin & Changhua Rich

YH-1 Setting standards toward the quality of Chinese compulsory education

Tao Xin, *Beijing Normal University, China*

The National Assessment of Educational Quality (NAEQ) aims to assess how well students in compulsory education have mastered knowledge and skills, and to investigate relevant factors that might influence students' academic performance. The domains of Chinese language, mathematical, scientific, arts and physical literacies are covered in terms of mastery of the national curriculum criterion. Questionnaires were developed to assess students' mental and physical wellbeing, the family-, teacher-, and school-related factors. The NAEQ was built in 2007. It has been a national policy-oriented assessment program after a seven-year pilot. In 2014, more than 200,000 students enrolled at the 4th and 8th grade from 31 provinces were measured in the subjects of mathematics and physical health. More than 20,000 school teachers and principals were also measured in terms of school environment, teaching behaviors, and other relevant factors. The findings not only provided a clear picture for the public about student academic progress, but also assist us in understanding what kind factors might influence student performance.

YH-2 Different growth measures on different vertical scales

Dongmei Li, *ACT, USA*

Vertical scales have been used by testing programs for decades to facilitate the tracking of student performance over time. With the recent emphasis on the measuring of student growth for accountability purposes in educational policy, scores from vertically scaled tests have been used to evaluate school or teacher performance. Though there are many statistical models that can be used to measure growth without a vertical scale, having vertically scaled scores make it possible to use statistics such as the simple gain score to measure growth in an absolute sense on the vertical scale. However, since vertical scales are constructed by linking scores of tests from different grade levels through various data collection designs and statistical methods, it is important to understand that any changes in the processes of vertical scaling could have led to a somewhat different vertical scale. That is, with scores from the same vertical scale, there are different measures of student growth and behind each operational vertical scale, there is a range of plausible vertical scales. Extending on previous research on the statistical relationships between different growth measures on the same scale and the relationships between the same growth measure on different vertical scales, this study will further investigate how the relationships between different growth measures would change on different plausible vertical scales, both theoretically and empirically, using real data and/or simulated data based on the a vertical scaling data collection. Results from these investigations will inform practical decisions based on growth results from vertical scales.

YH-3 A Chinese longitudinal study: Growth modeling with value-added interpretations

Xin Li, *ACT, USA*
Hongyun Liu, *Beijing Normal University, China*

One of the purposes of education is to foster learning, and to achieve changes in achievement (Willett, 1994), so investigating change in individual achievement over time is of central importance in educational research (e.g., Seltzer et al., 2003). Studying individual change not only allows researchers to measure and document individuals' progress, but also to monitor and evaluate educational systems, such as school performance and effectiveness of educational interventions over time. For decades, researchers have been using and improving growth modeling techniques for effectively and accurately measuring individual changes over time (e.g., Bryk & Raudenbush, 1987; Muthén & Khoo, 1998). The current study adopts several growth models, including the projection model and student growth percentile model, to a longitudinal dataset that captures five measurements for thousands of Chinese high school students from 40 schools. Results of the analyses are used to answer the questions of how individual students perform relative to peers with similar score histories, whether there are any differences in growth for different groups (i.e., girls and boys), and so on. With value-added interpretations, results can also be used within systems for teacher and school evaluation. In addition, considering the fact that students are nested within a large number of classrooms (i.e., more than 300), and the availability of characteristics of individuals' backgrounds, this study can be extended to include modeling intra-individual and inter-individual differences in growth using multilevel latent variable growth modeling.

YH-4 An investigation of the impact of extracurricular learning on student science achievement in China

Danhui Zhang, *Beijing Normal University, China*
Jiao Can, *Shenzhen University, China*
Xiuna Wang, *Beijing Normal University, China*
Yang Cui, *Beijing Normal University, China*

Learning activities outside of school have been regarded as taking more and more important roles in enhancing students' development in different aspects. This study is aimed to better understand questions related to the impact of extracurricular science learning behaviors on 8th grade student science achievements. A nationwide large scale data set collected in 2012 in China, including achievements in biology, physics, and earth and space science, along with student and teacher questionnaire, were all employed in the current study. Hierarchical linear modeling was conducted to examine the relationship between the extent to which that student engaged into extracurricular science learning activities and their academic performance in science. Students' different demographic characteristics, as well as the background information describing both teachers and schools were all controlled in the model. Further analysis was also conducted to explore whether school location, family social economic status, student gender, and teacher characteristics will influence student engagement in extracurricular learning activities. Particularly, we intend to explore the existence of disparities between urban and rural schools, high and low SES students, male and female student with respect to access to the extracurricular activity, which might be considered as the indicator of "opportunity gap." Suggestions for education policy decision making and implications for further studies were also discussed based on the findings.

YH-5 Generalizability of assessing school accountability under matrix sampling

Xiaohong Gao, *ACT, USA*

Over the past decade, policy makers, educators, parents, and students around the world have been overwhelmed with assessing educational accountability (e.g., NAEP and PISA). Matrix sampling is often used in school level assessment for its broad sampling of content domains and items. The emerging of the new Common Core State Standards has created another wave of testing. For school accountability assessment, not only are test items, raters, and occasions sources of measurement error, students can also contribute to measurement uncertainty when decisions about schools are generalized over all students and across multiple years (Cronbach, Linn, Brennan, & Haertel, 1997). Although matrix sampling had been used in school level assessment, very few studies have examined score generalizability under matrix sampling. Most of the studies have either used single form data or aggregated scores over students and items when evaluating school score generalizability. Generalizability theory provides a powerful conceptual framework in analyzing measurement precision under complex designs such as matrix sampling. The purpose of this study is to contribute to the knowledge base of school score reliability by conceptualizing multiple sources of measurement error associated with school scores and examining generalizability of a large scale assessment under matrix sampling. Since a single generalizability analysis usually cannot disentangle and estimate all potential sources of error, multiple univariate and multivariate generalizability analyses are conducted. Generalizability of composite school scores over multiple forms as well as scores of single forms are evaluated. The study further explores how to improve measurement precision of aggregated scores.

Y3 Mixture/Latent Class Models

Y3-1 An application of a random mixture nominal item response model in a latent transition analysis for investigating instruction effects

Hye-Jeong Choi, *University of Georgia, USA*
Allan S. Cohen, *University of Georgia, USA*
Brian A. Bottge, *University of Kentucky, USA*

The purpose of this study will be to apply a random mixture nominal item response model in a latent transition analysis for investigating instruction effects. The host study design was a pretest-posttest, school-based cluster randomized trial. A random mixture nominal item response model will be used to identify students' error patterns in mathematics at the pretest and the posttest. Instruction effects will be investigated in terms of students' transitioning in error patterns. That is, error patterns will be compared between students in the Enhanced Anchored Instruction (EAI) condition with students in a Business-As-Usual (BAU) instructional condition in inclusive settings. We will also compare error patterns in

mathematics of students with math disabilities and students without math disabilities following an instructional intervention.

Y3-2 Program evaluation with multilevel and multivariate longitudinal outcomes

Depeng Jiang, *University of Manitoba, Canada*
Rob Santos, *Healthy Child Manitoba Office, Canada*
Teresa Mayer, *Healthy Child Manitoba Office, Canada*
Leanne Boyd, *Healthy Child Manitoba Office, Canada*

Many intervention programs, such as the PAX Good Behavior Game (PAX) program implemented in Manitoba province-wide schools, often have multiple outcome variables (e.g., Emotional Symptoms, Conduct Problems, Hyperactivity/Inattention, Peer Relationship Problems, and Prosocial Behavior). These variables were also reported for multiple time points (e.g., pre and post intervention) where data for participants are organized at more than one level (students nested in schools). In this paper, we propose to use the latent variable framework to evaluate the intervention program with multilevel and multiple longitudinal outcome variables. Data from the Manitoba PAX Study serve as an illustration. First, Latent Transition Analysis (LTA) at the student level helps us to examine how students at different levels of antisocial behaviors or lack of pro-social behaviors transition from pre- to post-intervention and whether the intervention program affect this transition probability. Then, Latent Class Analysis (LCA) at the school level helps us examine how school profiles of antisocial behaviors or lack of prosocial behaviors change from pre- to post-intervention. The strengths and limitations of this approach are discussed.

Y3-3 Impact of misspecified turning points on heterogeneous growth trait estimation

Wen Luo, *Texas A&M University*
Ling Ning, *Texas A&M University, United States*
Fuhui Tong, *Texas A&M University*

Piecewise growth mixture modeling (PGMM) can be used to investigate growth and change of subpopulations consisting of distinct developmental phases (Muthén, 2008). The major difficulty in specifying a PGMM is how to optimally locate a turning point (or transition point, or knot). A brief literature review showed that empirical studies using PGMMs relied exclusively on theoretical considerations to specify a priori a turning point and the turning point was often set at the time of intervention. However such considerations may not be always reasonable, because the turning point may occur after the intervention due to delay in response to intervention. We conducted a comprehensive Monte Carlo simulation study to explore the impact of misspecification of turning points on growth trait parameter estimation and classification accuracy under various data scenarios. Based on previous findings regarding growth mixture modeling, five design factors were considered, including (a) the degree of trajectory separation, (b) the number of repeated measures, (c) mixing percentages of latent classes, (d) sample size, (e) degree of misspecification of the turning point. The preliminary results showed that the degree of how growth trait parameter estimation and classification accuracy are impacted depends on the severity of misspecification of turning points and the impact gets particularly worse when the degree of trajectory separation is low, keeping the other design factors constant. The study will draw applied researchers' attention to the importance of a correct specification of turning points in modeling heterogeneous growth traits.

Wen Luo (was Jing Ning)

Y3-4 Mixture higher-order diagnostic classification model with covariates

Yoon Soo Park, *University of Illinois at Chicago, USA*
Young-Sun Lee, *Columbia University, USA*
Kuan Xing, *University of Illinois at Chicago, USA*

Diagnostic Classification Models (DCMs) classify examinees into attribute mastery profiles and have useful instruction and assessment implications for students and educators. Among various DCMs developed, the higher-order deterministic, inputs, noisy, "and" gate model (HO-DINA; de la Torre & Douglas, 2004) provides diagnostic information about the test taker's attribute mastery. An attractive aspect of the HO-DINA model is the ability to estimate attribute difficulty and attribute discrimination parameters by specifying a higher-order latent trait in addition to the DINA item parameters, guessing and slip. However, if attribute characteristics vary for some latent subgroups of test takers, such information can provide meaningful context for educators. This study examines an extension of the HO-DINA by incorporating a finite mixture distribution that estimates different attribute-level parameters for latent subgroups of the data, by proposing the Mixture HO-DINA (MHO-DINA). This study also presents methods to parameterize covariates into the MHO-DINA model that serves to explain differences in latent subgroup classification. This study is divided into two studies that examine real-world data and simulations that investigate the stability of the MHO-DINA and extensions which incorporate covariates. Real-world analysis using Trends in International Mathematics and Science Study (TIMSS) mathematics data (25 items and 7 attributes) showed that the MHO-DINA model fits better than the HO-DINA. Furthermore, including covariates, such as access to educational resources or science scores, helped to explain differences in attribute characteristics. Simulation studies examining the recovery of parameters for varying sample sizes and number of attributes showed stability in parameter estimates and latent classes.

Y3-5 Person-mixture item-response models for Likert scale data

Jesper Tijmstra, *Tilburg University, The Netherlands*
Minjeong Jeon, *The Ohio State University, USA*
Maria Bolsinova, *Utrecht University, The Netherlands*

Psychological constructs such as personality and attitudes are frequently measured with Likert scale items, which often include a middle response category. In applied settings, the middle response category is often coded numerically and in ascending order (e.g., strongly disagree, disagree, neutral, agree, and strongly agree). This assumes that a decision to select the middle category should be seen as indicative of the personality trait or attitude that is measured. However, respondents could also have chosen the middle category based on other and qualitatively different response processes, such as response tendencies. If such a response is taken by the model to be indicative of the personality trait or attitude that the scale is meant to measure, this will result in biased parameter estimates and incorrect standard errors. To deal with the different possible processes that lead persons to select the middle category, we propose a person-mixture item response model, in which persons either treat the middle category as being qualitatively similar to the other categories (the standard model), or treat it as qualitatively different (e.g. due to response tendencies), for which an IRTree model is used that includes a second latent variable explaining the choice for the middle category. The mixture model is estimated using a Gibbs sampler, and the posterior probability of persons belonging to the models is estimated. The application and performance of the procedure will be illustrated based on empirical and simulation studies to show that the proposed approach works in practice.

H6 Handling Missing Data in Psychometrics [Invited Symposium] Organizer & Chair: Anders Skrondal

H6-1 Treating item-level missing data in SEMs when indicators are parcels

Victoria Savalei, *University of British Columbia, Canada*
Mijke T. Rhemtulla, *University of Amsterdam, The Netherlands*

In many modeling contexts, the variables in the model are linear composites of the raw items measured for each participant: for instance, regression and path analysis models rely on scale scores, and Structural Equation Models (SEMs) often use parcels as indicators of latent constructs. If missing data occur at the item level, currently no analytic method exists to handle such missing data with minimal loss of efficiency. Item-level multiple imputation is the only approach available that does not lose any information, assuming the number of imputations is large. In this article, we develop an analytic approach for handling item-level missing data that is a variant of the two-stage methodology (Savalei & Bentler, 2009; Yuan & Bentler, 2000; Yuan & Lu, 2008), and that is the analytic equivalent of item-level multiple imputation. A large simulation study compares the new two-stage approach to the currently available methods for handling item-level missing data: the scale-level FIML and the "available-case" FIML. A comparison to item-level MI is also conducted in select conditions. We find that the two-stage approach performs best, and we recommend its implementation in popular software, its further study and use by practitioners.

H6-2 Modelling missing values in cross-national surveys: A latent variable approach

Irini Moustaki, *London School of Economics, UK*
Jouni Kuha, *London School of Economics, UK*
Myrsini Katsikatsou, *London School of Economics, UK*

In survey research, the aim is often to measure some underlying trait(s) of the respondents through their responses to a set of questions. In the paper, we focus on cross-national surveys, where the main research objective is to compare the distribution of the latent variables across countries. We focus on the modelling of item nonresponse in such surveys, and studying its effects on cross-national comparisons. We consider models which are extensions of standard multigroup latent variable models, extended in such as a way as to model the missing data mechanism together with the latent constructs and their measurement. The model for the missing data mechanism will serve two purposes: first to characterize the item non-response as ignorable or non-ignorable and consequently to study the patterns of missingness and characteristics of non-respondents across countries, but also to study the effect that a misspecified model for the missing data mechanism might have on the substantively interesting parts of the model, including the cross-national comparisons. Results will be presented from the European Social Survey.

H6-3 Missing data in IRT models

Cees Glas, *University of Twente, The Netherlands*

Missing data in IRT models both present problems and possibilities. On one hand, IRT models provide flexible item administration designs where many variables are missing. On the other hand, missing data present problems for statistical analyses, especially when data are not missing at random, that is, when the ignorability principle of Rubin does not hold. However, IRT models can be enhanced to overcome the problems arising from violation of ignorability. A number of examples of possibilities, problems and solutions is given. Handling sparse designs is illustrated with an IRT choice model used in health research. Patients were asked to give their preferences regarding potential therapies which have varying attributes. Confronting patients with all possible choice combinations would lead to an enormous number of choices, but an item administration design with linear restrictions on the parameters solves the problem. The second topic is an overview of model-based procedures to handle nonignorable missing data due to item nonresponse. Here, an IRT model for the observed data is enhanced with an IRT model for the propensity of missing data. The choice of the latter model is discussed and it is shown how the ensemble of the two models can be concurrently estimated. The final topic is an amalgamation of the two prior topics: an IRT choice model with a stochastic design driven by patients' interests in the attributes varied in the choice experiment, i.e., the choice items. Model estimation and model fit are outlined and real data examples are presented.

H6-4 A simulation study on the performance of multiple imputation using regression with optimal scaling

Joost van Ginkel, *Leiden University, The Netherlands*
Anita J. van der Kooij, *Leiden University, The Netherlands*
Mariëlle Linting, *Leiden University, The Netherlands*

Multiple imputation has become a widely accepted method for handling missing data. Many of the available multiple-imputation procedures make distributional assumptions about the data and assumptions about linear relationships among the numerical variables. Van Ginkel, Van der Kooij, & Linting (2012) proposed a multiple-imputation method that makes no distributional assumptions and no assumptions of linear relations among variables, based on regression using optimal scaling (Gifi, 1990). In this presentation a simulation study is carried out in which the method by Van Ginkel et al. (2012) is compared with already existing multiple-imputation methods. Different violations of linear relations are studied. Results show that multiple imputation using regression with Optimal scaling is better at recovering nonlinear relations among variables than the currently existing multiple-imputation methods.

H6-5 Simple tests of Missing At Random (MAR) in multilevel models

Anders Skrondal, *Norwegian Institute of Public Health, Norway*

It is well known that statistical models can be estimated consistently by maximum likelihood if data are Missing At Random (MAR). However, it is conventional wisdom that the MAR assumption generally cannot be tested. An exception is in generalizations of standard multilevel models, such as shared-parameter models for data Not Missing At Random (NMAR), where a specific type of violation of MAR is captured by special parameters. Unfortunately, consistent estimation of these models hinges on vital but unverifiable assumptions about the missingness mechanism. Instead, we propose tests of MAR that are straightforward to implement in many longitudinal datasets by standard modeling. The approach can detect any violation of MAR and does not invoke more assumptions than multilevel modeling without missing data.

H1 Model Selection & Model Averaging

H1-1 Regression modelling with I-priors

Wicher Bergsma, *London School of Economics, UK*

As is well-known, the maximum likelihood method overfits regression models when the dimension of the model is large relative to the sample size. To address this problem, a number of approaches have been used, such as dimension reduction (e.g., multiple regression selection methods, or the lasso method), subjective priors (which we interpret broadly to include random effects models or

Gaussian process regression), or regularization. In addition to the model assumptions, these three approaches introduce, by their nature, further assumptions for the purpose of estimating the model. The first main contribution of this talk is an alternative method which, like maximum likelihood, requires no assumptions other than those pertaining to the model of interest. Our proposal is based on a new information theoretic Gaussian proper prior for the regression function based on the Fisher information. We call it the I-prior, the 'I' referring to information. The method is no more difficult to implement than random effects models or Gaussian process regression models. Our second main contribution is a modelling methodology made possible by the I-prior, which is applicable to classification, multilevel modelling, functional data analysis, and longitudinal data analysis. For a number of data sets that have previously been analyzed in the literature, we show that our methodology performs competitively with existing methods.

H1-2 Bayesian model averaging over directed acyclic graphs with implications for prediction in structural equation modeling

David Kaplan, *University of Wisconsin - Madison, USA*
Chansoon Lee, *University of Wisconsin - Madison, USA*

This paper examines Bayesian model averaging as a means of improving the predictive performance of structural equation models. Structural equation modeling from a Bayesian perspective addresses the problem of parameter uncertainty through the specification of prior distributions on all model parameters. In addition to parameter uncertainty, it is recognized that there is uncertainty in the choice of models themselves insofar as a particular model is chosen based on prior knowledge of the problem at hand. One approach to addressing the problem of model uncertainty lies in the method of Bayesian model averaging. We expand the framework of Madigan and his colleagues as well as Pearl by considering a structural equation model as a special case of a directed acyclic graph. We then provide an algorithm that searches the model space for sub-models that satisfy the conditions of Occam's razor and Occam's window and obtains a weighted average of the sub-models using posterior model probabilities as weights. Our simulation studies indicate that the model-averaged sub-models provided better posterior predictive performance compared to the estimation of the initially specified model as measured by the log-scoring rule.

H1-3 Model selection in random item mixture IRT model

Meereem Kim, *University of Georgia, USA*
Youn-Jeng Choi, *Massachusetts Institute of Technology, USA*
Hye-Jeong Choi, *University of Georgia, USA*
Allan S. Cohen, *University of Georgia, USA*

The purpose of this study is to examine the utility of model selection methods for Random Item Mixture IRT Models (RI-MixIRTMs). Even though items are generally considered as fixed effects in Item Response Theory (IRT), it is more theoretically convincing to take into account the randomness of items in a model (De Boeck, 2008). Moreover, it is possible to include a covariate(s) for an item in a RI-MixIRTM, which is helpful to assign examinees into different latent groups (Wang, 2011). The benefits of IRT models can only be fully realized if the correct model is chosen (Hambleton, Swaminathan, & Rogers, 1991). The present study is designed to compare model selection results for RI-MixIRTMs. Although some research has been done on model selection methods for a MixIRTM (e.g., Kim, et al., 2014; Li, et al., 2009), research on model selection methods for RI-MixIRTMs is somewhat lacking. Kim et al. (2014) examined model selection methods for the random item mixture Rasch model, but only AIC and BIC were considered for model selection. This study extends Kim et al. (2014) to include the random item mixture 2- and 3-parameter models and PsBF, DIC, and PPMC as model selection criteria. Based on results from Kim et al. (2014), it is expected that BIC will work well in all conditions for all RI-MixIRTMs, all test lengths, sample sizes, and numbers of latent groups, whereas AIC is expected to work less well under the conditions simulated.

H1-4 Bayesian analysis of dichotomous latent trait models

Huiping Wu, *University of Macau, Macau*
Shing-On Leung, *University of Macau, Macau*

Latent Trait Models (LTMs) for binary data analysis need to address two issues: model fit and model selection. A Bayesian model diagnostic tool, Relative Entropy - Posterior Predictive Model Checking (RE-PPMC) with limited information statistics, is suggested to assess the goodness-of-fit for LTMs. While more factors provide better fit, a Bayesian model selection method is introduced for determining the optimal number of factors. To approximate the high-dimensional integral required in the parameter space, we propose the Laplace method with polar coordinate transformation, which also solves the problem of rational indeterminacy. Simulated data sets with different item numbers, degrees of sparseness, sample sizes, and factor dimensions are studied to investigate the performance of the proposed procedure. In addition, two real data sets are analyzed. One is Social Life Feelings (SLF) data, previously studied by many scholars, and the new methods identify two clear-cut factors. The other example is a mathematics test from China national matriculation (known as Gaokao). The model determination provides sufficient grounds for unidimensionality, general skill, and accomplishment in mathematics.

H1-5 Testing autocorrelations: Comparing the standard asymptotic method and resampling techniques

Zijun Ke, *Sun Yat-sen University, China*
Zhiyong (Johnny) Zhang, *University of Notre Dame, USA*

Autocorrelation, which provides a mathematical tool to understand repeating patterns in series data, is often used to facilitate the identification of model orders of time series models, e.g., Moving Average (MA) models and AutoRegressive Moving Average (ARMA) models. The standard asymptotic method of testing autocorrelation assumes normality and may fail in finite samples. Resampling techniques such as surrogate data, simple bootstrap, and moving block bootstrap are competitive alternatives. In this study, we used Monte Carlo simulations to compare the standard asymptotic method with all the aforementioned resampling techniques. In addition, for resampling techniques, we considered both the percentile method and the bias-corrected and accelerated method for confidence interval construction. To evaluate the performance of various methods, two types of type I error rates and power were used: individual (results were calculated for autocorrelation at each lag) and overall (results were calculated for autocorrelations at lags with nonzero population values as a whole). Simulation results showed that surrogate data and simple bootstrap yielded better performance than the other methods, especially in small samples and nonnormal data. The asymptotic method is preferable if sample size is large given its simplicity and popularity.

H5 Reliability

H5-1 Reliable and precise measurement: Caveats and a research agenda

Klaas Sijtsma, *Tilburg University, The Netherlands*

No matter which measurement model one prefers, the eternal issues for each measurement instrument remain: validity and reliability. Validity is highly complicated and almost elusive to a degree that measurement specialists find it easier to concentrate on the uses of the instrument rather than the question what it measures. Reliability is a more technical topic and thus more tangible but the psychometric literature contains many confusing “factoids" that need correction and unsolved problems that need a solution. My presentation focuses on the three main approaches to reliability, viz. classical test theory, factor analysis, and generalizability theory, and discusses agreements and differences, both theoretically and practically. I also provide a list of misconceptions about reliability. Finally, I present a brief discussion of issues that need further exploration. Examples are less biased reliability estimates using latent class analysis, standard error versus bias in reliability estimation, and influence of reliability on power of statistical hypothesis testing.

H5-2 Factors causing reliability underestimation of coefficient alpha

Pieter Ruben Oosterwijk, *Tilburg University, The Netherlands*
L. Andries van der Ark, *University of Amsterdam, The Netherlands*
Klaas Sijtsma, *Tilburg University, The Netherlands*

The attention that coefficient alpha (α) has received in the psychometric literature over the last two decades has, for a large part, been focused on disadvantages of alpha. Alternatives to alpha that have been shown to be superior, where accompanied with examples that are favorable to the alternative reliability estimation method (e.g., negative inter-item covariance for Guttman’s Lambda 2). A systematic analysis of factors influencing the performance of alpha under conditions resembling typical psychological research was performed. Alpha is defined in a way that it is easiest to see how changes in the covariance matrix are having an effect on alpha. Factors include item variance size, inter-item relations and the number of items. This study shows under what conditions there is still a place for alpha in psychological research and when factors like negative inter-item relations and large heterogeneity in item variance causes alpha to break down and alternatives should be used.

H5-3 Estimation methods for single-item reliability

Eva A. O. Zijlmans, *Tilburg University, The Netherlands*
L. Andries van der Ark, *University of Amsterdam, The Netherlands*
Jesper Tijmstra, *Tilburg University, The Netherlands*
Klaas Sijtsma, *Tilburg University, The Netherlands*

Test constructors and test users are interested primarily in the total score on a test rather than scores on individual items. Consequently, reliability is estimated for the total score but not for single item scores. However, item-score reliability might also be of interest, especially to select items appropriate in test construction (Meijer, Sijtsma & Molenaar, 1995). In this study, different methods for estimating item-score reliability are compared with respect to their statistical properties and their practical usefulness. The first method considered is adapted from the Latent Class Reliability Coefficient (LCRC; Van der Ark, Van der Palm & Sijtsma, 2011) for item-score reliability, and uses the latent class model to estimate item-score reliability. The second method also uses latent class analysis: class and response probabilities can be obtained from the estimated latent class model and then used to obtain an item-reliability estimate related to classification consistency. The third method is based on the nonparametric Mokken approach to item response theory (Mokken, 1971), and was first proposed by Meijer et al. (1995). The bias and the standard error of the three item-score reliability estimation methods were investigated and compared by means of a simulation study.

H5-4 Evaluation of five reliability estimates under multidimensional test scenarios

Terry Ackerman, *University of North Carolina at Greensboro, USA*

Reliability is a ubiquitous concern in educational assessment and psychological research. Since Spearman's (1904) definition of reliability, many reliability indices have been developed. In a *Psychometrika* article, Sijtsma (2009) examined several measures of reliability and found that the estimate most often used by practitioners, Cronbach's α, may be very misleading and is not a good choice for practitioners to use. He noted that α was does not change while dimensionality increases, whereas other greatest lower bound estimates increased dramatically. This study expands on Sijtsma's work. This paper investigates the estimation bias of five reliability indexes: Cronbach's (1951) α, Guttman's (1945) λ_2 and λ_4, Bentler's (1972) greatest lower bound, *glb*, and McDonald's (1970) ω. Specifically this study examines how bias in the five lower bound estimates of reliability varies under different simulated two-dimensional conditions. The fully-crossed research design examined four factors: (1) Correlation between the ability dimensions, $r = .0$, 05, 0.8; (2) dispersion of ability distribution, $\sigma_1 = \sigma_2 = 1$ and $\sigma_1 = \sigma_2 = 2$; (3) multidimensional item discrimination level: low, middle, and high; and, 4) number of items: 20 and 40. Generating item parameters came from a real standardized assessment. The generation program provided the true reliability and the frequency distribution of observed scores and true scores. The five reliability estimates were then computed using "psych" package in R (Revelle, 2014).

H5-5 A structural equation modeling approach to estimation of reliability for a test with items having different numbers of ordered categories

Seohyun Kim, *The University of Georgia, USA*
Zhenqiu (Laura) Lu, *The University of Georgia, USA*
Allan S. Cohen, *The University of Georgia, USA*

The Classical Test Theory (CTT) definition of reliability is often described as the proportion of the variance due to true score variance over the total score variance. A common and widely used approach to estimating this reliability is coefficient alpha (Cronbach, 1951). This approach may not be useful, however, when the factor structure of a test is complex. Coefficient alpha assumes, for example, that a test is unidimensional. When a test has a complex, multidimensional structure, the above assumption will not be satisfied and the reliability estimate will be misleading. A Structural Equation Modeling (SEM) approach to the estimation of reliability has been proposed to overcome some of this difficulty (e.g., Raykov & Shrout, 2002). The SEM approach has two subcategories: one type is for continuous item responses and the other for categorical responses. For categorical responses, previous research has focused on responses with the same number of categories. In reality, however, it is the case that the number of categories will vary across items in a test. In this paper, we focus on the SEM approach to the reliability estimation for a tests consisting of varying numbers of ordered categories. An empirical example will be used to illustrate the differences in reliability using coefficient alpha, a linear SEM reliability estimate, and a non-linear SEM reliability estimate. Simulation studies motivated by this example will be designed to examine why differences in reliability estimates occur for the different coefficients.

XH Computerized Adaptive Testing

XH-1 A quick item selection method in computerized adaptive testing for ranking items

Chia-Wen Chen, *The Hong Kong Institute of Education, Hong Kong*
Wen-Chung Wang, *The Hong Kong Institute of Education, Hong Kong*

Ranking items have been widely used in non-cognitive tests such as personality tests and career interest tests. The Rasch model for ipsative tests with multidimensional pairwise comparison items was recently developed and their corresponding CAT algorithms were investigated (Chen & Wang, 2013, 2014). Moreover, this model has been extended to account for ranking items in which more than two statements are to be ranked (Qiu & Wang, 2015). To facilitate this new model for ranking items in a CAT environment, we investigated how a ranking item is selected in a reasonably short time. As this model tends to be high-dimensional and there are a huge number of possible ranking items in an item bank, standard item selection methods that are based on Fisher item information matrix become infeasible in real time. We proposed a quick-and-dirty method for item selection and evaluate its performance with simulations. The result showed this method could select a ranking item in a short time without sacrificing too much efficiency. In the future, we will further implement control procedures for item exposure.

XH-2 Investigation of constraint-weighted item selection procedures in polytomous CAT

Ya-Hui Su, *National Chung Cheng University, Taiwan*

Since examinees are given different sets of items from a large item bank, Computerized Adaptive Testing (CAT) not only enables efficient and precise ability estimation but also increases security of testing materials. The construction of assessments usually involves fulfilling a large number of non-statistical constraints, such as content balancing, key balancing, item exposure control, etc. To improve measurement precision, test security, and test validity, the maximum priority index approach can be used to monitor many constraints simultaneously and efficiently in CAT. Many CAT studies were investigated for dichotomously scored items. However, only few CAT studies were investigated for polytomously scored items. It was found that polytomous CAT needed fewer items than dichotomous CAT did because polytomous items are more informative. Many issues in polytomous CAT still need further attention. Therefore, the purpose of the study is to investigate constraint-weighted item selection procedures in polytomous CAT.

XH-3 Nearly optimal computerized adaptive testing for mastery test

Xiaoou Li, *Columbia University, USA*

Computerized adaptive tests are tests whose items are tailored to examinee's individual ability level. By selecting test items appropriately according to the examinee's previous responses, the accuracy of the inference could be improved and the test length could be shortened. In this paper, we investigate properties of optimal adaptive sequential designs for mastery tests that have minimal Bayes risk. This paper combines several major techniques in statistics and applies them to design optimal mastery test. In particular, we employ techniques in computerized adaptive testing, sequential probability ratio test, stochastic control and empirical Bayes.

XH-4 The application of restrictive stochastic item selection methods in multidimensional computerized adaptive testing

Xiuzhen Mao, *Sichuan Normal University, China*
Yating Wang, *Sichuan Normal University, China*
Zhihui Fu, *Shenyang Normal University, China*

Multidimensional Computerized Adaptive Testing (MCAT) has enjoyed tremendous growth in recent years. Most researchers are interested in the item selection method and are mainly focused on the following problems: (1) improving the accuracy of estimates for both domain capability and composite score, (2) controlling item exposure by the utilization of the stratification, Sympson-Hette, Stocking-Lewis, and maximum prior index methods; and (3) dealing with content constraints by adopting the shadow test or maximum prior index methods. The present study pays attention to item exposure control in MCAT and carries out two simulation experiments. First, the item exposure rate and item pool utilization distributions have been checked and compared by their item selection according to the D-optimality rule, posterior expected KL information, mutual information and continuous entropy. Then, two kinds of restrictive stochastic methods, which were popular in traditional CAT and Cognitive Diagnostic CAT (CD-CAT), have been implemented into the process of item selection. Their performance in terms of estimate accuracy and item exposure were further compared with that of the stratification method. The results showed that (1) the stratification method can improve the exposure rates of most underexposed items but cannot guarantee that all the item exposure rates will be below the allowed maximum exposure rate; (2) the restrictive methods can improve the evenness of item exposure a great deal. In summary, the restrictive methods produced higher exposure rates than the stratification method with similar accuracy of capability estimates.

XH-5 Sequential design for computerized adaptive testing that allows for response revision

Shiyu Wang, *University of Illinois at Urbana-Champaign, USA*
Georgios Fellouris, *University of Illinois at Urbana-Champaign, USA*
Hua-Hua Chang, *University of Illinois at Urbana-Champaign, USA*

In Computerized Adaptive Testing (CAT), items (questions) are selected in real time based on the observed responses, so that the ability of the examinee can be estimated as accurately as possible. This is typically formulated as a non-linear, sequential, experimental design problem with binary observations that correspond to the true or false responses. However, most items in practice are multiple-choice and dichotomous models do not make full use of the available data. Moreover, CAT has been criticized for not allowing test-takers to review and revise their answers. In this work, we propose a CAT design that is based on the nominal response model and in which test-takers are allowed to revise their responses at any time during the test. We show that as the number of administered items goes to infinity, the proposed estimator is (i) strongly consistent for any item selection method and revision strategy and (ii) asymptotically normal when the items are selected to maximize the Fisher information at the current ability estimate and the number of revisions is smaller than the number of items. Such asymptotic results were verified by an idealized simulation study. Furthermore, we also conducted simulation studies based on a realistic item pool to show the robustness of our proposed design to several deceptive test-taking strategies.

X3 Choice/Comparative Response Data

X3-1 A Bayesian multinomial probit model for the analysis of choice data

Duncan K. H. Fong, *Pennsylvania State University, USA*

A new Bayesian multinomial probit model is proposed for the analysis of panel choice data. Using a parameter expansion technique, we are able to devise a Markov Chain Monte Carlo algorithm to compute our Bayesian estimates efficiently. We also show that the proposed procedure enables the estimation of individual level coefficients for the single-period multinomial probit model even when the available prior information is vague. We apply our new procedure to consumer purchase data and reanalyze a well known scanner panel dataset that reveals new substantive insights.

X3-2 Item parameters in forced-choice personality assessments: Does context in which items appear matter?

Yin Lin, *University of Kent, UK*
Anna Brown, *University of Kent, UK*

Self-report questionnaires are the only way to measure people in many applications. The forced-choice response format, where respondents rank statements within blocks, leads to fairer assessments through eliminating uniform response biases and enhancing resistance to impression management. A recently developed Thurstonian IRT model (Brown & Maydeu-Olivares, 2011) allowed proper scaling of forced-choice data, and enabled applications such as computerized adaptive testing to enhance measurement efficiency. However, when assembling tests in an adaptive fashion, one necessarily assumes that item parameters stay constant regardless of the place the item takes in the test. In a forced-choice adaptive test, item A can be placed in a block together with different items depending on the current latent trait estimates. For this to be possible, item parameters have to be invariant with respect to the block design. The question is whether it is realistic to assume that varying context and potential item interactions have negligible impact on item parameters. This empirical study examines the impact of context on item parameters in forced-choice designs. Historical data of the Occupational Personality Questionnaire is used. The first sample (N=22,610) completed a triplet version of the forced-choice assessment. The second sample (N=62,639) completed a quad version, with an additional item per block. The samples were calibrated independently and the IRT parameters were equated before comparison. Results show strong linear relationships between the recovered parameters, with slopes/intercepts correlating across formats at 0.88/0.96 respectively. Conceptual analysis of identified DIF items will be reported and summarised, and implications for adaptive forced-choice assessments discussed.

X3-3 An investigation of enhancement of ability estimation using a nested logit model for multiple-choice items

Tour Liu, *Beijing Normal University, China*

An item response model called the Nested Logit Model (NLM) for multiple-choice data was used in this research. A simulation study and an empirical study were designed to investigate the enhancement of ability estimation by using the NLM in multiple-choice tests. Both simulation study results and empirical study results indicated that the NLM could enhance the accuracy and the stability of person parameter estimation, because more information from distractors was added in. But the accuracy of person parameter estimation showed little differences in 4-choice items, 5-choice items and 6-choice items. Moreover, the NLM could extract more information from low-level respondents than from high-level ones, because they had more distractor chosen behaviors. Furthermore, respondents at different trait levels would be attracted by different distractors in an empirical study of a Chinese Vocabulary Test for Grade 1 by using the changing traces of distractor probabilities caculated from the NLM. It is suggested that the responses of students at different levels might reflect the students' vocabulary development process.

X3-4 Optimal assembly of forced-choice questionnaires

Safir Yousfi, *German Federal Employment Agency, Germany*
Anna Brown, *University of Kent, UK*
Frauke Liers, *Friedrich-Alexander Universität Erlangen-Nürnberg, Germany*

Forced-choice questionnaires have been shown to improve personality assessment by reducing response biases common to Likert scales. Recent progress in IRT modeling has overcome the limitations of ipsative scoring, to date the most prevalent method of analyzing forced-choice data. Recommendations for assembling reliable forced-choice questionnaires from an item bank with known item parameters are available (Brown & Maydeu-Olivares, 2011). Nevertheless, designing a forced-choice questionnaire remains a challenge for test developers as numerous options for combining items in blocks have to be evaluated with respect to many different criteria. It is shown that modelling forced-choice test assembly as a mathematical optimization problem offers a way to address all these issues within a common framework. The main objective of force choice test assembly is to maximize multidimensional test information in order to recover accurately the absolute standing on psychological traits of interest. Other considerations except measurement precision are: (1) increasing cognitive complexity as the size of blocks increases; (2) matching items within blocks on social desirability; and (3) balanced content coverage. These can be taken to account as constraints in the optimization process. We demonstrate a working optimization procedure implemented in R; and illustrate by comparing two forced-choice tests measuring Big Five factors of personality assembled from the same items - an "optima" test, and a manually assembled one. The proposed approach allows not only automating the assembly of forced-choice questionnaires, but also provides an invaluable tool for research, which allows us to derive new recommendations for test assembly by hand.

X3-5 Comparing pairwise comparison and rank-order judgments scaling of teacher competences

Gökhan Kumlu, *Hacettepe University, Turkey*
Sinan Yavuz, *Hacettepe University, Turkey*
Nuri Doğan, *Hacettepe University, Turkey*
Gülfem Dilek Yurttaş, *Gazi University, Turkey*

Reaching the goals of educational programs is dependent on teacher's qualifications because teachers are the most effective players on educational system. Six fields have been identified in general qualifications. One of them is the teaching-learning process qualification, which has seven sub-qualifications; 1) planning the course, 2) preparing material, 3) organizing learning environments, 4) organizing extracurricular activities, 5) diversifying the program according to individual differences, 6) time management, 7) attitude management. In this study, pairwise comparison, rank-order judgments scaling methods, and scaling results consistency have been compared. The aim of the study is to compare two different types of scaling methods and investigate the consistency between methods. 335 pre-service teachers from 3 different universities were used. A teaching-learning process qualifications evaluation form has been used as a data collecting instrument. On the preparation of this form, seven sub-qualifications of teaching-learning process qualifications have been organized for pair-wise comparison and rank-order judgments scaling. Collected data have been investigated by using the rank-order judgments model and Thurstone's Case V model. To determine the consistency of scales, mean error, and chi-square statistics have been calculated. Attitude management has the same rank on two different scaling methods. Other sub-qualifications ranks are close, except diversifying the program according to individual differences sub-qualification. According to these results, types of stimuli on two different scales have no effect on high level rank of teaching-learning process qualifications. There are very few scaling comparison studies in the literature (O'Neil & Chissom, 1993; Albayrak & Gelbal, 2012).

X7 Standard Setting & Automated Scoring

X7-1 Standard setting: From educational assessment to health outcomes measurement (and back?)

David Thissen, *The University of North Carolina at Chapel Hill, USA*

Standard setting methods that associate ranges of tests scores with verbal descriptions have become ubiquitous in educational assessment. Because scale scores based on Item Response Theory (IRT) are an integral part of most commonly used standard setting methods, standard setting has become one of the prominent uses of IRT. Recently, standard setting methods have come to be used with health outcomes measures as well, first to establish score ranges associated with labels describing symptom severity ("mild", "moderate", and "severe"), and then to estimate the "Minimally Important Difference" (MID) on the score scale. This presentation reviews standard setting methods in health outcomes measurement, with emphasis on the development of the "scale-judgment method" to estimate the MID. The scale-judgment method combines ideas derived from "body of work" methods for educational standard setting, IRT, and classical psychometric models for judgment and choice to estimate the value of the minimally important difference between scores. That value, in turn, is used as an aid in the interpretation of the score scale for health outcomes measures.

X7-2 Quality control of standard setting using many-faceted and diagnostic models

Myoung Hwa Kim, *Korea Institute for Curriculum and Evaluation, South Korea*
Yoonsun Jang, *University of Georgia, USA*

The major feature of standard setting is that it is based on experts' subjective judgments. That is, experts have to conceptualize the "minimally competent examinees." It might be hard to coherently retain this concept, not only between experts but even for a single expert. In order to validate the cut-scores, we need to show that the standard setting method is appropriately designed and successfully implemented by the standard setting panelists. Therefore, the goals of this study are as follows. First, this study evaluates the procedures of standard setting according to six sets of criteria (i.e., panelist severity or leniency, the halo effect, response sets, restriction of range, interaction effects, and differential facet functioning) suggested by Engelhard (2013). In a standard setting context, the MFR model can incorporate multiple facets. In our study, three facets (i.e., panelists, items, and standard setting round) are included. Furthermore, the results of standard setting are evaluated through a comparison of the results of the classification of students. For controlling the quality of results of standard setting, the proportion of mastery for each domain from using the Diagnostic Classification Model (DCM) is compared to the proportion of mastery based on the cut-score for each domain. Among several types of DCMs, the Log-linear Cognitive Diagnostic Model (LCDM), provided by Henson, Templin, & Willse (2009), is used because the LCDM is a more flexible modeling framework.

X7-3 Adaptive body of work: A new standard setting procedure in complex exams

Xin Luo, *Michigan State University, USA*
Mark D. Reckase, *Michigan State University, USA*

Standard setting can be defined as a procedure for determining a cut-score that examinees must exceed to demonstrate that they have met an academic requirement. The accuracy of the results of a standard setting process influences the pass/fail decision directly and plays a vital role in educational fairness. Traditional standard setting procedures focus more on dichotomous items. The Body of Work (BoW) method is proposed to set cut-scores in constructed-response tests. Although the BoW is an appealing standard setting procedure due to its technical and practical adequacy for complex exams mainly consisting of constructed-response items, little research has been conducted to investigate its accuracy and consistency. In addition, the preparation of the work folders that are required by the method is tedious and time-consuming. Also, previous research indicated that the pin-pointing stage in BoW is unlikely to improve the cut-score estimate. Therefore, how to improve the efficiency of the BoW is worthy of investigation. This study adopts an adaptive strategy for the BoW to reduce the number of required work folders and improve the efficiency of standard setting. Specifically, three research objectives are addressed: (1) the idea of active learning in machine learning is introduced to facilitate a mathematical proof and a simulation study on why pin-pointing does not work; (2) adaptive BoW is compared with traditional BoW and the former is expected to achieve the same precision as the latter with fewer folders; and (3) factors influencing the results are investigated. Variables investigated include the number and distribution of folders in the range-finding stage, the folder selection method in adaptive pin-pointing, and the stopping rule under the adaptive strategy.

X7-4 Improving the effectiveness of the pre-screening filtering system of an automated essay scoring engine

Jing Chen, *Educational Testing Service, USA*
Mo Zhang, *Educational Testing Service, USA*

Automated scoring engines are being used to evaluate the writing quality of essays. The advisory flags of a scoring engine are used as a pre-screening filtering system to identify essays that are not scorable by the engine, either because the engine is likely to issue an invalid score or because the essay is designed to unduly inflate the score by taking advantage of aspects of automated scoring (i.e., "gaming-the-engine"). In this study, we examine the effectiveness of the advisory flags of a scoring engine that is widely used to score a variety of writing assessments and investigate ways to enhance the capability of the advisory flags to detect unscorable essays. Ten types of advisory flags are being used in the scoring engine. Each advisory flag is triggered by some cutoff values set on a particular feature score of the essay. Some advisory flags are classified as fatal flags that force the engine to exclude the essay from automated scoring, while the others are classified as non-fatal flags that only provide warnings. We investigate how to tailor the flag settings (e.g., adjust the cutoff values that trigger each flag; determine which advisory flag is fatal) for different writing tasks to optimize the performance of the advisory flags. In addition, we develop new flagging tools based on the unusualness of the essay feature scores to detect essays that are unsuitable for automated scoring. A strong pre-screening filtering system will support the validity of automated essay scoring.

Dissertation Prize Speaker: Michelle LaMar

YH-1 Cognitive models for understanding student thinking in complex tasks

Michelle LaMar, *Educational Testing Service, USA*

Complex tasks require complex cognition for which straight-forward unidimensional models are unlikely to be a good approximation. As current educational reform emphasizes integrated performance tasks, and modern technology offers rich streams of performance data, the use of more complex cognitive models in psychometrics may be increasingly useful and feasible. Recent work will be presented using a cognitive model borrowed from computer science, the Markov Decision Process (MDP), as the basis for a flexible psychometric framework. The MDP measurement model enables inferences about student goals, beliefs, and strategic thinking, either individually or in combination. Issues of parameter estimation and interpretation are explored through simulation, while empirical studies illustrate the validity, practicality and limitations of the approach.

Keynote Speaker: Yutaka Kano

YH-1 Developments in multivariate missing data analysis

Yutaka Kano, *Osaka University, Japan*

It is well known that missing data can cause serious bias in statistical inference if they are ignored. In this talk, we will provide two recent developments in missing data analysis. One is a proposal of an analysis method of data with a huge amount of missing values, say 90 percent of data points. This situation can happen if subjects can choose items to be responded in a questionnaire study. We apply it to make factor analysis of a real data set in an Internet survey research. The second is a theoretical study on effects of inclusion of auxiliary variables to reduce a bias of the MLE without any missing-data mechanism for NMAR missingness. Recently the method for missing data has been reported to be very useful. Some Mote Carlo studies in the literature have shown that inclusion of auxiliary variables can successfully reduce the bias greatly. We first provide a general framework to study the bias of the MLE for NMAR missingness, and then apply it to evaluate effects of auxiliary variables. Surprisingly, there are cases where inclusion of auxiliary variables enlarges the bias.

Parallel Sessions, Tuesday PM1

YH Psychometrics in East Asia [Invited Symposium] Organizer: Mark Wilson; Chair: Wen-Chung Wang

YH-1 Psychometric developments in the Chinese mainland

Tao Xin, *Beijing Normal University, China*

As an important branch of psychology devoted to tests and measurements, psychometrics began in the late 19th century in London through the influences of Galton and his associates. In the next one hundred years, psychometrics has obtained rapid theoretical development and widespread practical application all over the world (particularly in the US and Europe). However, in the Chinese mainland, the area of psychometrics did not really start to develop until the last decade of the 20th century. Since then, some Chinese mainland scholars began to focus on psychometric issues and conduct related studies at four leading psychometric laboratories, the representative experts from which are: Prof. Shuliang Ding at the Jiangxi Normal University; Prof. Tao Xin at Beijing Normal University; Prof. Tao Jian at Northeast

Normal University; and Prof. Xiangdong Yang at East China Normal University. For example, the work of Prof. Ding and his colleagues has centered on the development of item-selection algorithms in Computerized Adaptive Testing (CAT) (including dichotomously-scored CAT and polytomously-scored CAT), improvement of Q-matrix theory in cognitive diagnostic assessment, and discussion of new methods of test equating in the past twenty years. Prof. Xin's group has conducted a series of studies on the exploration of online calibration methods/designs, non-statistical constraints (e.g., item exposure control and content balancing) and automatic estimation of Q-matrix in Cognitive Diagnostic CAT (CD-CAT), development of a new cognitive diagnostic method (i.e., the generalized distance discriminating method), and polytomous extensions of cognitive diagnostic models. In summary, the psychometric researches done by Chinese mainland scholars are still relatively weak, the international academic influence of related work is still very limited, and thus there is still a long way for us to go.

YH-2 Psychometrics in Japan

Kohei Adachi, *Osaka University, Japan*

It is after 1960 that Japanese have published papers in *Psychometrika*, which shows that Japan does not have a long history for psychometrics. However, in 1952, Dr. Hayashi had proposed an optimal scoring procedure for categorical data. Since 1970, a number of books for factor analysis with related multivariate procedures have been published, but in Japanese unfortunately. They may include the original methodology and findings to be distributed internationally. Data analysis procedures, in which Japanese psychometricians have been/are particularly interested, include factor analysis, multidimensional scaling, optimal scoring, structural equation modeling, and item response theory. An approach to be noted is a matrix-intensive one for multivariate data analysis, which was mainly made by Dr. Yanai. One of the recent trends may be that young psychometricians like Bayesian approaches. Two big events in the last two decades were IMPS 2001 and 2007 held at Osaka and Tokyo, respectively. A domestic society related to psychometrics is the Behaviormetric Society of Japan. Psychometrics interests pure statisticians in Japan, which is exemplified by the fact that a number of younger psychometricians have got prizes in the meetings of the Japanese Society of Computational Statistics. The Japan Association for Research on Testing and Japanese Classification Society are also related to psychometrics.

YH-3 Past and present of psychometric work in Korea

Soonmook Lee, *Sungkyunkwan University, Korea*
Ahyoung Kim, *Ewha Womans University, Korea*
Min-Kyeong Kang, *Sungkyunkwan University, Korea*

Psychometric work in Korea can be categorized into four main areas: behavioral assessment, measurement and testing, behavioral statistics, and research methodology. Behavioral assessment involving judgmental evaluation and referral systems has the longest history that can be traced back to the 1st century B.C. in the three dynasties of Korean peninsula. In those days behavioral assessment was performed in the form of evaluating qualifications of candidates for state education systems or for government posts. Such practices are still the focus of attention in current psychometric work of Korea. Measurement and testing began during the Unified Silla Dynasty (788 A.D.) from a strong need to recruit talents from across the country using a national written examination system. Today, measurement and testing in Korea has established its place in academic institutes and many testing companies. The evaluation of soldiers' mental health, employment-related testing, clinical testing and diagnoses, as well as academic entrance exams are the major applications. Although the third and fourth areas, behavioral statistics and research methodology, were introduced to Korea much more recently (1950-1960) they are receiving equally as much attention as the previously mentioned areas.

YH-4 The development of psychometrics and psychological testing in Taiwan

Bor-Chen Kuo, *National Taichung University of Education, Taiwan*

In this talk, the history and recent development of psychometrics and psychological testing in Taiwan will be briefly introduced. In addition, a content analysis of articles published in *Psychological Testing*, which is the major psychometrics and psychological testing journal in Taiwan, during the past six decades, is used to demonstrate the important trends and developments. In conclusion, some future directions will be mentioned.

Y3 Model Fit/Selection in SEM

Y3-1 Small sample statistics for SEM: New corrections to the likelihood ratio statistic for nonnormal data

Ge Jiang, *University of Notre Dame, USA*
Ke-Hai Yuan, *University of Notre Dame, USA*

Test statistics are essential for evaluating overall model fit in Structural Equation Modeling (SEM). The classical likelihood ratio statistic, TML, performs well with normally distributed data at large sample sizes. However, normal data seldom occur in practice. This presentation will introduce three new corrections to the likelihood ratio statistic aiming to yield improved performance with nonnormally distributed and/or small sample sized data. A Monte Carlo study is conducted to compare the performance of TCOR1, the rank-corrected statistic, TCOR2, the average-constant statistic, and TCOR3, the average-p-value statistic, against existing test statistics in the Confirmatory Factor Analysis (CFA) model. The results show that all of the corrections outperform existing test statistics for severely nonnormal data at small sample sizes. Under different conditions, these corrections give optimal performance respectively and are recommended accordingly.

Y3-2 The problem with having two watches: Assessment of fit when RMSEA and CFI disagree

Keke Lai, *University of California, Merced, USA*

The Root Mean Square Error of Approximation (RMSEA) and the Comparative Fit Index (CFI) are two widely applied indices to assess fit of structural equation models. Because these two indices are viewed positively by researchers, one might presume that their values would yield comparable qualitative assessments of model fit for any data set. When RMSEA and CFI offer different evaluations of model fit, we argue that researchers are likely to be confused and potentially make incorrect research conclusions. We derive the necessary as well as the sufficient conditions for inconsistent interpretations of these indices. We also study inconsistency in results for RMSEA and CFI at the sample level. Rather than indicating any systematic misspecification in the model, the two indices can disagree because (a) they evaluate, by design, the magnitude of the model's fit function value from different perspectives, (b) the cutoff values for these indices are arbitrary, and (c) "close" fit and its relationship with fit indices are not well understood. In the context of inconsistent judgments of fit using RMSEA and CFI, we discuss the implications of using cutoff values to evaluate a model's adequacy and to plan the sample size for a study employing SEM to analyze the data.

Y3-3 Mardia's skewness and kurtosis tests for normality in covariance structure models under high dimension with a small sample size

Jiajuan Liang, *University of New Haven, USA*

Mardia's (1970) multivariate skewness and kurtosis statistics are widely applied to testing normality in validating the normal assumption on statistical models. Covariance structure models are mostly developed under the normal assumption. Yuan & Bentler (1997) proposed the general format of covariance structure models $X = A\zeta$, where the matrix A is a function of a basic vector of parameters, ζ may represent measured, latent or residual random or fixed variables. Testing the normality assumption on the latent variables in ζ can be transferred to testing normality of the observed variable X. When the sample size (say, n) is larger than the dimension of X (say, p), many existing tests for normality in the literature are available. Most existing tests for normality were constructed from the sample covariance matrix requiring $n > p$. Mardia's (1970) skewness and kurtosis are among those tests constructed from the sample covariance matrix. Yuan & Bentler (1997) pointed out that the model format $X = A\zeta$ contains as its special cases for most structural models, include the factor analysis model, errors-in-variable model, and linear structural equation models. In this paper we will employ the idea of principal component and the theory of spherical distributions to develop projected Mardia's skewness and kurtosis. These projection tests are applicable to the cases of both $n > p$ and $n \leq p$. The empirical performance of the projection tests is studied by Monte Carlo simulation and illustrated by a real example.

Y3-4 Assessing the size of model misfit in covariance structure models

Alberto Maydeu-Olivares, *University of Barcelona, Spain*

As in any other statistical procedure, when assessing the fit of a covariance structure model, if the null hypothesis is rejected, it is necessary to report the magnitude of the misfit using a confidence interval. In recent years, it has been advocated to replace this null hypothesis framework by an interval estimation approach. Within this approach, a confidence interval for an effect size of model misfit is directly estimated. How shall we assess the size of model misfit? I advocate using standardized effect sizes that can be readily interpreted. Covariance structure modeling is a multivariate technique. As a result, there is an effect size for each residual covariance. I advocate using the standardized residual covariances to gauge the model size misfit, and the square root of the unweighted sum of the squared standardized residual covariances (SRMSR) to gauge the overall model misfit. Statistical theory is provided to construct confidence intervals for these effect sizes. I show that the sample SRMSR is asymptotically a biased estimate of the population SRMSR and provide an unbiased estimator. I contrast the results obtained with those of the RMSEA. The RMSEA is also an effect size of model misfit. However, because the RMSEA is unstandardized, its interpretation is unfeasible without cut-off values. No cut-off values are needed for the SRMSR as it is a standardized statistic.

Y3-5 Errors and biases in the reporting of the fit of structural equation models

Jelte M. Wicherts, *Tilburg University, The Netherlands*

Like any statistical procedure, the application of Structural Equation Modeling (SEM) is sensitive to errors and biases. Here we discuss the most likely errors and biases in applying and reporting of SEMs, and present the results of a systematic study of how authors report SEM results in the scientific literature. We retrieved from ISI Web of Science 3125 peer-reviewed articles that referred to the manuals of AMOS, LISREL, or Mplus, from which we drew a random sample of 242 articles that reported the fit of a total of 1286 SEMs. Results show that the most common fit measures were exact fit test (90%), followed by RMSEA (80%) and CFI (73%). We present the distributions of the fit measures and show that RMSEAs are more likely to be reported when they are lower than the typical thresholds of .05 and .08. We also checked reported RMSEAs by re-computation. We were unable to replicate reported values of about a quarter of reported RMSEA values (discrepancies>.005), with most reported values being lower than values recomputed on the basis of DFs, Chi-squares, and sample sizes. Downward rounding was particularly common around .05, which is reminiscent of results found in studies of misreporting of p-values of null hypothesis significance testing. I discuss the implications and highlight potential solutions. The latter includes the mandatory reporting of correlation/covariance matrices when applying SEM and syntaxes as part of supplementary files, which would allow for independent verification of the SEMs in published reports.

H6 Modeling Responses and Response Times to Psychological Tests [Symposium] Organizer & Chair: Dylan Molenaar

H6-1 Fast versus slow multiplication strategies

Abe Hofman, *University of Amsterdam, The Netherlands*
Ingmar Visser, *University of Amsterdam, The Netherlands*
Brenda Jansen, *University of Amsterdam, The Netherlands*
Maarten Marsman, *University of Amsterdam, The Netherlands*
Han van der Maas, *University of Amsterdam, The Netherlands*

Strategies for solving multiplication problems differ in speed. Whereas retrieval is fast, backup strategies (e.g., repeated addition) are slow. Furthermore, fast and slow strategies also differ qualitatively. From an item perspective, it is therefore expected that effects, such as the tie-effect, are more or less prominent in slow and fast strategies. From a person perspective, children are expected to differ in execution of these strategies. In the present study we used a tree-based item response-modeling framework1 to investigate whether the proposed qualitative distinctions in fast and slow processes can be detected. Both processes, and a third process that determines whether the response is fast or slow, are modeled with a Rasch model, disentangling item and person effects for both processes. We analyzed responses of 18,000 children on a set of single digit multiplication items, collected with Math Garden (an online computer-adaptive training environment for learning mathematics). Preliminarily results showed qualitative differences between the fast and the slow process. Although the order of the item difficulties was comparable, the tie-effect was larger in the fast than in the slow process. Concerning the person perspective, the fast and slow processes correlated but not perfectly, r = .81, indicating that performance differs between processes. Moreover, girls were more able than boys on both the fast and the slow process, whereas boys were more often fast than girls. The results emphasize the quantitative and qualitative differences between strategies for solving single digit multiplication, and provide possibilities of tailored feedback on learning multiplication.

H6-2 What dependencies between response times and response accuracies can tell

Paul De Boeck, *The Ohio State University, USA*

Based on data from a variety of cognitive tests it seems there is a particular pattern of residual dependencies between response times and response accuracies after controlling for the latent variables from the hierarchical model of van der Linden. The residual dependencies between response time and accuracy are negative up to a turning point on the item response probability scale from where on they become positive. Above the turning point short response times are associated with correctness and below the turning point the association is reversed. These results cannot be explained as speed-accuracy tradeoff effects. They can instead be explained based on the following two assumptions. (1) The competence level of subjects varies across items. Competence is a rotated latent variable that combines ability and speed, in the tradition of Spearman and in line with a scoring rule recently described by Maris and van der Maas. (2) The probability scale for item responses has a natural zero point below which slow tends to be associated with correct. The zero point can be seen as the latent "I don't know" state from Thissen & Steinberg's multiple-choice model. An alternative interpretation is based on qualitative differences between fast and slow processing.

H6-3 Response mixture modeling of responses and response times

Dylan Molenaar, *University of Amsterdam, The Netherlands*
Daniel Oberski, *Tilburg University, The Netherlands*
Jeroen K. Vermunt, *Tilburg University, The Netherlands*
Paul De Boeck, *The Ohio State University, USA*

It can be argued that the most dominant approach to model responses and response times is the so-called collateral information approach (e.g., Van der Linden, 2007; Thissen, 1983). However, within this approach, the speed of the test taker is assumed to be constant across all test items (Van der Linden, 2009). In ability testing, this assumption will likely be violated if the test takers use different strategies that require different amounts of time to apply. In present talk a new method called response mixture modeling is presented to account for the differential use of strategies when solving ability test items. Response mixture modeling differs from traditional mixture modeling in that each item response is classified into one of a fixed number of strategies (latent classes). In traditional mixture modeling, the full vector of responses is classified into classes.

H6-4 Response time models: All for one and one for all

Gunter Maris, *Cito, The Netherlands*

There is a range of process models that account for how and when someone comes to an answer or a choice. These process models are quite distinct. We demonstrate that such models for explaining response time and accuracy in two alternative forced choice tasks as the diffusion model, race models, and urgency gating models are indistinguishable. The set-up we consider involves two runners (or processes) running against each other in a competition. The behaviour of both runners is described by processes $X(t)$ and $Y(t)$ that indicate the length each has run after t time units. We assume that our stochastic processes $X(t)$ and $Y(t)$ are non-decreasing with time. As with any competition, the purpose of the running is to name a winner. The different models for this type of data correspond to different ways to decide who wins the race. A diffusion type of process ends when one person leads with a certain distance, a race model when the first runner reaches a certain distance, and an urgency gating race ends when two runners run towards each other (the one having covered the largest distance winning). Although these different ways to reach a decision seem to be different, we can for any one particular model set up models under the other two schemes that are empirically indistinguishable. Hence, the mere observing of when someone gives a certain response can not inform us about the process which generated this response.

H6-5 Response times in computer adaptive practice and monitoring systems

Han van der Maas, *University of Amsterdam, The Netherlands*

One way to deal with the speed accuracy trade-off in ability testing is to use explicit scoring rules, such that subjects know how speed and accuracy are weighted. An additional advantage is that Response Time (RT) can be used to estimate ability, which is important when relatively easy items are administered in a computer adaptive practice and monitoring system for education. We will present results from one such system, called Math Garden, which is very popular in the Netherlands. Children use this system to practice math, teachers to monitor progress, and researchers to collect data. The setup of this new computer adaptive approach will be explained and the use of scoring rules will be evaluated in relation to different approaches to incorporate RTs in psychometric models.

H1 Clustering & Classification

H1-1 COSA 2.0: Clustering objects on subsets of attributes revisited

Jacqueline J. Meulman, *Leiden University, The Netherlands*
Maarten Kampert, *Leiden University, The Netherlands*
Jerome Friedman, *Stanford University, USA*

The original COSA paper by Friedman and Meulman was published more than ten years ago. Its motivation was given by data from the life sciences, where the number of attributes (variables) is often much larger than the number of objects. In those cases, it is very likely that groups of objects cluster only on a limited number of attributes, where the subsets of important attributes is not the same for all clusters. Since its publication, the COSA paper has both been cited very often, as well as described as very hard to understand. Also, its complicated software implementation did not encourage application of the method. Since there is still a very large demand for clustering in high-dimensional data, and the potential of COSA has been far from exploited, we have started further work on COSA by closely studying (and possibly improving) its properties, expanding its features, and making it widely accessible in a user-friendly implementation in an R package (rCOSA). Because the main output of COSA is a "cluster-happy" dissimilarity matrix, we included a variety of proximity analysis methods to be used side-by-side to discover hard to find clusters in the data. We will also indicate circumstances in which COSA is not the most appropriate clustering method.

H1-2 A general method for clustering in a reduced subspace

Maria Brigida Ferraro, *Sapienza University of Rome, Italy*
Paolo Giordani, *Sapienza University of Rome, Italy*
Maurizio Vichi, *Sapienza University of Rome, Italy*

We propose a new method for the simultaneous reduction of units and variables in a data matrix. Reduced K-Means (RKM) and Factorial K-Means (FKM) are two well-know techniques used in this context. Both techniques involve principal component analysis and K-means but they work in a different way. On the one hand, RKM maximizes the between-clusters deviance without imposing any condition on the within-clusters deviance. On the other hand, FKM minimizes the within-clusters deviance without imposing any condition on the between one. Hence, RKM and FKM give different results: the partition obtained by RKM may contain isolated but heterogeneous clusters while the one obtained by FKM may include homogeneous but not isolated clusters. FKM can be used when RKM fails, and vice versa. For this reason we propose to combine the two techniques in a general model through a linear convex combination. In doing so, we approach the problem in a fuzzy framework. We investigate the adequacy of the proposal by means of simulation and real case studies.

H1-3 In search of an optimally valid criteria set for diagnosing alcohol use disorder

Melanie Wall, *Columbia University, USA*
Cheri Raffo, *Columbia University, USA*
Deborah Hasin, *Columbia University, USA*

In the field of psychiatry, the current classification system is the Diagnostic and Statistical Manual of Mental Disorders, 5th edition (DSM-5). The DSM-5 provides diagnostic criteria that are associated with each mental disorder. Following the prototypal theory, not all listed criteria for a specific disorder must be present in order to receive a diagnosis but rather a specified subset of criteria must be present. To be diagnosed with an Alcohol Use Disorder (AUD) under the DSM-5, an individual meeting any two of the 11 criteria receives a diagnosis. The question we will address is whether the validity of the diagnosis using the rule of 2 out of 11 can be improved upon using some other rule based on the same or a subset of the 11 criteria. In this presentation we will: 1) define validity in the absence of a gold standard by formalizing ideas of content validity from psychometrics 2) compare different measures of diagnostic performance (e.g., sensitivity, specificity, misclassification rate, and diagnostic odds ratio from biostatistics), and 3) demonstrate an empirical search algorithm incorporating uncertainty through bootstrapping that can identify an optimally valid rule for diagnosing alcohol use disorder. The data we will use comes from the National Epidemiologic Survey on Alcohol and Related Conditions (NESARC) conducted by the National Institute on Alcohol Abuse and Alcoholism in 2001/2002.

H1-4 Classification of writing styles using keystroke logs

Mo Zhang, *Educational Testing Service, USA*
Jiangang Hao, *Educational Testing Service, USA*
Chen Li, *Educational Testing Service, USA*
Paul Deane, *Educational Testing Service, USA*

In this study, we applied a hierarchical vectorization method to quantify the Keystroke Logging (KL) information for essays written under a timed testing condition. With the new feature variables, we examined how well they can help to improve the prediction of essay scores. We also performed cluster analysis based on the cosine similarity of the KL information, by which we hope to identify different writing styles. Similar to hierarchical clustering analysis, the hierarchical vectorization method builds a hierarchical tree structure based on the time intervals between the adjacent words (known as Inter Word Interval, IWI) using simple linkage. At each tree level, four characteristic variables are extracted to represent the information at that level: location of the IWI, the value of the IWI, the sum of total variance of time from each of the clusters (i.e., intra-cluster variance), and the variance of time among the clusters (i.e., inter-cluster variance). By appending the feature variables from different levels up to a specified depth level, a vector can be created to represent the keystroke (in particular, IWI) information for a single essay. The data we used are captured by a well-designed KL system that records: (a) action (type of behavior; e.g., deletion), (b) duration (time elapsed or text length that a given action covers; e.g., burst), (c) location (where in the text a given action occurs; e.g., within-word), and (d) time point (when a given action occurs in the writing process; e.g., at the beginning of the composition).

H1-5 The δ-machine

Mark de Rooij, *Leiden University, The Netherlands*

We introduce the δ-machine, a statistical learning tool for classification based on dissimilarities or distances, δ, between inputs. Compared to other statistical learning tools, which are often black boxes, this machine has a clear interpretation in terms of distances towards a prototype or exemplar. We introduce the machine, discuss its properties, derive variable importance measures and partial dependence plots for the machine, and show toy as well as empirical examples. Detailed interpretations of the machinery will be discussed.

H5 IRT for Paired Samples & Longitudinal Data

H5-1 Modeling global and local dependence in paired samples

Kuan-Yu Jin, *The Hong Kong Institute of Education, Hong Kong*

In psychological tests or social surveys, measurements are often made on paired samples. The paired samples are grouped due to certain similarity (e.g., couples, twins), which may cause dependence on their responses. It has been widely agreed that the measured latent traits of paired samples should be correlated, and is known as the Global Person Dependence (GPD). However, after the latent traits are controlled, the item residuals of paired samples may not be correlated, suggesting the presence of Local Person Dependence (LPD). In this study we extend the standard Rasch model to account for the GPD and LPD simultaneously. The parameters of the models can be estimated with the freeware WinBUGS. Two simulation studies were conducted to evaluate the parameter recovery of the new model and the consequences of model misspecification. Results showed that the parameters of the new model was recovered fairly well. Fitting unnecessarily complicated models to data without LPD did little to harm parameter estimation. However, ignoring the LPD by fitting the standard Rasch model yielded biased estimates for the item parameters, correlation between latent traits, and test reliability. Two empirical examples are provided to demonstrate the implications and applications of the new models.

H5-2 A heteroscedastic multilevel item response theory model for twin data

Inga Schwabe, *University of Twente, The Netherlands*
Stéphanie van den Berg, *University of Twente, The Netherlands*

Results of international comparisons such as PISA and TIMSS show that Dutch students are underperforming when compared internationally. While overall performance is not bad, there is only a small percentage of excellent students in the Netherlands when compared internationally. This suggests that, currently, Holland's most gifted students are not nurtured to their maximum potential. Talent might be the result of a particular combination of innate talent and the right environmental conditions in which such talent can flourish; a process also described as *gene-environment interaction*: Depending on their genotype (innate talent), students might respond differently to environmental stimuli such as for example the school environment. In this presentation, we present a heteroscedastic multilevel model that can be used with twin data to test for such an interaction between genetic and environmental influences. We show that ignoring differences in measurement precision (e.g., higher measurement error for high performing students) can lead to spurious effects and how this potential bias can be remedied by incorporating an item response theory model into the analysis. An application of the new approach is illustrated using raw item scores of Dutch twins on a national educational achievement test.

H5-3 Psychometrics in genetics

Stéphanie van den Berg, *University of Twente, The Netherlands*

In genetics, the relationship is studied between observable traits (phenotypes) and the genetic code (genotype). Only 0.1 % of the genomic sequence in our DNA is different across individuals, but this variation is responsible for significant proportions of variance in, for example, intelligence, personality, and educational test scores.

This presentation will discuss a number of psychometric issues that have to be taken into account in genetic inference: heterogeneous measurement error, scaling and scale transformation, measurement variance, and test score equating. Ignoring the psychometrics will lead to spurious gene-environment interaction, spurious gene-gene interaction, biased claims about the relative importance of the family environment and inheritance, and may lead to reduced statistical power in gene-finding studies. A number of psychometric models will be presented specially developed for genetic and genetically-informative data. These can be seen as variations and extensions of multilevel IRT models and generalized linear mixed models.

H5-4 Alternative methods of latent growth modeling for examining intervention effects

Jenn-Yun Tein, *Arizona State University, USA*
Heining Cham, *Fordham University, USA*

Intervention studies are longitudinal in nature. Often researchers collect follow-up data after posttest for evaluation of short-term or long-term program effects. With pretest, posttest, and two or more follow-up assessments of the outcomes, researchers can examine the intervention effects on the developmental trajectories of the outcomes over time using latent growth curve modeling (Muthén & Curran, 1997). Latent Growth Modeling (LGM) estimates whether individual growth trajectories of the outcome variable (i.e., within-person estimations) vary from person to person and whether the inter-individual variation (i.e., between-person estimations) is systematically related to the intervention condition. The conventional LGM approach gives a single average estimate for each of the growth factors (e.g., initial level and slope components) and a single estimation of variance for the growth parameters, and assumes a uniform influence of the intervention on the variance and growth parameters. The conventional LGM is not practical and cannot reflect the possibility of different types of trajectories or different typologies of trajectories across participants. This study investigates three alternative methods of growth curve modeling for testing intervention effects: the multi-group piecewise LGM method (Sandler, Cham, et al., 2015), the latent growth mixture modeling method (Muthén et al., 2002), and the discontinuous functional forms of growth method (Singer & Willett, 2003). The presentation also discusses the pros and cons that are associated with each of the methods. Examples with real data will be provided.

H5-5 A comparison of dynamic factor analysis, principal component analysis, and independent component analysis for an EEG study

Guangjian Zhang, *University of Notre Dame, USA*
Wen Li, *Florida State University, USA*

We compare and contrast three multivariate methods (dynamic factor analysis, principal component analysis, and independent component analysis) for an EEG study in which participants are exposed to three types of emotional stimuli: fear, disgust, and neutral. Factor analysis and principal component analysis are established methods in the social and behavioral sciences. Independent component analysis is a recently developed method, but it is gaining popularity among neuroscientists who are interested in modeling complex physiological data like fMRI and EEG. Implications of the comparison are also discussed.

XH Linking & Data Fusion

XH-1 The comparison of multidimensional equating methods in correlation between dimensions and different group ability in NEAT design

Jiyoung Jung, *Yonsei University, South Korea*
Guemin Lee, *Yonsei University, South Korea*
Dasom Hwang, *Yonsei University, South Korea*

[Title changed 7/7/2015. Abstract not updated] Many researchers have agreed that psychological or educational tests are sensitive to multiple traits, implying the need for Multidimensional Item Response Theory (MIRT). Nevertheless, most IRT linking methods that have been used in practice are based on unidimensional IRT models. To investigate more accurate and efficient linking procedures for equating, previous studies have tried to propose and develop linking procedures under MIRT frameworks. However, few studies of comparison of MIRT linking methods based on the bi-factor model under nonequivalent groups have been developed yet. The purpose of this study is to compare results of accuracy and efficiency from three kinds of linking methods (Stocking-Lord, Mean/Mean, Mean/Sigma) under a bi-factor MIRT. According to previous studies of MIRT linking methods by using simple and complex structure models, it is found that Stocking-Lord is better than others under the Non-Equivalent Groups with Anchor Test (NEAT) design in most of studies (Hirsch, 1989; Li & Lissitz, 2000; Oshima et al., 2000; Thompson et al., 1997). In this study, simulated data will be generated and used to investigate statistical characteristics of bi-factor MIRT linking results with various conditions. This study aims to investigate the effect of MIRT linking methods with regard to different anchor-test conditions.

XH-2 Data integration methods and implementing IRT models for test linking: When multiple versions of a measurement tool have been administered for longitudinal child assessment

Katerina Marcoulides, *Arizona State University, USA*
Kevin Grimm, *Arizona State University, USA*

Meta-analysis, where statistics from published work is analyzed, is a common way of summarizing research of a specific research question. Recently, data fusion has been proposed as an alternative to meta-analysis. The two have the same goal of synthesizing a collection of research with a specific research question, but a major difference between the two is what is analyzed. In data fusion, data from multiple studies is analyzed versus reported statistics used in a meta-analysis. Data from the various studies are combined and analyzed as a single dataset. There are many benefits to using data fusion methods. For example, new data collection is unnecessary. Additionally, by combining multiple data sets you increase the sample size, which increases the statistical power. Combining multiple data sets also creates a more heterogeneous sample, which can increase generalizability. However, there are some drawbacks of the data fusion method. First, you need the actual data from multiple studies. Additionally, the method of measurement needs to be the same because data are being combined. If different tests are used to measure the same construct, then researchers need to find a way to link the tests. This study utilizes the benefits of the data fusion approach by combining data from six different longitudinal studies, and examines the associations between covariates such as EocioEconomic Status (SES), gender, and race, and how they are associated with developmental changes. These developmental changes are measured using the Woodcock-Johnson Psycho-Educational Battery-Revised (WJ-R; Woodcock & Johnson, 1990) and the Woodcock-Johnson Psycho-Educational Battery-III (WJ-III; Woodcock, McGrew & Mather, 2001) were each measured in three of the six studies. Both versions have an Applied Problems (AP) and a Letter-Word Identification (LWID) section. Each version shares some common items; these common items are used to link the two versions of the Woodcock Johnson. An item response model was fit to all the item responses as if they form a single test. Since the common items have the same ten parameters, the uncommon items are scaled appropriately allowing the theta scores to be comparable (allowing a one to one mapping between theta scores and total scores on the two versions of the test). This project has multiple purposes. First, it applies an example of data fusion methods to longitudinal data. Then, it uses item response theory test linking to address the issue with multiple version of a measurement tool. Finally, we want to show that more complex (and more appropriate) models can be fit using the combined data sets that we otherwise would not be able to fit using data from any one of the six data sets.

XH-3 A comparison of calibrated projection and conventional IRT based linking

Eisuke Segawa, *Northwestern University, USA*
Benjamin Schalet, *Northwestern University, USA*
David Cella, *Northwestern University, USA*

Calibrated Projection (CP; Thissen et al., 2011) uses two-dimensional confirmatory item factor analysis loading linked items on one factor (linked latent variable or linked LV) and anchor items on another (anchor LV). One-dimensional IRT based linking (IRT1) is a special case of CP where the correlation between the two LVs is assumed to be 1. Linking in CP and IRT is based on the posterior of anchor LV and common LV (for both linked and anchor items), respectively, given linked items. The posterior in CP is decomposed into two parts: (1) the posterior of the linked LV given linked items and (2) the conditional distribution of the anchor LV given the linked LV (projection of anchor LV on linked LV). Comparison of the decomposed posterior in CP with posterior in IRT1 reveals that they differ because IRT1 (1) uses the common LV instead of anchor LV and (2) does not use the projection. Our simulation study showed that even with a reasonably high correlation, e.g., $\rho = 0.87$, on average the differences between IRT1 and CP in score were positive and 20% of CP and differences in standard error were negative and 39% of CP. Because CP does not impose the unrealistic assumption, it is free from the bias, and its implementation is as easy as IRT1. These findings suggest that relative to IRT1, CP may be more advantageous when the correlation deviates significantly from 1.0.

XH-4 Data fusion of heterogeneous data sets: The Tucker3-PCA model as an illustrative example

Tom F. Wilderjans, *Leiden University, The Netherlands*
Elisa F. Bernal, *Universidad de Salamanca, Spain*
Purificación Galindo-Villardón, *Universidad de Salamanca, Spain*
Eva Ceulemans, *KU Leuven, Belgium*

In different fields of science, challenging research questions often imply that different blocks of (heterogeneous) information pertaining to the same research units are to be analysed simultaneously (i.e., data fusion), with these different blocks possibly being of a different nature and emanating from different sources. Examples can be found in psychology (e.g., simultaneous analysis of behavioural and questionnaire data) but also in other fields, like bio-informatics (e.g., fusion of fluorescence spectroscopy, nuclear magnetic resonance and liquid chromatography-mass spectrometry data). One such challenging question pertains to the disclosure of the joint structure underlying coupled data blocks. To tackle this question, we propose a data fusion strategy that consists of factorizing each data block with an appropriate decomposition method (e.g., PARAFAC, Tucker3, PCA) and imposing the parameters of the common mode, which are estimated by using the information present in all data blocks, to be equal for all data blocks this mode belongs to. In this talk, as an illustration of the proposed data fusion approach, we will present the new Tucker3-PCA model, which is a multiway multiblock component model for the fusion of a three-way real-valued data array and a two-way real-valued data matrix that share a single mode. An alternating least squares algorithm to fit the Tucker3-PCA model to data is described and evaluated in a simulation study. An application of the model to the fusion of behavioural (i.e., situation-specific behaviour profiles) and questionnaire (i.e., person dispositions) data is discussed.

XH-5 A Bayesian “fill in” method for correcting for publication bias in meta-analysis

Han Du, *University of Notre Dame, USA*
Fang Liu, *University of Notre Dame, USA*
Lijuan Wang, *University of Notre Dame, USA*

Publication bias occurs when the strength and/or direction of the results between published and unpublished studies differ. It can threaten the validity of the systematic reviews and summaries of the results of a research topic. Conclusions based on the meta-analysis of published studies without correcting for publication bias are often too optimistic and biased toward significance or positivity. Statistical methods have been developed to adjust/correct publication bias under different missingness assumptions. We propose a novel Bayesian fill in method, BALM, for estimating and adjusting publication bias with considering the p-values of a test statistic from all possible studies (published and unpublished) under certain assumptions. The aim of this talk is to provide researchers with a rigorous and accessible method to correct for publication bias in meta-analysis. We will first discuss several existing Publication Bias Correction Methods (PBCM), including their underlying assumptions, advantages, and limitations. We then present the procedure of our proposed method, BALM. We focus on testing standardized mean differences using a random-effects meta-analysis. This approach can adjust publication bias occurring in two scenarios when non-significant results are suppressed and when results in the unexpected direction are suppressed. Finally, the performance of our proposed approach is investigated via a simulation study with real sample size information, compared with several commonly used PBCMs. Our approach is also applied to a real meta-analysis to correcting for publication bias.

X3 Nonparametric & Nonstandard IRT

X3-1 Multilevel Mokken scale analysis

L. Andries van der Ark, *University of Amsterdam, The Netherlands*
Daniela Crisan, *Tilburg University, The Netherlands*
Marcia S. Andrade, *University of Amsterdam, The Netherlands*
Janneke E. van de Pol, *Utrecht University, The Netherlands*

Snijders (2001) proposed two-level Mokken scale analysis for dichotomous data, but the idea remained largely unnoticed. In this presentation, we discuss two-level Mokken scale analysis for dichotomous data, and present a generalization to polytomous data, the standard errors of the scalability coefficients, and software. Furthermore, we demonstrate the usefulness of two-level Mokken scale analysis by analyzing students’ evaluations of high-school teachers.

X3-2 The sources of variance overlap between personality and behavioral measures

Samantha Bouwmeester, *Erasmus University Rotterdam, The Netherlands*

In the domain of developing personality there is quite some debate about the relationships between personality and behavior. Before the interesting question concerning the direction of the relationships between personality traits and internal and external behavior of the developing child is investigated one should explore the source of the variance overlap. That is, it is not clear whether the observed correlations between personality and behavioral measures are explained by similar items in the two scales or by true construct overlap. The aim of the present study is to explore this question in order to shed light on the source of the variance. Nonparametric IRT will be used to investigate the dimensionality of the personality and behavioral measures and to perform item analyses. The personality and behavioral data of a large group of children stem from a longitudinal study of Prinzie and colleagues.

X3-3 Derivation of logistic model from item response patterns

Yuan Sun, *National Institute of Informatics, Japan*
Naoya Todo, *National Institute of Informatics, Japan*
Yi Sun, *University of Chinese Academy of Sciences, China*

Today, the logistic model is most frequently used item response model in item response theory. Historically, Lord (1952) introduced item response model based on the normal cumulative distribution function (Normal Cumulative Model, NCM). However, because the logistic function could approximate normal cumulative distribution function and it was easier to handle mathematically, Birnbaum (1968) proposed using logistic function as item response model (Logistic Model, LM). Although LM was initially introduced as an alternative of NCM, in this paper, we tried to demonstrate that when we derive the item response model based only on item response patterns, we necessarily obtain the LM. The results showed that we were successful in deriving a person's response probability based only on item response patters (Person Response Probability, PRP) and from the PRP we derivate LM as the item response model. We can regard this LM as a Two Parameter Logistic Model (2PLM), and the item parameters (discrimination parameter and difficulty parameter) in this LM are the same as in the 2PLM. Furthermore, in the course of this demonstration, we introduced a new definition of the subject parameter and then related the average amount of information to the subject parameter. This newly discovered relationship will let us reconsider item response theory from different points of view. In addition to these results, we also tried to develop a new parameter estimation technique based on this LM derivation.

X3-4 Continuation ratio model in item response theory and selection of models for polytomous items

Seock-Ho Kim, *University of Georgia, USA*

In the continuous ratio model under item response theory, continuation ratio logits are used to model the probabilities of obtaining ordered categories in polytomously scored items. The continuation ratio model is an alternative to other models for ordered category items such as the partial credit model, the graded response model, and the generalized partial credit model. The theoretical development of the model, descriptions of special cases, maximum likelihood estimation and Bayesian estimation of the item and ability parameters, and model comparisons with information based criteria are presented. Illustrations and comparisons of the models for ordered category items are presented using empirical data.

X3-5 Is the four-parameter model suited for achievement tests as well as psychological tests?

Yue Liu, *Sichuan Research Institute of Education Sciences, China*
Hongyun Liu, *Beijing Normal University, China*

The four-parameter item response theory model (4PM) was first introduced by Barton and Lord. The 4PM allows each item's upper asymptote to be less than 1 to account for the possibility that even a high ability respondent may on occasion answer an easy question incorrectly. In previous research, this model, representing both upper and lower asymptotes provided good fit for some psychological tests, such as the Minnesota Multiphasic Personality Inventory (MMPI) and so on. However, for achievement tests Barton and Lord found that the 4PM failed to improve the likelihood, or to significantly change any ability estimates, for the datasets collected by ETS (i.e., SAT Math etc.). Therefore, is it really inappropriate to use the 4PM model in achievement tests? In this study, we apply the Rasch model, 2PM, 3PM, 3PM_R (3PM with reversing scores on each item), 4PM, 4PM_c (4PM with equal guessing parameters) and 4PM_d (4PM with equal slipping parameters) to the Taylor Manifest Anxiety Scale and the Standard Maths test for senior one students. Meanwhile, the dataset collected in Maths tests was used to construct two different kinds of distributions: approximately normal distribution and negatively skewed distribution. Based on the AIC and BIC criteria, the 4PM is needed to accurately characterize item response behavior on both psychological tests and achievement tests. Relative to the 3PM, the estimated discrimination parameters are larger and the difficulty parameters are smaller in the 4PM. Moreover, model choice also makes a substantial difference for persons with extreme trait values.

X7 Measurement in Practice

X7-1 Tomorrowland: A computer adaptive measure of delay discounting

David Stillwell, *University of Cambridge, UK*

Michael Palkovics, *University of Vienna, Austria*
Michal Kosinski, *Stanford University, USA*
Vaishali Mahalingam, *University of Cambridge, UK*

Measuring delay discounting with multiple time delays and amounts is time consuming. We develop and validate an efficient and psychometrically sound computer adaptive measure of delay discounting based on a dataset of N = 4,190 participants. Study 1 discusses the development of a binary search-like algorithm to measure delay discounting, and presents the results of a simulation study comparing the newly developed algorithm to item response theory-based computer adaptive testing and a standard measure. Study 2 presents evidence of concurrent validity with a standard measure of delay discounting, and convergent validity with addictive behaviour and the BIS-11 questionnaire measure of impulsivity. The new measure is quicker than standard measures, includes a range of time delays, can be applied to multiple reward magnitudes, and shows excellent concurrent and convergent validity. Delay discounting has been linked to various behavioural, health and social outcomes, including academic achievement, social functioning, substance use; thus, highlighting the importance of the paradigm.

X7-2 An effectiveness comparison of IRT-based DIF detection using a scale purification procedure and a DIF-free-then-DIF strategy

Shenghong Dong, *Jiangxi Normal University, China*
Xiaofen Cai, *Jiangxi Normal University, China*
Chu Tang, *Jiangxi Normal University, China*

The main goal of the study was to examine the power and robustness of nine different DIF assessment methods, “SIB-ST”, “SIB-SP”, “SIB-PA”, “IRT-LR-ST”, “IRT-LR-SP”, “IRT-LR-PA”, “DFIT-ST”, “DFIT-SP”, and “DFIT-PA”, under the graded response model. Four independent variables were manipulated in a Monte Carlo simulation study, including sample sizes, the DIF pattern, the percentage of DIF items, and the DIF seriousness. The outcomes were the type I error and the power of DIF assessment. The main conclusions of this study were: (1) For all nine methods for DIF detection, the power and type I error increased as the sample size increased. The SP and PA procedure effectively helped to control the type I error and to improve power. (2) Regarding DIF seriousness, the nine methods performed better in the moderate seriousness conditions than in two types of light DIF seriousness with uniform or mixed DIF items. (3) For all nine methods, the power and type I error increased as the number of DIF items increased. (4) Methods using scale purification and DFTD performed better with a high percentage of DIF items and large sample size than the standard procedures, but when the number if DIF items and sample size were small, the general procedures perform betterd. (5) Considering both power and type I error, IRT-LR with scale purification and DFTD performed better than the other methods.

X7-3 Multidimensional item response analysis of a number sense assessment with mathematics learning disabilities students

Hye Kyung Lee, *University of California, Berkeley, USA*

An instrument to assess number sense ability was developed for General Education (GE) students and Math Learning Disability (MLD) students in elementary and early middle school. This study presents results from the second pilot study (2012), which included a total of 222 students (102 GE students and 120 MLD students). Multidimensional item response model was used to (a) investigate the empirical evidence supporting the four-dimensional design of the instrument, (b) explore the performances between GE and MLD students, and (c) examine the validity of the learning progression of number sense underlying the instrument. The model and item fit statistics support the use of the multidimensional design of the instrument. Cognitive developmental lags were identified in the MLD students.

X7-4 Development of an American standard-based instrument for assessing principal leadership: Investigating the psychometric properties of the Chinese principals’ perspectives

Mingchu Neal Luo, *Emporia State University, USA*

Principal performance assessment and feedback have received national attention in recent years in both the US and China. However, measuring principal leadership remains a challenge and high-quality assessments of principal performance are lacking (Elliott & Clifford, 2014). The purpose of this study is to develop and validate a scale of instrument, the Principal Leadership Index (PLI), to assess principal leadership practices within the framework of the building level leadership standards adopted by the Council for the Accreditation of Educator Preparation (CAEP, 2011) in the US. Items of the PLI were initially developed by the researcher and the experienced school administrators followed by expert reviews in the field of school leadership. The instrument was used to collect empirical data from 73 Chinese principals, which were investigated with exploratory factor analyses, reliability tests, and multilevel analyses. Results reveal four unique constructs for the leadership dimensions, showing robust psychometric properties with high levels of reliability and validity. The proposed scale can be used for both assessment and professional development to promote principal standard-based leadership behaviors.

X7-5 Estimation accuracy of IRT calibration with small sample sizes

Yue Zhao, *The University of Hong Kong, Hong Kong*
Wai Chan, *The Chinese University of Hong Kong, Hong Kong*

In educational and psychological measurement, large sample size (e.g., a minimum of 500) is desirable for accurately estimating parameters using polytomous IRT models (Kieftenbeld & Natesan, 2012; Reise & Yu, 1990). In the practice of Patient-Reported Outcome (PRO) measures, however, it is cost and labor intensive to collect large-size dataset in clinical settings. In a review of a series of published IRT studies in PRO measures, Coles et al. (2014) pointed out the lack of psychometric guidelines of sample sizes for the development of PROs for use, and reported that the sample size for published studies using IRT was as low as 100. Additionally, the skewed latent trait distribution and the extremely high discrimination parameters in PRO measures (Reise & Walker, 2009) bring the complexity for the item calibration. The study is hence designed to evaluate the capability of the recovery of IRT item and person parameters in various conditions with small sample sizes, through Monte Carlo simulations, on the basis of a realistic PRO dataset. Findings indicated that, for the most part, it is psychometrically applicable for an IRT analysis in PRO measures with a sample size below 500, but the extent to which the impact of small sample sizes on the parameter recovery varies over a number of factors. It is our hope that such a study could provide some practical guidance in the sample size issue so as to assure the appropriate use of IRT models and to improve the applications of IRT in PRO measures.

Invited Speakers: Li Cai, Jee-Seon Kim

YH-1 The utility of multidimensional item response models in educational assessment and evaluation studies

Li Cai, *University of California, Los Angeles, USA*

This presentation is focused on demonstrating the utility of multidimensional (and multilevel) item response models in educational assessment and evaluation settings. The central theme is that the flexibility of an expanded modeling framework enables the specification of models that respond to features of design, rather than forcing studies to fit into molds made up of standard measurement model choices. The first application involves a novel approach to analyze data from a multi-site randomized experimental study of the impact of learning games in middle schools. The second application describes a multidimensional model-based approach to evaluate the statistical properties of Student Growth Percentiles (SGPs), a widely used growth measure in educational evaluation and accountability systems in the US.

H6-1 Causal inference with observational multilevel data: Challenges & strategies

Jee-Seon Kim, *University of Wisconsin - Madison, USA*

This talk discusses issues and challenges for causal inference with observational multilevel data and presents strategies to remove selection bias and obtain a consistent estimate of treatment effects. Statistical methods are presented for multilevel propensity score analysis, where treated and untreated units are matched based on their estimated probabilities to receive the treatment. Unlike other multilevel matching methods, homogeneous classes of clusters are identified with respect to the selection process and then units are matched across clusters but within the homogeneous classes. This resulting multilevel latent-class logit model approach overcomes major weaknesses of existing methods and provides a flexible tool for investigating treatment effects when treatment assignment is not random. The strategy is particularly effective for handling different selection processes and/or heterogeneous treatment effects, where there might be different reasons for receiving the treatment and/or the treatment effects might vary for different subgroups in the data.

Parallel Sessions, Tuesday PM2

YH Computerized Multistage Testing: Theory and Applications [Symposium] Organizer & Chair: Duanli Yan

YH-1 Overview of test assembly methods in multistage testing

Yi Zheng, *Arizona State University, USA*

Chun Wang, *University of Minnesota, USA*
Michael J. Culbertson, *University of Illinois at Urbana-Champaign, USA*
Hua-Hua Chang, *University of Illinois at Urbana-Champaign, USA*

As a typical practice of MultiStage Testing (MST), multiple parallel panels need to be preassembled before test administration. Each panel consists of several stages and each stage contains one or more modules (i.e., item sets) assembled at different difficulty levels. During administration, an examinee is randomly assigned one of the parallel panels and takes one module from each stage, which constitutes the pathway the examinee travels throughout the test. This complex structure generates new challenges for test assembly because of the following multifaceted demands: (1) modules in one stage must have distinct information curves to sufficiently differentiate pathways through the test, (2) all pathways must be sufficiently parallel across the parallel panels, and (3) all pathways in all parallel panels must meet all nonstatistical design constraints (such as content balancing and enemy items). This becomes highly demanding especially when the item bank is limited. Although Automated Test Assembly (ATA) algorithms have been developed to assemble traditional linear tests, they must be adapted to the MST framework. This presentation first discusses how MST assembly differs from linear test assembly. Then an overview will be given as to how various existing ATA methods can be adapted to assembling MST. We will also present a new paradigm for MST called on-the-fly assembled MST, which borrows well-established item selection algorithms in Computerized Adaptive Testing (CAT) to dynamically construct individualized modules for each examinee during the test. Finally, we will discuss several possible directions for future development in MST assembly.

YH-2 Routing and scoring, estimations in multistage testing

Duanli Yan, *Educational Testing Service, USA*
Charlie Lewis, *Educational Testing Service, USA*
Alina von Davier, *Educational Testing Service, USA*

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. As in CAT, IRT and CTT can be applied to MST; however, the structural differences between CAT and MST raise important considerations. Unlike the item-by-item adaptation in CAT, MST has fixed paths along which an examinee may move through the test. Along these paths are sets of items (modules) that are administered as intact units. The rules that determine which modules are administered to an examinee are called routing rules. Generally these rules are either information-based or determined by number-correct thresholds. This paper illustrates both IRT-based and CTT-based MST routing. Proficiency estimation, or scoring, is also common to both MST and CAT. 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.

YH-3 Application of multistage testing in large-scale certification examinations

Oliver Zhang, *College Board, USA*
Krista Breithaupt, *Medical Council of Canada, Canada*
Donavan Hare, *University of British Columbia, Canada*

The past decade or so has been very productive related to adaptive models for test administration. It is clear that much of this work builds on the extensive application of CAT test designs and attendant challenges to the degree that adaptive testing models and computerized delivery has become a hallmark of modernization and forward-thinking service delivery for large-scale high-stakes testing. Comparable to the popularity of CAT in the early 1990s, we are seeing proof of concept in many programs where Multistage Testing (MST) models have been adopted. One category of such programs is the large-scale certification and licensure examinations, best represented by the Uniformed Certified Public Accountants Examinations. This paper charts the development and implementation of the CPA examination from a paper-based exam to a computer-delivered multistage adaptive test (MST). After an overview of its inception supported by an extensive research program, the chapter focuses on key decision making regarding scoring and AuTomated Assembly (ATA) models, item bank development and inventory planning, as well as challenges and practical solutions for exposure control and test security in general. It has been evident that the MST design allows pre-construction of modules for adaptive administration, while preserving the benefit of increased precision in test scores that can be derived from targeted test construction. Some implications for operational implementation are discussed, along with suggestions for future research. The paper concludes with a few more recent developments since the MST delivery model originally launched in 2004.

YH-4 Software tools for multistage testing simulations

Kyung (Chris) Han, *Graduate Management Admission Council, USA*

As with CAT, simulation techniques play critically important roles in MST development and evaluation. Because MST implementation and administration differs significantly from typical CAT, many existing CAT simulation software packages, such as SimulCAT (Han, 2012), CAT-Sim (Weiss & Guyer, 2012), and Firestar (Choi, 2009), are incapable of or inefficient while conducting MST simulations. A new MST simulation software MSTGen is introduced. This paper begins with a new MST method that replaces the preassembled test module with a test module assembled on the fly after each stage. In this method, a test module for each stage is shaped to come as close as possible to the normal density function of the interim proficiency estimate and its standard error. We call it ‘multistage test by shaping' (MST-S) and refer to the traditional multistage test as MST by routing (MST-R). MST-S offers the advantages of both MST-R and CAT. With MST-S, the difficulty of a test module always centers on the latest interim proficiency estimate, which means it potentially can administer a test module that is more efficiently adapted to the individual compared with MST-R. Because test items are not necessarily limited to a certain stage but instead are available for use in any stage, the number of items required to implement MST-S can be much smaller than those required in MST-R. This paper shows a series of simulations to compare the performance of MST-S and MST-R and CAT. MSTGen will be demonstrated during this presentation.

Y3 Stata IRT Software Demo

Y3-1 Item response models using Stata 14

Chuck Huber, *Stata Corporation, USA*

Item Response Theory (IRT) models can be used to evaluate the relationships between a latent trait of interest and the items intended to measure the trait. With IRT, we can also determine how the instrument as a whole relates to the latent trait. This talk will demonstrate how to fit a variety of IRT models using the new `irt` command introduced in Stata 14. These models include one-parameter logistic (1PL), two-parameter logistic (2PL), and three-parameter logistic (3PL) models for binary data; partial credit, generalized partial credit, graded response and rating scale models for ordinal data; nominal response models for unordered categorical outcomes; and hybrid models that include combinations of binary, categorical and ordered responses. We subsequently demonstrate how to present results with custom reports and graphs. Some items may exhibit Differential Item Functioning (DIF) for different groups and we will also demonstrate how to identify DIF using the new `difmh` command.

H6 Sponsors Session

H6-1 Quasi Linear-on-the-Fly Test (LOFT) when item parameters are not available

Hujuan Meng, *GMAC, USA*
Kyung (Chris) Han, *GMAC, USA*

H6-2 Research and innovation at Pearson

Matthew Gaertner, *Pearson Research & Innovation Network, USA*

Pearson's Research & Innovation Network was established in 2012 to produce rigorous, policy-relevant scholarship and invent capabilities and tools that support engaging, meaningful, and personalized learning. This session will overview the Network's mission, its senior scientists, and its major research initiatives. By way of illustration, we will discuss the College Readiness Index – a comprehensive system for synthesizing academic, affective, and contextual indicators to provide early college readiness diagnoses for students and help educators design targeted interventions. For additional background on the College Readiness Index, please visit http://researchnetwork.pearson.com/cri.

H6-3 Differential item functioning detection using Mantel-Haenszel procedure in ACER ConQuest

Xiaoxun Sun, *Australian Council for Educational Research (ACER), Australia*

Many methods have been developed for investigating Differential Item Functioning (DIF). Holland and Thayer developed an approach using Mantel-Haenszel statistics (Mantel & Haenszel. 1959) in detecting DIF for dichotomous items. They compared a focal group with a reference group. This presentation will describe the extension of Mantel-Haenszel procedure (Holland & Thayer, 1988) for polytomous items and with multiple focal groups if needed. This extended procedure has been implemented in ACER ConQuest. Examples of applying the procedures in ACER ConQuest will be illustrated. The DIF analysis results using extended Mantel-Haenszel procedure showed a more conservative DIF detection rates compared with the standardized difficulty difference method except for very easy or very difficult items. When calculating Mantel-Haenszel statistics, it requires a partition of k matched groups based on ability ranges for both reference and focal groups. However, there is no consensus on how to determine the number of matched groups. This presentation will address the challenge on the choice of the number of matched groups used. For standard Mantel-Haenszel procedure, items are usually classified into three categories (Zwick, Thayer, & Lewis, 1999). The categories are classified based on two factors: the absolute value of the statistic and whether or not the value is statistically significant. The method of item classification in applying extended Mantel-Haenszel procedure will be discussed.

H6-4 The validity of Scenario Based Testing (SST) in large-scale campus recruitment assessment

Alex Tong, *ATA Inc., China*

Scenario Based Testing (SST) is designed to evaluate candidates on various situations that occur in a simulated workplace. When candidates complete SST, they are put in a situation very similar to an actual workplace to solve multiple tasks through different items. The skill level, which is the most important factor to evaluate whether the candidate fits the job, is being tested during this process. The result of the study shows that SST in the competitive selection has higher validity and predicts the actual performance of candidates in the future work more effectively. For the first time in large-scale campus recruitment assessment, computer-based SST was used to assess basic work ability. In this 2013 study which took place in China, the number of examinees was more than fifty thousand. Attend this session to hear the results, and how SST overturns traditional assessment in a competitive selection process. Learn why SST should be the significant new trend in the assessment area.

H5 Non-Standard Response Types

H5-1 Thurstonian scaling of compositional questionnaire data

Anna Brown, *University of Kent, UK*

A Thurstonian IRT model (Brown & Maydeu-Olivares, 2011; 2013) was developed to overcome the problems of ipsative data in multidimensional forced-choice questionnaires. Another comparative format yielding ipsative data is the compositional format, where respondents have to distribute a fixed number of points (for instance, 100) between items. The present research extends the Thurstonian modeling approach to compositional data. Relative strengths of preferences in compositional blocks of size n are fully described by $\tilde{n} = n - 1$ ratios of points given to items to an arbitrarily chosen referent item k (for example, the last item in the block), y_i/y_k. These pairwise ratio data are distributed approximately log-normally; the log transformation $y_{\{i,k\}} = \ln(y_i/y_k)$ is applied to the ratios to achieve continuous and approximately normally distributed outcome variables (Aitchison, 1982). Following recommendations of Martín-Fernández, Barceló-Vidal, & Pawlowsky-Glahn (2003), any zeros in item scores are replaced with the smallest detectable proportion before computing the log-ratios, to avoid zero and infinity ratios. The log-ratios $y_{\{i,k\}}$ can be thought to represent arbitrarily scaled pairwise differences of items' utilities (Thurstone, 1927). The mean and covariance structure of y is analyzed using Structural Equation Modeling (SEM), assuming n latent utilities, t underlie the differences $y_{\{i,k\}}$ in each block, and that second-order factors (attributes the questionnaire measures) underlie the latent utilities. A comparison of analyses based on compositional and Likert-type responses to a Big Five measure using a sample of $N = 326$ students confirmed the suggested response process, and will be used to illustrate the approach.

H5-2 The impact of bounded measurements on the relation between the mean and the standard deviation

Merijn Mestdagh, *KU Leuven, Belgium*

Psychopathology is characterized by disturbed emotion dynamics. This is currently a hot topic in psychological research. Summary statistics, like the standard deviation, are extracted from time series of emotion measures, and these statistics are then related to levels of depression or borderline. However, in this research, emotions of participants are commonly measured in time on a bounded scale. As a result, the mean and the standard deviation may show a large correlation. This correlation can lead to problems of multicollinearity, and consequently, in difficulties when interpreting the regression model. Moreover, it is questionable that this relation limits itself to a linear dependency. We propose a new method that leads to results that are better interpreted, that avoids problems of multicollinearity and that controls for the possible nonlinear relation between variability measures and the mean.

H5-3 A confirmatory factor model for the investigation of cognitive data showing a ceiling effect: An example

Karl Schweizer, *Goethe University Frankfurt, Germany*

The ceiling effect is a method effect that is observed in items that are too easy for almost all or even all members of the sample. Such an effect leads to skewness and impairs the suitability of maximum likelihood estimation in confirmatory factor analysis because of the mismatch of the data distribution on one hand and the variables of the model of measurement on the other hand. Such a mismatch is considered with respect to the case of cognitive scores. It is assumed that the ceiling effect observed in such scores can be described by means of a binomial distribution. Constraints for factor loadings that reflect the ceiling effect, as it is described by means of a binomial distribution, are considered for balancing the ceiling effect in confirmatory factor analysis. This possibility of dealing with a ceiling effect is investigated in data obtained by means of Exchange Test, a measure of working memory involving five different treatment levels that are characterized by different degrees of difficulty. Using an empirical dataset it is demonstrated that controlling for the ceiling effect by appropriate weights added to the constraints considerably improves the model-data fit. Furthermore, simulated data were constructed according to the cognitive data and were investigated. The results indicate that this way of controlling for the ceiling effect improves the model-data fit.

XH Cognitive Diagnosis Models II

XH-1 Item selection strategies based on attribute mastery probabilities in CD-CAT

Xiaofeng Yu, *Jiangxi Normal University, China*
Zhaosheng Luo, *Jiangxi Normal University, China*
Chunlei Gao, *Jiangxi Normal University, China*
Yujun Li, *Jiangxi Normal University, China*
Yafeng Peng, *Jiangxi Normal University, China*

The key to a Cognitive Diagnostic Computerized Adaptive Testing (CD-CAT) system is the item selection strategies. Some popular strategies are developed based on the Kullback-Leibler (KL) information and Shannon Entropy (SHE). Typically, during CD-CAT, these familiar methods would use a cutoff point to transform the attribute mastery probabilities' provisional values to binary values, but at the initial stage, the cutoff point method may lead to a larger deviation. A method that can take advantage of the probabilistic information with regard to attributes may offer a better alternative. This paper proposed two item selection strategies based on the provisional value of the attribute mastery probabilities, as follows: (1) the first strategy, which is called as PPWKL, is based on the KL information, and can lead to a maximum difference of the sum of attribute mastery probabilities; (2) the PPWKL considers the fact that not all patterns are equally likely, but overlooks the fact that the distances between different patterns and the current estimate are not of equal importance. Therefore, PPWKL can be weighed by the inverse of the distance between $\widehat{\alpha}$ and any possible latent states, which is called PHKL. Three simulation studies were carried out, one was the fixed length of CD-CAT, and the second was the variant length CD-CAT, and the last was the short length CD-CAT. The simulation results indicate that PPWKL and PHKL have better comprehensive performances in fixed and variant length CD-CAT, they can retain a good measurement accuracy, and also improve the utilization ratio of the item pool.

XH-2 Multi-stage testing with cognitive diagnosis

Chunlei Gao, *Jiangxi Normal University, China*
Zhaosheng Luo, *Jiangxi Normal University, China*
Xiaofeng Yu, *Jiangxi Normal University, China*
Yujun Li, *Jiangxi Normal University, China*
Yafeng, *Jiangxi Normal University, China*

Multi-Stage Testing (MST) is a form of computerized testing which aims to overcome the disadvantages of CAT. It incorporates most of the advantages of CAT and linear testing. Cognitive Diagnosis Assessment (CDA) aims to determine whether or not examinees have each of many attributes or skills underlying responses to test items. In contrast to IRT, CDA provides a more detailed evaluation of the strengths and weaknesses of students. This study presents Cognitive Diagnostic Multi-stage Testing in order to increase the application of CDA. CD-MST is a new idea which combines CD with MST. CD-MST has many advantages. It has the functions of cognitive diagnosis and is adaptive; compared with CD-CAT, it has speed superiority, and allows the examinees to go back to check and revise; it is also a flexible testing model, the test takers can arrange the testing designs according to their demand. CD-MST can solve some practical problems and has theoretical and practical value. The research used two simulations to illustrate the CD-MST. First, we tried to find out if the designs of CD-MST could influence the results of CD-MST. Second, we wanted to contrast the CD-MST and the CD-CAT. Two contributory factors were considered, the quality of the item bank and the item selection strategy. We also investigated the speed of CD-MST and CD-CAT. The results showed that, compared with CD-CAT, CD-MST was more influenced by the quality of the item bank, the recovery rate of CD-MST was as high as CD-CAT when the high quality item bank was used. CD-MST can take 2/3 less time than CD-CAT.

XH-3 Comparing approaches in cognitive diagnostic computerized adaptive testing

Miao Gao, *Nanjing Normal University, China*
Ren Liu, *University of Florida, USA*
Anne Corinne Huggins-Manley, *University of Florida, USA*

Obtaining both continuous test scores (θ) and categorical attribute mastery profiles ($\boldsymbol{\alpha}$) in a single administration is a test outcome desired by many educational practitioners, and the combination of Cognitive Diagnostic modeling and Computerized Adaptive Testing (CD-CAT) offers an efficient way to provide this information. Three distinct CD-CAT approaches have been proposed: the Shadow Test Approach (STA), the Constraint-Weighted Approach (CWA), and the Aggregated Information Approach (AIA). However, the question of which approach produces more accurate θ and $\boldsymbol{\alpha}$ estimates under different conditions has not been addressed in the literature. This simulation study compares the θ recovery rate and the accuracy of θ estimates across these three approaches while varying item selection methods and psychometric models. Results show that the CWA produces the most accurate θ and $\boldsymbol{\alpha}$ estimates across most conditions, and the STA tends to yield accurate θ but biased $\boldsymbol{\alpha}$ estimates. AIA largely fluctuates in its accuracy in θ and $\boldsymbol{\alpha}$ estimation when different attribute-level weights are assigned. However, AIA has the unique advantage that fitting separate models instead of one model for the dual purposes would help link responses from different test forms, groups or times. Detailed comparison results are discussed to guide practitioners in choosing an appropriate CD-CAT approach when implementing.

XH-4 Evaluating performance of differential item functioning detection methods for DIF data in DINA model

Önder Sünbül, *Mersin University, Turkey*
Seçil Ömür Sünbül, *Mersin University, Turkey*

This study aims to evaluate the performance of Differential Item Functioning (DIF) detection methods (Mantel-Haenszel, Logistic Regression, Lord's Chi Square and Raju's Signed Area) for Cognitive Diagnosis Model (CDM) DINA type data which contain differentially functioning items in terms of Type I Error and power rate. For the Type I error study, same *s* and *g* parameters were used for both focal and reference groups. Performance of DIF detection methods were investigated for sample size (250, 500, 1000, 2000, 3000) and correlation between attributes (0.2, 0.5, 0.8) for the Type I error study. For the power study, the *s* and *g* parameters were manipulated to cause differences between focal and reference groups. For this purpose, data were generated according to sample size (250, 500, 1000, 2000, 3000), correlation between attributes (0.2, 0.5, 0.8), amount of DIF (0.075, 0.15), and number of attributes measured by differential functioning items (1, 2, 3). The "CDM" package was used for data generation and "difR" package was used for DIF detection methods in the R programming language. Main effects and interaction effects of simulation criteria were evaluated for DIF detection methods for *s* and *g* parameters separately. The results of the Type I error study showed that none of simulation criteria had any significant effects for the Type I error rates in terms of main and interaction effects. The results of the power study showed that all of the simulation criteria, except the correlation between attributes, had significant main and interaction effects for DIF detection methods.

X3 Rater Effects in IRT

X3-1 Parameter recovery and estimation of the longitudinal hierarchical rater model

Jodi M. Casabianca, *The University of Texas at Austin, USA*
Mark Bond, *The University of Texas at Austin, USA*
Brian W. Junker, *Carnegie Mellon University, USA*

Rater effects in education testing and research have the potential to impact the quality of scores in constructed response and performance assessments. The Hierarchical Rater Model (HRM; Patz, et al., 2002) yields estimates of latent traits that have been corrected for individual rater bias and variability. This research presents an extension of the HRM called the Longitudinal HRM (L-HRM) that includes an autoregressive time series component as well as a parameter for overall growth for longitudinal ratings. This paper confirms and extends partial preliminary results presented at the 2014 IMPS, providing a complete set of findings from a Monte Carlo simulation study used to determine how the L-HRM performs under various conditions for a 5-item test. We varied the number of time points, (3, 7), the number of raters, (3, 6), and sample size (250, 500) for a linear and a logistic trend. We fitted

the L-HRM as a Bayesian model using MCMC estimation in JAGS (Plummer, 2003) via the R2jags package (Su & Yajima, 2012). Parameter recovery results reveal negligible bias for most parameters across conditions. In addition to simulation study results, we will discuss issues with computation and estimation, study limitations, and preliminary results from research currently in progress to improve and further evaluate the L-HRM.

X3-2 The multifaceted IRT Model for examinee-selected items

Wen-Chung Wang, *Hong Kong Institute of Education, Hong Kong*
Xue-Lan Qiu, *Hong Kong Institute of Education, Hong Kong*

In addition to mandatory items that all examinees must answer, examinees may be requested to answer a fixed number of items from a given set of items (e.g., answering 2 among 5 items). These so-called Examinee-Selected (ES) items bring challenge to standard IRT models because unselected items may be missing not at random. A new class of IRT models have been proposed for ES items (Wang, Jin, Qiu, & Wang, 2012; Wang & Liu, 2015). These models have two facets: person and item. Often, ES items are in a constructed-response format and are marked by raters. Therefore, in addition to item difficulty and person ability, rater severity also plays a role. Moreover, a rater may not be able to hold a constant severity throughout the whole rating process. This study thus proposed a new and general multifaceted IRT model to account for rater effects in ES items. Preliminary simulation studies show that the parameters of this new model could be well recovered by the freeware JAGS. An experiment with the "Choose-one-answer-all" design (Wang, Wainer, & Thissen, 1995) was conducted to collect complete data in ES items. The new model was validated by the empirical data.

X3-3 Applications of computerized adaptive testing in evaluating rater effects

Zhuoran Wang, *University of Minnesota, USA*
Chun Wang, *University of Minnesota, USA*
Tian Song, *University of Minnesota, USA*
Edward W. Wolfe, *Pearson, USA*

Scores are assigned to responses by raters in many scenarios, such as essay grading, oral test scoring, and artistic performance scoring. In order to get a credible rating, rater quality should be monitored. Current rater monitoring procedures focus on identifying rater effects using a fixed set of validity papers. To improve the efficiency of those procedures, we propose to use an adaptive strategy by selecting the validity responses adaptively such that fewer validity essays are needed to accumulate enough information about raters. In this study, we based our adaptive design on the Rating Scale Model (RSM), in which the essay parameters are assumed pre-estimated and known (which serve as the "item" parameters in a classical CAT). An essay selection algorithm is proposed to maximize the accuracy of detecting rater effects, and the specific types of rater effect we consider is the severity/leniency effect. A simulation study is conducted to compare the performance of CAT and traditional fixed set of essays in detecting rater effects. The results show that adaptive tests have a more accurate rater parameter estimation compared to linear tests of the same length. The practical implications of the simulation results are further discussed.

X3-4 Adaptive rater monitoring

Chun Wang, *University of Minnesota, USA*
Tian Song, *University of Minnesota, USA*
Edward W. Wolfe, *Pearson, USA*

A common concern about the scores assigned by human raters in educational settings is the degree to which those scores contain measurement error due to the subjectivity of raters' judgments. Due to these concerns, the quality of human ratings is routinely monitored during the scoring process. Current rater monitoring procedures often ask raters to score a fixed set of validity papers (that have been assigned a consensus score by experts), and then compare their ratings to the experts' score and identify any rating patterns that deviate from true scores, which are commonly referred to as rater effects. To improve the efficiency of those procedures, one way is to adaptively select the most informative validity papers for each rater based on the rater's ratings on the previously administered validity papers. This way, the number of validity papers assigned to each rater will decrease, whereas the amount of information accrued for each rater stays roughly the same. In this paper, a Computerized Adaptive Testing (CAT) algorithm is developed in the context of rater monitoring, and two simulation studies are conducted to evaluate the performance of the adaptive system: (1) a fixed-length CAT, which shows with the same number of validity papers, the proposed method can reach higher precision in terms of more accurate recovery of rater effects and higher classification accuracy; (2) a variable-length CAT, which shows the proposal method needs fewer number of validity papers to accrue the same amount of information.

X7 Statistical Learning

X7-1 Test assembly using a quantum particle swarm optimization approach

Jiayuan Yu, *Nanjing Normal University, China*
Wei Hang, *Jiangsu Zhuo Yun Information Technology Co. Ltd, China*

In psychological measurement, test assembly is often a combinatorial optimization problem with constraints. The Quantum Particle Swarm Optimization (QPSO) approach is a technique used to find the optimal solution to a problem that is difficult to solve analytically. This article introduces the use of QPSO to perform automated test assembly. In the simulation study, QPSO was applied to item selection and test assembly for the Chinese Language Proficiency Test (HSK). It was also compared with the construction of a test using the Particle Swarm Optimization (PSO) approach. Based on Item Response Theory (IRT), we developed a simulation tool to generate a virtual item bank containing 7000 three-parameter items. The values of the three parameters were randomly generated. We established the objective function according to the maximum information principle. Using the orthogonal design method, we found the optimal parameter combination of QPSO. Performance of test assembly was evaluated using test information at the cutoff score, the flatness and robustness. The simulation results showed that the optimal parameter combination of QPSO were inertia weights w1 and w2 equal to 1.2 and 0.3 respectively, with the number of particles and iterations equal to 40 and 300 respectively. Comparing the performance indicators of QPSO and PSO for test assembly, it was found that there were no significant differences in flatness. However, in terms of test information and robustness, QPSO was found to be superior to PSO. Thus, from the above three indicators, we conclude that the quantum particle swarm optimization approach is a good method for test assembly.

X7-2 The use of support vector machines in cognitive diagnostic assessments

Ying Cheng, *University of Notre Dame, USA*
Cheng Liu, *University of Notre Dame, USA*

Cognitive diagnostic modeling in education measurement has attracted much attention from researchers in recent years. Its applications in real assessments, however, have not gained as much traction. One reason might be model fit–even though a myriad of Cognitive Diagnostic Models (CDMs) have been developed, it is difficult to identify the model that best fits the data, and it is not uncommon that none of the models fit all items on a test. Computational complexity is another issue due to the large number of possible latent classes (2^K, where K is the number of attributes.) We propose to use the Support Vector Machine (SVM), a popular supervised learning model, to make classification decisions on each attribute (i.e., if the student masters the attribute or not) given a training dataset. By using SVM we convert the problem of fitting a CDM, of calibrating the CDM parameters from a calibration sample, and of obtaining a test taker's latent profile into a quadratic optimization problem in hyper dimensional space. No $O(2^K)$ computational time is needed when we evaluate new test takers. Our simulation study shows that an accuracy rate around .95 could be achieved with only a training sample of size 50, when 6 attributes and 30 items are simulated, given there is no error in labeling the latent profile in the training sample. This method has great promise to significantly increase the usability of cognitive diagnostic modeling in education measurements.

X7-3 Random forests for non-homogeneous zero-inflated Poisson processes

Denis Larocque, *HEC Montreal, Canada*
Walid Mathlouthi, *HEC Montreal, Canada*
Marc Fredette, *HEC Montreal, Canada*

Random forests (Breiman, 2001) is one the most popular and powerful nonparametric modeling methods. Originally proposed for continuous and categorical responses, this approach has since been generalized to more complicated settings including survival data (Zhou & McArdle, 2014). Count data is another type of response encountered frequently in practice. In this paper, we propose a random forest method for Zero-Inflated Poisson (ZIP) data. The key idea is to link two forests, one to model the zeros, and one to model the count of non-zero observations. The method is also extended to the case of non-homogeneous ZIP data. Simulation studies show that the new method performs well compared to basic Poisson forests and to a forest of ZIP trees using the method of Lee & Jin (2006). To appear in *Psychometrika*.

X7-4 A Dantzig selector composite likelihood approach for multivariate generalised linear mixed models

Vassilis Vasdekis, *Athens University of Economics and Business, Greece*
Kostas Florios, *Athens University of Economics and Business, Greece*
Irini Moustaki, *London School of Economics, UK*
Dimitris Rizopoulos, *Erasmus University Medical Center, The Netherlands*

Analysis of multivariate longitudinal data allows studying several items across time. Estimation of Generalised Linear Mixed Models (GLMM) using full information maximum likelihood can be difficult in this type of data due to the need for computation of a large number of integrations. Composite likelihood approaches provide a solution in this difficult situation. The Dantzig selector performs variable selection and model fitting in linear regression and generalised linear models. This is accomplished by using an L1 penalty to shrink the fixed effects model coefficients towards zero. We extend the Dantzig selector to fit multivariate GLMMs and develop a computationally efficient algorithm which is based on a linearization of the composite score function. Simulations show good performance for estimation and model selection.

Wednesday, July 15

Parallel Sessions, Wednesday AM

YH Advances in Modeling and Analysis for Educational Testing [Invited Symposium] Organizer & Chair: Jingchen Liu

YH-1 Detecting aberrant behavior in standardized testing

Gongjun Xu, *University of Minnesota, USA*
Chun Wang, *University of Minnesota, USA*

The modern web-based technology greatly popularizes computer administered testing, also known as online testing. Compared with traditional Paper-and-Pencil (P&P) tests, computer administered tests provides better control over user-role access to test and test results due to its feature of immediate scoring. However, because online tests are usually administered continuously within a certain testing window, many items are likely to be exposed and compromised, posing a new type of test security concern. This paper proposes a new mixture hierarchical item response theory model, using both response accuracy and response time information, to help detect aberrant behavior and item compromise. The new model-based approach is compared to the traditional residual-based fit statistic in both simulation study and real data example. Results show that the mixture model approach considerably outperform the residual method, in particular when the proportion of aberrance is high.

YH-2 The effect of anchoring vignette scoring on reliability and validity

Matthias von Davier, *Educational Testing Service, USA*

Anchoring vignettes are used in surveys adjust for individual tendencies in the use of response formats. These types of tendencies are often referred to as response styles, and are assumed to distort the 'true' answers of the respondent towards the use of a preferred set of response categories, for example, as seen in the tendency to choose extreme categories (e.g. Rost, Carstensen & von Davier, 1996). Scores obtained from anchoring vignettes are used to correct for these response tendencies (e.g. King et al., 2004) either by using a latent variable model (King et al., 2004, Rabe-Hesketh & Skrondal, 2002), or a non-parametric re-scoring rule (King et al, 2004, 2007).

The mechanics behind the transformation of responses by anchoring vignettes is using an estimate of individual response thresholds for each respondent. While this may be done in a hierarchical manner using covariates, the transformation of observed responses varies across respondents. This talk examines the effect of anchoring vignette scoring on reliability and validity of the results based on simulated and real data.

YH-3 From computerized adaptive testing to adaptive learning and beyond

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

This presentation provides a survey of 30 years' progress in Computerized Adaptive Testing (CAT). We start with a historical review of the establishment of a psychometrical foundation for CAT. Then, we address a number of issues emerging from large scale implementation and show how theoretical works can be helpful to solve the problems. A new focus will be on Cognitive Diagnostic (CD) CAT that can be utilized as diagnostic tools for schools to classify students' mastery levels for a given set of cognitive skills that students need to succeed. In addition, we will show how CD-CAT can be very helpful to support individualized learning on a mass scale. Lastly, we will ruminate on and discuss some possible future directions of research on CAT. We believe CAT has a great potential to work extremely well in various applications of MOOCs and also in online educational data mining.

YH-4 A hypothesis testing procedure for Q-matrix entries

Matthew S. Johnson, *Columbia University, USA*

de la Torre (2008) introduced a method, which he called the EM based δ-method, for empirically evaluating the Q-matrix within the DINA framework. The method compares the pseudo-empirical proportions of correct responses between groups of examinees that either have or do not have the (proposed) required skills for an item. In this talk I present the asymptotic distribution of these pseudo-empirical differences in the proportions correct, denoted by $\delta_j(\mathbf{q})$. I then investigate, through simulation, the power and Type I error rates for hypothesis testing procedures for the correct specification of the q-vector for a given item, i.e., $H_0 : \mathbf{q}_j = \mathbf{q}_{j0}$, against a simple alternative hypothesis, $H_1 : \mathbf{q}_j = \mathbf{q}_{j1}$, and for testing the correct specification of single entry in the Q-matrix, e.g., $H_0 : q_{jk} = 0$. Finally I demonstrate the method by applying it to Tatsuoka's fraction subtraction data.

YH-5 Regularized latent class analysis with application in cognitive diagnosis

Jingchen Liu, *Columbia University, USA*

Diagnostic classification models are confirmatory in the sense that the relationship between the latent attributes and responses to items is specified or parameterized. Such models possess a good interpretability and each component of the model usually has a practical meaning. On the other hand, parameterized diagnostic classification models are sometimes oversimplified and they are not flexible enough to capture all the data patterns. Thus, model lack of fit is often significant. To achieve a good fit, more sophisticated parametric or nonparametric families of models are often needed. However such models usually lack interpretability due to model complexity. In this talk, we try to obtain a compromise between interpretability and goodness of fit by imposing regularization on a latent class model. This approach starts with minimal assumptions on the data structure other than the discreteness of the latent variables and reduces the model complexity by imposing regularization on the parameters to reach better interpretability. An expectation-maximization algorithm is developed for the computation.

Y3 Recent Developments in Latent Variable Models for Complex Data Structures [Symposium] Organizer: Silvia Cagnone; Chair: Irini Moustaki

Y3-1 Remedies for degeneracy in Candecomp/Parafac

Paolo Giordani, *Sapienza University of Rome, Italy*

Three-way data arrays usually refer to a set of observation units on which a number of quantitative variables are collected during several (time) occasions. The Candecomp/Parafac (Carroll & Chang, 1970; Harshman, 1970) model can be seen as an exploratory latent variable model allowing us to study such arrays by extracting unobservable factors. These factors can be seen as latent variables underlying the observed data. Unfortunately, the use of Candecomp/Parafac can be difficult due to the risk of obtaining degenerate solutions, i.e. highly correlated factors. A possible remedy to this problem is to constrain the factors to be orthogonal. However, this strategy solves the problem from a theoretical point of view, but fails in practice. In fact, the orthogonality constraint is often a very strong assumption preventing the recovery of the true and unknown factors underlying the data. For this reason, some alternative strategies for avoiding Candecomp/Parafac degeneracy are discussed. These are the so-called Candecomp/Parafac with Lasso constraints (Giordani & Rocci, 2013), the Candecomp/Parafac with ridge regularization (Giordani & Rocci, 2013) and the Candecomp/Parafac with SVD penalization (Giordani & Rocci, 2015).

Y3-2 Parsimonious representations of finite mixture models for multivariate mixed responses

Marco Alfó, *Sapienza University of Rome, Italy*
Paolo Giordani, *Sapienza University of Rome, Italy*

We specify a flexible regression model for multivariate mixed responses, conditional on a set of unobserved, outcome-specific, latent parameters, that represent the effects of unobserved individual-specific heterogeneity. We adopt a finite mixture representation, which is a semi-parametric version of parametric methods based on Gaussian quadrature. It consists in leaving unspecified the random parameter distribution and in estimating it through a discrete distribution on a finite number of locations. This approach is based on unidimensionality of the latent structure, with the same components describing heterogeneity within (and dependence between) margins. When the task is testing for dependence, or several outcomes are considered, this issue may pose some problems. Starting from Alfó & Rocchetti (2013), a general representation of the multivariate distribution of the random parameters is introduced, where outcome specific heterogeneity is separated from dependence between outcomes. The outcome-specific random parameter distribution is estimated by a discrete distribution on a limited outcome-specific number of locations and the "global" random parameters distribution is based on joining all the possible combination of (outcome-specific) locations. The joint probabilities of the multivariate random parameter distribution is stored in a multiway array, with dimension equal to the number of outcomes. While it is flexible enough, the proposed parameterization suffers from a complexity which is exponential in the number of outcomes. The array can be synthesized by means of the Parafac (Carroll & Chang, 1970; Harshman, 1970) to reduce the complexity of the adopted parameterization.

Y3-3 Dynamic, random-coefficient based, missing data models

Maria Francesca Marino, *Sapienza University of Rome, Italy*

Marco Alfó, *Sapienza University of Rome, Italy*

Longitudinal studies are often used to describe the individual evolution over time in terms of a regression model built to analyse the relation between a response variable and a set of covariates. Due to the presence of multiple measurements for each individual, observed data are likely to be associated. Individual-specific sources of unobserved heterogeneity are typically considered in the model specification to describe this association. We propose a latent Markov model for the longitudinal response, where a potentially non-ignorable missing data mechanism, in the form of irretrievable drop-out, may be explicitly taken into consideration. Unobserved individual-specific features may influence the missing data process as well, but may be shared, correlated or independent on the unobserved heterogeneity we have considered in the primary model. We specify a mixed effect regression model for the missing data indicator variable, where sources of unobserved heterogeneity are described by (a second set of) individual-specific random parameters. The potential dependence between the primary longitudinal and the missing data models is due to the dependence between the sources of unobserved heterogeneity in the two equations; in particular, these are assumed to be conditionally independent given an upper-level discrete latent variable. The "joint distribution" of, and therefore the dependence between, the two sources of unobserved heterogeneity arises due to the presence of this upper-level structure. By adopting this model specification, we define a MNAR model which nests the corresponding MAR counterpart, giving also the way to compare the parameter estimates derived by using the two approaches.

Y3-4 A multivariate latent variable model for analysing longitudinal mixed data

Silvia Cagnone, *University of Bologna, Italy*
Cinzia Viroli, *University of Bologna, Italy*

A latent variable model for the analysis of multivariate mixed longitudinal data is proposed. It extends the factor mixture model discussed by Cagnone & Viroli (2012) to longitudinal data. The model is based on the introduction of two hidden variables: a continuous latent variable for modeling the association among the observed variables at each time point and a latent discrete variable that follows a first-order Markov chain with the aim of taking into account the unobserved heterogeneity. The aim of the proposed model is twofold: it allows us to perform dimension reduction when data are of mixed type and it performs model based clustering in the latent space. We derive an EM algorithm for the maximum likelihood estimation of the model parameters. The method is illustrated by an application to a longitudinal dataset on aging and health.

Y3-5 The latent variable ALT model: A general framework for longitudinal data

Silvia Bianconcini, *University of Bologna, Italy*
Kenneth A. Bollen, *University of North Carolina at Chapel Hill, USA*

In recent years, longitudinal data have become increasingly relevant in many applications, heightening interest in selecting the best longitudinal model to analyze them. General longitudinal models provide researchers a means to select the most appropriate model for a given application. In this regard, Bollen & Curran (2004) developed the Autoregressive Latent Trajectory (ALT) model, here denoted as observed variable ALT. The model grew out of the aim to capture the desirable features of both latent growth curve and autoregressive models, having the ability to discriminate between these two approaches to model panel data. The purpose of this paper is to develop the latent variable ALT model as a generalization of the observed variable ALT. We show how the latent variable ALT model under constraints can specialize to a wide variety of other longitudinal models. Hence, if theory or prior work dictate the model, then latent variable ALT is likely capable of specialising to that structure. On the other hand, if there is little guidance on the best model to be selected, latent variable ALT provides a way to empirically compare a wide variety of models and determine the most appropriate for the data. Furthermore, the latent variable ALT model provides a framework which reveals the connections between many longitudinal models that were previously considered as distinct.

H6 Item Generation/Item Banks

H6-1 Exploiting properties of a feasible set to improve item pool utilization

Dmitry I. Belov, *Law School Admission Council, USA*
Madison Williams, *Law School Admission Council, USA*
David Kary, *Law School Admission Council, USA*

In high-stakes testing programs, the number of non-overlapping tests (i.e., tests with no items in common) assembled from a given item pool is a critical measure of pool utilization. Since items are constantly being withdrawn from and added to the pool, test developers need to know what type of new items are needed to improve pool utilization. In practice, the simultaneous assembly of multiple non-overlapping tests is intractable. Several heuristics will be discussed to address this problem with emphasis on heuristics that also predict the types of new items that will improve pool utilization. It will be shown that the performance of heuristics depends on how well they exploit the properties of the feasible set (i.e., a set where each element is a unique test). Mathematically these properties describe the distribution of cliques in a graph formed by elements of the feasible set, where an edge between two elements exists if and only if two corresponding tests are non-overlapping. The performance of these heuristics will be assessed using the Law School Admission Test (LSAT) specifications and item pool.

H6-2 Enhancing item pool utilization for designing multi-stage computerized adaptive tests

Lihong Yang, *Michigan State University, USA*
Mark D. Reckase, *Michigan State University, USA*
Mingcai Zhang, *Michigan State University, USA*

An ideal item pool consists of an appropriate number of item combinations that meet all content specifications of a test and provide sufficient estimation for examinees at different proficiency levels. With the increasing demand for Multi-Stage computerized adaptive Tests (MSTs) in operational testing, this paper evaluates the effectiveness of a new methodology to enhance item pool utilization for designing various MST panel designs. The performance of different MST panel designs assembled under different test configurations is compared and the most optimal design in terms of more accurate ability estimate and test efficiency is selected. The results from this research will provide information for efficiently designing optimal item pools for multistage computerized adaptive test, facilitating the MST assembly process, and improving scoring accuracy.

H6-3 Crowdsourcing item generation

Francis Smart, *Michigan State University, USA*
J. Lucas Reddinger, *University of California, Santa Barbara, USA*

Automated item generation theory informs the construction of large quantities of assessment items through automated means with pre-specified attributes. Yet due to the complexity of cognition, these methods have been difficult to implement for many complex cognitive tasks. At the same time, "crowdsourcing" has become a powerful tool for data generation in that tasks typically easy for humans and difficult for computers are distributed across a large network of human workers. Although crowds have been used extensively to calibrate items, little work has been done using crowds to construct items. Through the use of crowdsourcing methods, we have generated a bank of 18,000 raw items in the content domains of Earth Science, Biology, and US History. Half these items are multiple choice items with one correct answer and three distractors while the remaining are true/false statements. We also propose crowdsourcing methods to evaluate items in terms of cognitive demand, appeal, and difficulty, and to flag items for grammatical or technical revision. We construct an index using this information to select top items for revision. From a preliminary study of 1,500 raw crowdsourced Earth Science items, 61% are reported as either "liked" or "strongly liked", with less than 10% of items reported as requiring moderate or extensive edits. When scaled by Bloom's Taxonomy, 90% of items are reported as only testing the domains of remembering and understanding. In conclusion, the use of crowdsourcing to generate items is very promising.

H6-4 A comparative study of online item calibration methods in multidimensional computerized adaptive testing

Ping Chen, *Beijing Normal University, China*

Calibration of new items online has been an important topic in both psychological and educational measurement. Recently, four online calibration methods were developed in the context of Multidimensional Computerized Adaptive Testing (MCAT): M-Method A, M-OEM, M-MEM (Chen, Wang, Xin & Chang, 2013) and FFMLE-M-Method A (Chen & Wang, 2014). To more accurately and efficiently calibrate the new items, this paper proposes two new MCAT online calibration methods by combining Bayes modal estimation (BME; Mislevy, 1986) with M-OEM and M-MEM, and they are referred to as M-OEM-BME and M-MEM-BME, respectively. Simulation studies were conducted to compare the performance of the six methods in terms of item-parameter recovery under three levels of sample sizes ($N = 900$, 1,800 and 3,600) and three levels of correlations between dimensions ($R = 0$, 0.5 and 0.8), assuming the new items are randomly assigned to examinees. The simulation results showed that considering the prior distribution of the item parameters of the new items is helpful to improving the calibration precision for M-MEM, but this is not true for M-OEM. M-MEM-BME (M-OEM) worked best under all sample sizes when $R = 0$ ($R = 0.8$) and FFMLE-M-Method A had the best performance for a sample size of 3,600. For all six methods, a larger sample size (correlation between dimensions) produces more (less) precise parameter recovery.

H6-5 Exploring online calibration of polytomous items in computerized adaptive testing

Yi Zheng, *Arizona State University, USA*

Online calibration is a technology-enhanced strategy for pretesting new items in Computerized Adaptive Tests (CATs). Many CATs are administered continuously over a long-term and rely on large item banks. To ensure test validity, these item banks need to be frequently replenished with new items, and these new items need to be pretested before being used operationally. The concept of online calibration is to dynamically embed pretest items in operational tests and calibrate their parameters as response data obtained through the continuous test administration. The idea of embedding pretest items in operational tests is not new, but the added value of online calibration is dynamic. Online calibration allows for individualized treatment for each pretest item, such as adaptive item selection to match each pretest item with the most informative examinees, or flexible termination of individual pretest items from the pretesting stage. With rapidly growing computer capacities, online calibration could soon become a standard procedure in the management of long-term, large-scale CATs. The proposed study builds upon existing literature on methods for online calibration of dichotomously scored items. Previous studies have developed a variety of formulas, procedures, and computer algorithms for dichotomous item response theory models. This study extends those formulas, procedures, and algorithms to polytomous item response theory models, which are used to analyze items scored in more than two categories. Polytomous items, such as performance-based items, are becoming increasingly important in educational assessments. This study will fill in the blank in online calibration methods for polytomous items.

H1 Component & Scaling Analysis

H1-1 Multidimensional joint graphical display: Back to the basics

Shizuhiko Nishisato, *University of Toronto, Canada*

The basic premise of dual scaling/correspondence analysis lies in the simultaneous analysis of rows and columns of the data matrix, making it almost compulsory to provide graphical display of both rows and columns in the common Euclidean space. The traditional displays such as symmetric, non-symmetric graphs, CGS scaling and a variety of biplots are all flawed with respect to the basic premise: they do not provide a precise description of complex information in the data, hence failing in the purpose of graphical display. Unfortunately, “seeing is believing” does not apply to the current practice of joint graphical display in quantification theory. The current paper will discuss the logical problems and offers a justifiable alternative to joint graphical display of categorical data, the idea first proposed by Nishisato & Clavel (2010).

H1-2 Sparse core Tucker2 for computationally identifying the optimal model between parafac and Tucker2

Hiroki Ikemoto, *Osaka University, Japan*
Kohei Adachi, *Osaka University, Japan*

Three-way principal component analysis (3WPCA) techniques for a data array of objects by variables by sources can be formulated with the core matrices that explain the inter-sources differences in the relationships between the components underlying objects and those for variables. Among 3WPCA, extremes are are Tucker2 and Parafac, in that the core elements are unconstrained in the former and restricted to zeros except diagonal elements in the latter. In this presentation, we propose a technique intermediate between Tucker2 and Parafac. In this technique, a Tucker2 loss function is minimized subject to the constraint that a specified number of elements in core matrices are zeros: the optimal locations of zero elements and nonzero parameter values are estimated simultaneously. It can be named sparse core Tucker2 (ScTucker2), as the matrices with a number of zero elements are said to be sparse. We present an alternating least squares algorithm for ScTucker2 and present the procedure for selecting the suitable number of zero elements. The behavior of ScTucker2 is assessed in a simulation study and illustrated with real data examples.

H1-3 Applications of principal cluster components analysis to scale development and the validation of clustering by a resampling method

Takashi Murakami, *Chukyo University, Japan*

Principal Cluster Components Analysis (PCCA) is a procedure to cluster variables to define linear composites with maximal explained variances based on the matrix of correlations between variables. It has some preferable characteristics to the usual cluster analysis of variables as a method of constructing of several subscales as sum scores of item responses according to the concepts implied by the construct to be measured. We will explain that PCCA can be viewed as a method for obtaining a matrix of loadings with perfect simple structure, and we will show several simple propositions for PCCA, which are useful in practical usages. An illustrative example will be presented in order to show the process of determining the number of subscales and item selection in the analysis of data obtained by a preliminary survey, the use of a resampling method evaluating the stability of the clustering, and the appropriateness of the method for choosing the number of clusters.

H1-4 Component-based item response theory

Ji Hoon Ryoo, *University of Virginia, USA*
Kwanghee Jung, *University of Texas, USA*
Heungsun Hwang, *McGill University, Canada*
J. Patrick Meyer, *University of Virginia, USA*

Item Response Theory (IRT) is a widely used statistical modeling method in psychometrics, primarily to estimate item and person parameters to render the true ability (or endorsement) interpretable. Estimation generally utilizes one of three Maximum Likelihood Estimation (MLE) methods: conditional, full, and marginal MLE. Although popular, all three depend upon distributional assumptions such as multivariate normality and therefore require large amounts of data to achieve stable estimation, forcing researchers to deal with an estimation problem. We will incorporate Generalized Structured Component Analysis (GSCA) into the common IRT model to help address this small sample and estimation issue. GSCA differs from the conventional IRT by using the ordinary least square estimation method to determine a composite component score rather than a factor score for ability, allowing us to estimate parameters without the assumption of normality and thus compute parameters faster than MLE for both small and large data sources. The efficiency of GSCA has already been proven in the neuroscience field for applications such as fMRI analysis and in the genetic epidemiology. In this presentation, we will discuss a development of a theoretical foundation of component-based IRT using nonlinear-GSCA that is the extension of GSCA to categorical indicators, making it possible to apply GSCA to qualitative data such as nominal and categorical data.

H5 Lasso/Penalization & Trees

H5-1 Using LASSO penalization for explanatory IRT: An application on instructional covariates for mathematical performance in a large-scale assessment

Marije Fagginger Auer, *Leiden University, The Netherlands*

LASSO penalization can be used for models with very high numbers of covariates. The regression parameters are penalized and more of them become zero as the penalty increases, thereby resulting in covariate selection. LASSO has so far been applied in many (generalized) linear models, but has only recently been extended to Generalized Linear Mixed Models (GLMMs), allowing for the modeling of correlated observations. Since explanatory Item Response Theory (IRT) models can also be seen as GLMMs, this means that the lasso penalization is now also possible for explanatory IRT models. This LASSO IRT can be especially useful for educational large-scale assessment data. This type of data is typically analyzed using IRT (to allow for the linking of different item sets) and questionnaire data on many instructional variables is collected to investigate influences on performance. However, these effects are considered mostly in isolation in analyses, while LASSO would allow for simultaneous consideration of all covariates and for a determination of which covariates contribute most to performance. Therefore, in the present study we demonstrate the first use of LASSO penalized explanatory IRT in an application to a large-scale educational dataset. We describe the various steps involved in executing the technique in applying it to data from the 2011 Dutch national large-scale mathematics assessment, which included 1,619 sixth graders who did 21 multiplication and division items and 107 teachers who filled out an extensive questionnaire on their instructional practices. The effects of the resulting selected instructional covariates are discussed.

H5-2 An alternative to post-hoc model modification in confirmatory factor analysis: The Bayesian Lasso

Junhao Pan, *Sun Yat-sen University, China*

As a commonly used tool for operationalizing measurement models, Confirmatory Factor Analysis (CFA) requires strong assumptions that may lead to poor fit of the model to real data. The post-hoc modification model approach attempts to improve CFA fit through the use of modification indexes for identifying significant residual correlated error terms. We analyzed a 28-item emotion measure collected for n = 169 participants, and the post-hoc modification approach indicated that 86 item-pair errors were significantly correlated. The example showed the challenge in using a modification index, as the error terms must be individually modified as a sequence. Additionally, the modification approach cannot guarantee a positive definite covariance matrix for the error terms. We propose a method that enables the entire residual inverse covariance matrix to be modeled as a sparse positive definite matrix that contains only a few off-diagonal elements bounded away from zero. The method circumvents the problem of having to handle residual correlation terms sequentially. By assigning a Lasso prior to the inverse covariance matrix, this Bayesian method achieves model parsimony as well as an identifiable model. Both simulated and real data sets were analyzed to evaluate the validity, robustness, and practical usefulness of the proposed procedure.

H5-3 A penalized likelihood method for structural equation modeling with ordinal data

Po-Hsien Huang, *National Cheng Kung University, Taiwan*

Penalized Likelihood (PL) has now become an important method for many statistical learning problems. By implementing PL, the relationships among variables may be learned efficiently. Recently, Huang, Chen, & Weng (2014) developed a PL method for Structural Equation Modeling (SEM) and demonstrated its utility through both simulations and real data examples. However, the method is only suitable for SEM with continuous data by model assumption. In this study, we extend the PL method to be applicable to ordinal data. The main idea is to substitute a polychoric correlation estimate for the sample covariance in the PL criterion and hence the original procedure may still be applied. A simulation is conducted to evaluate the empirical performance of the extended PL method. Real data examples are also presented to demonstrate its applicability.

H5-4 Meta-CART: Integrating classification and regression trees into meta-analysis

Xinru Li, *Leiden University, The Netherlands*
Elise Dusseldorp, *Leiden University, The Netherlands*
Jacqueline J. Meulman, *Leiden University, The Netherlands*

Meta-analysis is an important tool to synthesize results from multiple studies in a systematic way. Interaction effects play a central role in assessing conditions under which the relationship between study features and effect size (the outcome variable) changes in strength and/or direction. Within the framework of meta-analysis, when several study features are available, meta-regression lacks sufficient power to detect interactions between them. To overcome this shortcoming, a new approach named "meta-CART" (Dusseldorp et al., 2014) introduced classification and regression trees (CART) in the field of meta-analytic data to identify interactions. The current implementation of meta-CART has its shortcomings: when applying CART, the sample sizes of studies are not taken into account, and the effect size is dichotomized around the median value. In our presentation, we will overcome these shortcomings by 1) weighting the study effect sizes by their accuracy, and 2) using the numerical values of the outcome variable instead of dichotomization. The new methodology will be compared to the current meta-CART in terms of Type I error, power, and recovery performance in a Monte Carlo simulation study. Our initial results are promising, and an extensive simulation study for different population effect sizes and heterogeneity magnitudes will be presented.

H5-5 Detection of treatment-subgroup interactions in clustered datasets

Marjolein Fokkema, *Leiden University, The Netherlands*
Niels Smits, *VU University, The Netherlands*
Henk Kelderman, *VU University, The Netherlands*

Identification of subgroups of patients for which treatment A is more effective than treatment B, and vice versa, is of key importance to the development of personalized psychological medicine. Several tree-based algorithms have been developed for detection of such treatment-subgroup interactions. In many instances, however, datasets may have a clustered structure, where observations are clustered within, for example, research centers or persons. We propose a new algorithm that allows for detection of treatment-subgroup interactions, as well as estimation of cluster-specific random effects. The algorithm uses model-based recursive partitioning to detect treatment-subgroup interactions, and a linear mixed-effects model for estimation of random effects, and can be used for continuous, as well as dichotomous outcomes. In a simulation study, we evaluate the performance of the algorithm, in terms of recovery of treatment-subgroup interactions and prediction of treatment differences, and compare it with that of existing methods. We will provide an illustration by application of the new algorithm to an existing dataset of treatment outcomes.

XH Differential Item Functioning

XH-1 Differential item functioning in Rasch models with recent methods: An application to educational data

Can Gürer, *UMIT- Health and Life Sciences University, Austria*
Gerhard Tutz, *Ludwig Maximilian University of Munich, Germany*

Many procedures have been developed to investigate DIF in questionnaire response data, among others the well-known Likelihood-Ratio test and the Mantel-Haenszel

statistic (see Magis et al. (2010) for a comprehensive overview). Although testing several categorical variables for DIF is straightforward with these methods, investigating continuous covariates for DIF effects has only recently been covered by newly developed methods. Moreover, these new approaches do not face problems of multiple testing as seriously as established methods. In this study, three recent modelling approaches are presented, all of which are capable of handling continuous and multiple covariates at the same time. The models are applied to educational questionnaire data collected from a sample of vocational training students in Germany. The three procedures include (a) the DIF lasso (Tutz & Schauberger, 2015), which utilizes a penalization technique, (b) Rasch trees (Strobl et al., 2013), and (c) item focused trees (Tutz & Berger, 2015), both applying classification and regression tree models but with differing kinds of tests for possible DIF effects. The three approaches are theoretically discussed and contrasted, as well as their results compared.

XH-2 Anchor item selection methods in DIF analysis

Jing Jiang, *Boston College, USA*
Zhushan Li, *Boston College, USA*

Differential Item Functioning (DIF) occurs when individuals from different populations show different probabilities of correctly responding to an item with the same latent trait. If DIF items exist in the assessment, the analysis results of individuals' responses will no longer reflect their true abilities alone. In DIF analysis, examinees in different groups should be placed on a common metric, which is treated as a prerequisite. The common metric consisting of anchor items is supposed to be invariant or DIF-free over groups, so as to produce accurate DIF analysis results. Thus, selecting ideal anchor items becomes important during the process. In the proposed study, several commonly used and newly developed anchor item selection strategies including both non-iterative (e.g., equal-mean-difficulty method, all others as anchors method, and constant-item method), and iterative procedures are summarized and compared. Furthermore, a comprehensive simulation study is conducted to evaluate each method. Especially, we focus on both uniform and non-uniform DIF, when groups differ on discrimination, difficulty, and pseudo-guessing parameters in a three-parameter logistic item response theory model. False alarm rates and hit rates for both uniform and non-uniform DIF are examined. The results show that an appropriate anchor selection strategy plays an important role in DIF analysis. At last, a practical example using a large-scale dataset is illustrated so as to provide visualized results of the optimal anchor selection strategy.

XH-3 A procedure to select more invariant items for the anchor in tests of differential item functioning

Can Shao, *University of Notre Dame, USA*
Quinn Lathrop, *University of Notre Dame, USA*
Ying Cheng, *University of Notre Dame, USA*

Differential Item Functioning (DIF) occurs when items behave differently for examinees at the same ability level that are in different groups. Choosing a proper group of invariant items as anchors is an essential step in the detection of DIF. If the anchor contains less invariant items than it potentially could, the power of DIF detection and other procedures such as testing the underlying cause of DIF is lowered. In this paper, we propose a procedure that aims to select longer and invariant anchors. Instead of just rank ordering the DIF scores and arbitrarily deciding the anchor size, we first construct a distribution of the DIF scores by bootstrapping, and then iteratively identify the candidate anchor by visual inspection or clustering. This procedure is therefore named the Bootstrap, Visualize and Build (BVB) procedure. The BVB procedure can work with any method for DIF detection and it allows practitioners to update the candidate anchor based on visual inspection or clustering instead of arbitrarily deciding the anchor size. A simulation study based on clustering is carried out to compare the performance of the proposed procedure with two existing procedures, All-Others-As-Anchors (AOAA) and One-Item-Anchors (OIA). Results indicate that the BVB procedure outperforms the other two procedures in selecting a longer anchor, with minimal risk of including DIF items in the anchor.

XH-4 Methods to enhance small-sample DIF estimation: Implications for minimum sample size requirements and flagging rules

Xiuyuan Zhang, *College Board, USA*
Amy Hendrickson, *College Board, USA*

The Mantel-Haenszel (MH) procedure is a widely used nonparametric statistical method to measure the amount of DIF present in an item. For operational programs at ETS, at least 200 members are required for the smaller group and at least 500 in total are needed at the test assembly or pilot testing phase. However, in practice, large samples are not always available and small-scale pilot tests are often conducted to provide the first opportunity for test developers to evaluate an item's level of difficulty, discrimination, and functioning among different

subgroups. When DIF analysis is conducted on a small sample of students, special techniques might be necessary to provide more accurate DIF estimation. The present study explored the impact on DIF estimation when applying two adjusting methods in small sample DIF analysis, log-linear smoothing and thick matching of criterion scores. The sample sizes examined were 50/200, 100/300, and 200/500 for the focal/reference groups. The study compared the MH DIF statistic and standardized proportion difference (STD P-DIF) for different conditions to the baseline population DIF statistics in order to evaluate which adjusting method yielded sufficient DIF estimation given small samples. Operationally for this testing program, MH D-DIF is used to flag items with DIF, but our previous study indicated that STD P-DIF might be more sensitive in DIF detection in some scenarios. Therefore, flagging rules would be also reexamined based on the comparison of MH D-DIF and STD P-DIF. Implications for minimum sample size requirements will also be discussed.

XH-5 Testlet effect on IRT-based DIF methods: Bayesian testlet response theory approach

Burhanettin Özdemir, *Hacettepe University, Turkey*

The purpose of this study is to examine the testlet effect on IRT-based DIF methods, since these traditional methods do not take Local Item Dependence (LID) caused by testlets into account. For this purpose, items from the listening and reading sections of the English Proficiency Test (EPT), administered by Hacettepe University in 2011, were analyzed to determine the degree of testlet effect with Bayesian testlet response theory approaches using the MCMC method. DIF values obtained by Lord's Chi-Square and two different Raju's Area methods were compared with respect to gender to test whether these procedures yielded similar results. In addition, results of Lord's Chi-Square, Raju's Signed Area and Unsigned Area methods based on traditional 3PL IRT model were compared to the results based on the 3PL-IRT model, which takes testlet effects into account, in order to determine how the testlet structure of the test effected the results of IRT-based DIF methods. DIF statistics and results for each method without testlet effect and with testlet effect were compared with respect to MAD, RMSD, and number of items detected as DIF. Results indicated that the number of items detected as DIF tended to increase for each DIF method when the testlet effect was taken into account. Overall, Raju's Signed Area method yielded more accurate results with relatively smaller RMSD and MAD values between IRT and IRT model based DIF statistics. On the other hand, Raju's Unsigned Area method yielded highest MAD and RMSD values which indicates that it was largely affected by the testlet structure of the test and yielded less accurate results.

X3 Scoring in IRT

X3-1 Helpful Bayesian inferences about examinee performances for new scale implementations

Michael Chajewski, *College Board, USA*

When new psychometric scales are developed two major operational concerns arise during the transition state associated with its respective implementation: 1) the stability of the scale given operational transition state issues (such as the introduction of new content, form changes, etc.), and 2) the ability to disentangle anomalies in examinee performances from changes in the target population. Previous work by Novick and Jackson (1974), Rubin (1983), and others demonstrated the advantages of using Bayesian inferences to model expected population performances in anticipating steady state score distributions. Iteratively updating prior information about the new examinee population can be leveraged to provide expected conditional scale score proportions. These, in turn, can be evaluated against observed administrations to garner a sense of departure from what the expected steady state distribution should look like. As the number of administrations increases, the prior expectation becomes more credible. Thus, observed deviations from the prior can be attributed to auxiliary agents. The herein proposed research evaluated two possible scenarios using operational transition state examinee data in demonstrating how: 1) a testing program can delineate a screening schedule for the evaluation of the scale score stability, 2) to create a transition state timeline for determining approximate intervals when steady state information about the scale should be known, and 3) how to utilize the aforementioned information in providing ongoing examinee screening for purposes of test integrity.

X3-2 Maximum likelihood score estimation solution for short tests and computerized adaptive tests

Kyung (Chris) Han, *Graduate Management Admission Council, USA*

A critical shortcoming of the Maximum Likelihood Estimation (MLE) method for test score estimation is that it does not work with certain response patterns, including ones that consists only of all 0's or all 1's. This can be problematic in the early stages of Computerized Adaptive

Testing (CAT) administration or when the test length is very short. To overcome this challenge, test practitioners often set lower and upper bounds of theta estimation and truncate the score estimation to be one of these bounds when the log likelihood function fails to yield a peak due to responses consisting only of 0's or 1's. Even so, this MLE with Truncation (MLET) method still cannot handle response patterns in which all harder items are correct and all easy items are incorrect. Bayesian-based estimation methods such as the Modal A Posteriori (MAP) method or the Expected A Posteriori (EAP) can be viable alternatives to MLE. The MAP or EAP methods, however, are known to result in estimates biased toward the center of a prior distribution, resulting in a shrunken score scale. This study proposed a new approach to MLE, called MLE with Fences (MLEF). In MLEF, a couple of imaginary "fence" items with fixed responses are introduced to form a workable log-likelihood function even with abnormal response patterns. The simulation study showed that, unlike MLET, the newly proposed MLEF could handle any response patterns, and unlike both MAP and EAP, resulted in score estimates that did not cause shrinkage of the theta scale.

X3-3 A transition to the theta-based scale for large scale assessments

Tianli Li, *ACT, USA*
Troy Chen, *ACT, USA*

Recently in large scale testing programs, there is an increasing demand for on-going test administrations that allow students to take tests at any time. To meet this demand, the traditional linear form development faces tremendous challenges, and single-item or module based multi-stage adaptive testing is becoming more popular. Consequently, under adaptive testing, each examinee takes a unique set of items and/or modules, and number-correct or percent-correct scoring is no longer comparable or suitable. Instead, a theta scoring method that is based on estimates of the latent trait (i.e., theta) derived under the Item Response Theory (IRT) models is preferred. This study examines various approaches to developing a theta-based scale that matches several psychometric properties of an existing number-correct scale. Those properties include desired statistical moments or scale score distributions and constant conditional standard error of measurement across abilities. In addition to those psychometric properties, this study will also examine scale score reliabilities and classification consistency. To generalize the results, this study will also include several practical conditions such as number of items on a test, number of reported score points and various test subjects. The approaches are evaluated in terms of the degree of their capacity of maintaining the desirable psychometric properties of existing number-correct scale. The empirical evidences from this study will help practitioners select appropriate theta-based scale that fits the particular situation of their transition. It will also help determine how much potential impact such transition will have on students' reported scores.

X3-4 Empirical comparison of classical test theory (CTT) and item response theory (IRT) in change assessment using outcome questionnaire-45 in the Dutch sample

Ruslan Jabrayilov, *Tilburg University, The Netherlands*

Item Response Theory (IRT) is considered as superior to Classical Test Theory (CTT) and its use has been advocated in change assessment and beyond. However, there is no direct comparison of the two theories in terms of their classification of patients based on their change scores after therapy. In this study we compare the two test-scoring methods based on empirical data collected from a Dutch outpatient sample. The questionnaire with which data was collected is the Outcome Questionnaire-45 (OQ-45), a widely-used instrument for assessing psychological distress.

X3-5 Using a test battery for examinee classification: An empirical study

Qing Xie, *The University of Iowa, USA*
Rongchun Zhu, *ACT, USA*
Xiaohong Gao, *ACT, USA*

Scores from test batteries are usually used for making high-stakes decisions for placement, admission, and certification. It is thus vital to develop appropriate composite score and make adequate classification decisions. It is not uncommon to form a composite score from distinct tests designed to measure different but related competences (Kane & Case, 2004). Simply aggregating single test scores into a composite score naturally raises a question on test dimensionality (Cao, 2008). The decision of using a composite score or a profile of single test scores depends on the relationship of the constructs measured by different tests. This study uses empirical data based on a test battery of three multiple choice tests. First, the test battery dimensionality will be assessed via disattenuated correlation coefficients between three pairs of reported scores of different tests, and confirmatory factor analysis comparing one-factor, bi-factor and multiple-factor models. The findings are expected to shed light on how the test scores are related to each other and how to construct a composite score. The second part of the study focuses on classification consistency using specific cut on either composite or the profile of single test scores. Different weights of summing single test scores and methods of setting cut scores will be compared. The implications will be discussed regarding to examinee classification based on a test battery.

X7 Power, Non-Normality & Missing Data

X7-1 Flawed intuitions about power

Marjan Bakker, *Tilburg University, The Netherlands*
Chris Hartgerink, *Tilburg University, The Netherlands*
Jelte M. Wicherts, *Tilburg University, The Netherlands*
Han van der Maas, *University of Amsterdam, The Netherlands*

Because of relatively small effect and sample sizes, many psychological studies are underpowered. Even if researchers understand the importance of well-powered research designs, their intuitions about power might be incorrect. In our first study we surveyed 291 psychological researchers concerning their power intuitions and found large discrepancies between the preferred amount of power and the power calculated based on their typical sample size, effect size, and level of alpha. Almost half of the respondents indicated to base sample size decisions on formal power analyses, but the use of power analyses was unrelated to power intuitions. In our second study we surveyed 214 psychological researchers and asked them to indicate the power, or to estimate the sample size needed to achieve a power of .80, in several research situations. Especially when the underlying effects are small (d = 0.20), respondents overestimated the power and underestimated the sample size needed. This result shows that intuitions about power are flawed and we recommend the reporting of formal power analyses.

X7-2 A Monte Carlo evaluation of parametric and permutation-based methods for two-group multivariate means comparisons when parametric assumptions are violated

Ming Huo, *Northeast Normal University, China*
Patrick Onghena, *KU Leuven, Belgium*

A Monte Carlo study was carried out to examine the performance of two-group multivariate means comparison under the conditions of homogeneous and heterogeneous variance-covariance matrices across two groups and multivariate normal and uniform distributions. The statistical test procedures under evaluation were permutation of raw data values, permutation of composite values derived by standardizing and combining the original values of the dependent variables for each observation, and the parametric Hotelling's T^2 test. Results suggest that the method of permutation of composite values has better performance in terms of Type I error rates and power than the methods of permutation of raw data values and parametric Hotelling's T^2 test under various conditions. With respect to Type I error rates, all three test procedures have inflated Type I error rates with unequal variance-covariance matrices across two groups and this inflation is larger with smaller sample sizes. Unequal variance-covariance matrices across two groups have an impact on the permutation of raw data values method for the most. The permutation of composite values method is most sensitive to the group mean difference. The power of the permutation of composite values method reaches nearly .90 when the group mean difference is .80 and the sample size is 20. On the other hand, the permutation of raw data values method and the parametric Hotelling's T^2 test have almost the same power rates across all the scenarios.

X7-3 A probability-based effect size (AG) robust to multivariate non-normality and heterogeneity of covariance matrices in one-way MANOVA

Johnson Ching-Hong Li, *University of Manitoba, Canada*
Marcello Nesca, *University of Manitoba, Canada*
Rory Waisman, *University of Manitoba, Canada*
Yongtian Cheng, *University of Manitoba, Canada*

This study develops an effect size (ES; A_G) that is robust to violations of data assumptions – multivariate normality and homogeneity of covariance matrices – in Multivariate Analysis Of VAriance (MANOVA) with two independent samples. MANOVA ESs – e.g., generalized eta squared (η_A^2), generalized omega squared (ω_A^2), generalized d (d_G; Steyn & Ellis 2009) – have been proposed in response to the severe flaw in null-hypothesis significance testing, that a large sample size often leads to a significant result ($p < .05$) regardless of the ES. However, these ESs rely on two assumptions: multivariate normality and homogeneity of covariance matrices across groups. Thus, this study develops a robust ES (A_G), a non-parametric estimator for a probability-of-superiority ES (CL_G); this development is based on Ruscio's (2008) CL and A in univariate ANOVA. A simulation study was conducted to evaluate the performance of η_A^2 ,ω_A^2 , d_G, CL_G, and A_G under different conditions: distribution (normal and five non-normal), number of dependent variables (2, 5, 8), d_G (.20, .50, .80,1.50), variance ratio (.25,1, 4), sample size (.50,150), sample-size proportion (.25,.50,.75), and correlation between variables (.5 and .8 for Groups 1 and 2), producing a total of $6 \times 3 \times 4 \times 3 \times 2 \times 3 \times 2 \times 2 = 5,184$ conditions. Each condition was replicated 2,000 times. The results showed that A_G was the most accurate estimator of true ES. The mean absolute percentage bias was 8.87%, which was within the criterion of 10%. However, these biases were 69.98%, 123.80%, 104.34%, and 14.07% for η^2_A ,ω^2_A , d_G, and CL_G respectively, which were substantially larger than that in A_G. The proposed A_G provides a useful method for researchers to report the most appropriate ES even when the key assumptions are violated.

X7-4 Imputation methods for missing non-normal data in structural equation modeling

Fan Jia, *University of Kansas, USA*
Wei Wu, *University of Kansas, USA*

Structural Equation Modeling (SEM) has wide applications in social and behavioral research. When data are multivariate normal and complete, the normal-theory-based estimators, such as Maximum Likelihood (ML) and Generalized Least Squares (GLS), are generally used in SEM because of their desirable asymptotic properties. Violation of multivariate normality may introduce problems in standard error estimates, and produce incorrect fit indices. More problems may arise in the presence of missing data. Modern missing data techniques, such as Full Information Maximum Likelihood (FIML) and Multiple Imputation (MI) are found effective when missing data are normal and ignorable (MCAR and MAR). When missing data are non-normal, robust FIML (Yuan & Bentler, 2000) is the most popular strategy in the SEM literature. However, MI has received much less attention in this context. For example, the standard MI method assumes multivariate normality (MI-MVN). Although MI-MVN was found to be robust to mild non-normality (Demirtas, Freels, & Yucel, 2008), it is not clear whether it yields biased results when data are severely non-normal. Multiple Imputation by Chained Equations (MICE) is also a promising imputation algorithm. It imputes data on a variable-by-variable basis, and can work with semi-parametric or nonparametric techniques, such as Predictive Mean Matching (PMM) and Random Forests (RF), which do not rely on any distributional assumptions. However, little is known about the performance of the MICE methods in SEM. The purpose of this study is to compare the performance of the imputation methods to robust FIML under a broad range of conditions.

X7-5 A hierarchical IRT approach to item nonresponse for Likert-type scales

Chen-Wei Liu, *The Hong Kong Institute of Education, Hong Kong*
Wen-Chung Wang, *The Hong Kong Institute of Education, Hong Kong*

Item NonResponse (INR), such as “Don't Know”, “Refusal”, “Hard to Say”, and “No Opinion”, occur when a respondent does not give a substantive answer to a particular question. Treating INR as missing at random is a common practice, but it could yield biased parameter estimates when they are not. In this study we classified responding processes into a hierarchy and proposed a new Item Response Theory (IRT) model for INR, in which additional latent traits were added to account for the hierarchical structure of responding processes. Simulation studies were conducted to evaluate parameter recovery when INR were ignorable or non-ignorable. The results showed that ignoring non-ignorable INR by fitting standard IRT models yielded severely biased parameter estimates especially when the latent traits were highly correlated; whereas the new model yielded unbiased estimates regardless of whether the INR were ignorable or not. The new model was fit to a real data of citizenship survey about democratic politics. The results demonstrated the superiority and feasibility of the new model for INR for Likert-type scales.

Keynote Speaker: Simon Wood

YH-1 Modeling with smooth functions

Simon Wood, *University of Bath, UK*

Deciding which variables are potential predictors in a regression model is often much easier than deciding how they should enter the model. Generalized additive models and their extensions address this imbalance by allowing models to be specified in terms of smooth functions of predictors, where the functions are then the object of inference. This talk will review the rich variety of smooth model components that can be used to construct such models, and how they can reliably be estimated, including estimation of the degree of smoothing. By taking an empirical Bayesian approach based on reduced rank splines and using Laplace approximate marginal likelihood to estimate smoothing parameters, a quite widely applicable framework can be constructed encompassing random effects, functional predictors and response distributions well beyond the usual single parameter exponential family, including multivariate and location-scale smooth regression models. The talk will cover practical application of this framework and its software implementation in R.

Career Award Speaker: Lawrence Hubert

YH-1 An old tale of two computational approaches to regression: Updated for the 21st century

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

Two different approaches from the early 1900s to the computational task of multiple regression are attributed to a psychometrician, Truman Lee Kelley, and a major political figure, Henry A. Wallace. The Kelley approach would today be called an alternating least-squares strategy; a convergence criterion is set but there is no fixed number of operations. For Wallace, it was the 1925 publication, *Correlation and Machine Calculation*, with George Snedecor that detailed the computational steps for solving the normal equations through Gaussian elimination (or the Doolittle method); the approach was algorithmic with a fixed number of operations. The Kelley iterative alternating least-squares approach is alive and well today in the contemporary psychometric literature in the form of the Kaczmarz-Dykstra iterative projection strategy for solving linear inequality constrained least-squares tasks. It has been used to find and fit unidimensional scales (linear and circular), multidimensional city-block scalings, ultrametric and additive trees, and various other order-constrained structures all fit to given proximity matrices.

Parallel Sessions, Wednesday PM

YH Roger E. Millsap Tribute Symposium: Perspectives on Measurement Invariance [Invited Symposium] Organizers & Chairs: Paul De Boeck & Jelte Wicherts

YH-1 Partial measurement invariance then and now: Methodological aspects and practical implications

Oi-Man Kwok, *Texas A&M University, USA*
Myeongsun Yoon, *Texas A&M University, USA*

Measurement invariance is an important condition for a measure or scale to be meaningfully compared, specifically on the corresponding latent factor mean or the mean of the corresponding observed composite variable, between different groups or populations. However, researchers often find themselves with a measure not fully but partially invariant between the groups they intend to compare. There are many studies on how to conduct the measurement invariance test with typical chi-square difference test or the delta goodness-of-fit statistics. Nevertheless, the "next-step" after confirming violations of measurement invariance is rarely discussed. In this presentation, we will discuss some new perspectives on the methodological issues of partial measurement invariance, along with the practical implication of evaluating partial measurement variance based on the approach proposed by Millsap & Kwok (2004).

YH-2 Testing strong factorial invariance in multilevel data

Suzanne Jak, *National University of Singapore, Singapore*

Within structural equation modeling, the most prevalent model to investigate measurement bias is the multi-group model. Equal factor loadings and intercepts across groups in a multi-group model represent strong factorial invariance across groups. With large numbers of groups, we could treat group as a random variable, and test strong factorial invariance across groups (clusters) in a multilevel structural equation framework. I refer to this test as the test for cluster bias. In the present study, I give an overview of the possibilities with the test for cluster bias, and extend the test for use with three-level data. The proposed method is illustrated with an investigation of strong factorial invariance across school classes and schools in a Dutch dyscalculia test, using three-level structural equation modeling.

YH-3 Score-based tests of measurement invariance

Ting Wang, *University of Missouri, USA*
Edgar C. Merkle, *University of Missouri, USA*
Achim Zeileis, *Universität Innsbruck, Austria*

Measurement invariance is typically assessed via likelihood ratio tests, requiring advance definition of the number of groups, group membership, and offending model parameters. We present a family of recently-proposed measurement invariance tests that are based on the scores of a fitted model (i.e., observation-wise derivatives of the log-likelihood with respect to the model parameters). This family can be used to test for measurement in- variance w.r.t. a continuous auxiliary variable, without pre-specification of subgroups. Moreover, the family can be used when one wishes to test for measurement invariance

w.r.t. an ordinal auxiliary variable, yielding test statistics that are sensitive to violations that are monotonically related to the ordinal variable (and less sensitive to non-monotonic violations). The tests can be viewed as generalizations of the Lagrange multiplier (or score) test and they are especially useful for identifying subgroups of individuals that violate measurement invariance (without pre-specified thresholds) and identifying specific parameters impacted by measurement invariance violations. We illustrate how the tests can be applied in practice in factor-analytic contexts (both traditional factor analysis and item factor analysis) using the R packages "lavaan" for model estimation and "strucchange" for carrying out the tests and visualization of the results. This work was supported by National Science Foundation grant SES-1061334.

YH-4 Identifying models in invariance testing for ordered categorical data

Hao Wu, *Boston College, USA*
Ryne Estabrook, *Northwestern University, USA*

The current practice of invariance testing for ordered categorical outcomes is to first identify a model with only configural invariance and then test the invariance of parameters based on this identified baseline model. This approach is not optimal because different identification conditions on this baseline model identify the scales of latent continuous responses in different ways. Once an invariance condition is imposed on a parameter these identification conditions may become restrictions and define statistically non-equivalent models. In this work we analyze the reparametrizations that leave the observed category probabilities unchanged and give identification conditions for models with invariance of different types of parameters without referring to a specific parametrization of the baseline model. Tests based on this approach have the advantage that they do not depend on the specific identification condition chosen for the baseline model.

YH-5 What we (could) learn from (failures of) measurement invariance

Jelte M. Wicherts, *Tilburg University, The Netherlands*

Tests of Measurement Invariance (MI) grew out of the practical and legal need to assure that cognitive items and tests function equivalently across different pre-existing groups (e.g., sex, ethnicity, nationality). MI tests are now widely considered a crucial step in the validation of scales and tests. Many (if not most) contemporary applications of MI tests are focused on supporting MI without formulating any substantive or psychometric expectations about why MI might fail. I argue that a lot can be learned from failures of MI, specifically when combined with modeling of nuisance factors. I discuss several examples of MI tests applied to cognitive and IQ tests that show that failures of MI can lead not only to practical improvements of standardized tests, but also to a deeper understanding of the nature of latent variables that are supposed to underlie test performance.

Y3 Multilevel IRT

Y3-1 Multilevel item factor analysis models for group-level inferences

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

This study provides a demonstration of the efficiency and comprehensiveness of the Multilevel Item Factor Analysis (MLIFA) framework using Full-Information Maximum-Likelihood (FIML) estimation. MLIFA models are useful in contexts when survey data are used to make inferences at the group-level, such as the measurement of school climate, teacher practice, or the academic proficiency of fourth-graders within a state, but the data consist of individual-level item responses on a set of dichotomous or polytomous items. Large-scale educational assessment researchers have developed a multi-stage analytic procedure that attempts to maximize the efficiency and precision of aggregated scores. I focus on the estimation of a MLIFA model with latent regression, which can be seen as a multilevel extension of the analytical models used in NAEP. This measurement framework allows for simultaneous estimation of item parameters, regression coefficients, and variance components, and can accurately account for multidimensionality, hierarchical nesting of individuals in groups, and a range of background covariates. Standard item response models (2PL, 3PL, nominal, etc.) can be utilized in this framework. I will evaluate the parameter recovery of the MLIFA model with simulated multilevel multidimensional item response data with covariates using the Metropolis-Hastings Robbins-Monro (MH-RM) algorithm implemented in flexMIRT®. under realistic data conditions. Additionally, an empirical demonstration with data from a student survey of instructional practice collected in six large school districts is provided to demonstrate the utility of this model for making classroom-level inferences about dimensions of teacher practice, controlling for classroom composition and other key covariates.

Y3-2 A multilevel cross-classified polytomous item response theory (IRT) model for complex person clustering structures

Chen Li, *University of Maryland, College Park, USA*
Hong Jiao, *University of Maryland, College Park, USA*
Dandan Liao, *University of Maryland, College Park, USA*

This study focuses on investigating native language and school clustering effects on English language learner's English proficiency using a multilevel cross-classified Item Response Theory (IRT) model. For this purpose, a new model will be proposed to deal with the complexity in polytomous item response data from a language test. In this model, item effects are modeled at level-1; level-2 models the person effects, and the clustering effects of school and first language are modeled at level 3 as cross-classified factors. This model is essentially a three level extension of the partial credit model (Masters, 1982). Previous multilevel IRT models only take into account of the clustering effect due to one factor. The proposed model in this study simultaneously models the clustering effects due to two factors that are cross-classified. Applying the proposed model, the clustering effects of first language and school are modeled at a higher level in addition to students' ability. A simulation study will be carried out to simulate item response data reflecting the nature of real data from an English language proficiency test. Multiple estimation methods will be explored including Markov chain Monte Carlo, marginal maximum likelihood, penalized quasi-likelihood, Laplace estimation, and marginal quasi-likelihood methods. Respective software programs such as OpenBugs, SAS, HLM, Mplus and R will be included for comparison of true model parameter recovery under different simulated study conditions. A real data set will be analyzed to demonstrate the application of the proposed model.

Y3-3 Measuring multidimensional growth: A higher-order IRT perspective

Chun Wang, *University of Minnesota, USA*
Steve W Nydick, *Pearson, USA*

Many latent traits in the social sciences display a hierarchical structure, such as intelligence, personality, or cognitive ability. Sophisticated IRT models have been recently proposed to model such a hierarchical latent trait structure. However, these currently available models mainly focus on data collected at a single time point, and few systematic efforts have extended these models to measure individual change (across multiple time points). This study focuses on the development, evaluation, and application of longitudinal extensions of the Higher-Order Item Response Theory (HO-IRT) model using the SEM formulation. Parameter recovery of the longitudinal HO-IRT model was evaluated via a simulation study. Conditions varied include correlation between latent traits (for two time points, $T = 2$) or residual variance (for four time points, $T = 4$), items loading on each dimension, and number of simulees. For each condition, item, person, and growth parameters are compared when using either joint (a.k.a., combined) or separate calibration methods. Necessary transformation equations are proposed for both methods. Simulation studies demonstrate that when $T = 2$, little information is lost when separately estimating item and person parameters at each time point as compared to calibrating the computationally and practically complicated complete longitudinal model. When $T = 4$, joint calibration yields much more accurate parameter recovery than separate calibration. A real data analysis shows the feasibility of determining higher-order and domain-specific growth trajectories. We hope this study provides useful statistical tools for reliably reporting and evaluating individual growth on a general (overall) trait and across several, more specific content domains.

Y3-4 Multilevel item bifactor models with nonnormal latent densities

Ji Seung Yang, *University of Maryland, College Park, USA*
Ji An, *University of Maryland, College Park, USA*

A multilevel item bifactor model is useful when modeling item response data with local dependency (e.g., testlet) that are collected under a multi-stage sampling design (e.g., Programme for International Student Assessment). When only general factor is decomposed into within- and between-level factors, an analytical dimension reduction technique (e.g., Gibbons & Hedeker, 1992) is still beneficial to achieve full information maximum likelihood estimation while avoiding a high dimensional integration of latent trait densities. However, the Gaussian distributional assumption for a general factor may not be reasonable due to some reasons such as the nature of construct and the sampling mechanism of respondents. This study extends the multilevel item bifactor model so that a nonnormal between-subject latent density can be properly modeled using either an empirical histogram (e.g. Mislevy, 1984) or a Ramsey-curve (e.g, Woods & Thissen, 2006; Monroe & Cai, 2014). A simulation study is conducted to evaluate parameter recovery and appropriateness of standard errors compared to the base model that does not take nonnormality of latent density into account. The varying data generation condition includes different shapes of latent densities and sampling conditions. A subset of empirical data set from Early Child Longitudinal Study for Kindergarten is analyzed for illustration.

Y3-5 Extensions of a group-level diagnostic assessment model for practical considerations

Chanho Park, *Keimyung University, South Korea*

Appropriate methodology should be applied when the purpose of assessment is to diagnose groups instead of individuals. Park & Bolt (2008) proposed such a model based on multilevel item response theory. Although the utility of the model was illustrated for tests designed to compare groups such as the Trends in International Mathematics and Science Study, the model needs to be extended for practical considerations. First, the model needs to accommodate test data consisting of polytomous items, since polytomous items are more increasingly used. Second, the model is more useful if longitudinal analyses can be conducted as well as cross-sectional comparisons. Such an extension of the model can be made by designing a method to maintain the base scale across multiple time points. Third, effects of different group sizes on the accuracy of model parameter estimates and whether to apply weights for the parameter estimates will be examined since the group size may differ among the participating groups. The extended model can also be viewed as a special case of a generalized linear mixed model. Theoretical considerations are discussed for each extension, followed by simulation analyses and real data applications.

H6 Multidimensional IRT

H6-1 Application of multidimensional item response theory to personality assessment

Yin Wah Fiona Chan, *University of Cambridge, UK*
Luning Sun, *University of Cambridge, UK*
Baosheng Loe, *University of Cambridge, UK*
David Stillwell, *University of Cambridge, UK*
John Rust, *University of Cambridge, UK*

Multidimensional Item Response Theory (MIRT) has been gaining increasing popularity in educational measurement. As a recently developed methodology, MIRT is of large help with modelling the relationships in a matrix of responses to a set of test items. Few attempts have been made in previous literature to explore the possibility of applying MIRT to personality assessment. In the current study, two personality measures, i.e., the 100-item IPIP proxy to NEO-PI-R (IPIP-100) and the Chinese Personality Assessment Inventory (CPAI) were taken as examples. The datasets consisted of more than 400,000 participants who took part in the myPersonality project and around 3,000 Chinese participants in the original CPAI standardisation sample. A series of MIRT analyses were performed in order to address the following topics: (a) Is MIRT a good method for both dichotomous and polytomous questions? (b) How can MIRT improve personality domain and facet scores? (c) What is the influence of the underlying structure model fit on the validity of the factor scores derived from MIRT? (d) Is the power of MIRT limited by sample size, given a certain test length? (e) Can multidimensional adaptive testing based on MIRT help generate brief measures of personality assessment?

H6-2 The interaction effect of the correlation between dimensions and item discrimination on parameter estimation when multidimensional data are scaled as unidimensional

Derya Çakici Eser, *Kirikkale University, Turkey*
Sakine Göçer Şahin, *Hacettepe University, Turkey*
Selahattin Gelbal, *Hacettepe University, Turkey*

There are some studies in the literature that have considered the impact of modeling multidimensional mixed structured tests as unidimensional. These studies have determined that the error associated with the discrimination parameters increase as the correlation between dimensions increases. In this study, the relation of items' angles on the coordinate system and the correlations between dimensions was investigated when estimating multidimensional tests as unidimensional. Data were simulated based on a two-parameter MIRT model. Angles of items were determined as 0.15, 0.30, 0.45, 0.60 and 0.75, respectively. The correlations between the ability parameters were determined as 0.15, 0.30, 0.45, 0.60, and 0.75, respectively, the same as the angles of the discrimination parameters. The ability distributions were generated from standard normal, positively, and negatively skewed distributions. In total, 75 (5 × 5 × 3) conditions were studied: five different conditions for the correlation between dimensions, five different angles of items, and three different correlations between dimensions. For all conditions, the number of items was fixed at 25 and the sample size was fixed at $n = 2,000$. Item parameter and ability estimation were conducted using BILOG. For each condition, 100 replications were conducted. The RMSE statistic was used to evaluate parameter estimation error, as well as ability estimation error when multidimensional response data is scaled using a unidimensional IRT model.

According to the results of this study, it can be concluded that the pattern of RMSE values, especially for discrimination parameters, are different from the studies in the literature in which multidimensional tests were estimated as unidimensional.

H6-3 Resolving multicollinearity using a two-tier information item factor analysis model

Aiden Loe, *University of Cambridge, UK*
Petar Čolović, *Univdersity of Novisad, Serbia*

The Reinforcement Sensitivity Theory (RST) is an important psychobiological conception in the modern psychology of individual differences (Gray & McNaughton, 2000). The original version of the theory implies that two basic emotional systems (the Behavioral Approach and the Behavioral Inhibition System) exist. The third system, named the Fight/Flight system, was introduced as an explanatory construct for the sensitivity to unconditioned aversive stimuli (Corr, Pickering & Gray, 1995; Corr, 2008). However, further findings of laboratory studies resulted in necessity for major revisions of the RST (Corr, 2008). This led to the restructuring of each of the construct definitions. As most of the revisions were conducted based on laboratory findings, this proved to be a challenging task for psychologists to redefine them into questionnaire items describing human behaviour. A recurring problem is with regards to multicollinearity. A recent study supported that the RSQ-39 version presented good psychometric properties using Confirmatory Factor Analysis (CFA). However, the multicollinearity issue was not resolved. In our study, we propose that a bifactor model may be adopted as a plausible approach. The redefinition of the RST calls for a change in the psychometric properties which otherwise, could not be captured by employing simple structure CFA. Furthermore, simulation studies are often used to compare the applications of different statistical models. Therefore, it is of interest for us to explore the challenges of employing two different forms of bifactor (Item Response Theory (IRT) vs. Structural Equation Modeling (SEM)) models to compare the strengths and weaknesses of each model in real data. Preliminary findings suggest that the bifactor IRT model presents more robust psychometric soundness and encapsulates theoretical concepts better than the bifactor SEM model. The presentation will conclude with implications of the bifactor models and discuss future direction for this work.

H6-4 A dual testlet response theory model that accounts for dual local item dependence

Yong Luo, *National Center for Assessment in Higher Education, Saudi Arabia*
Khaleel A. Al-Harbi, *National Center for Assessment in Higher Education, Saudi Arabia*

Testlet Response Theory (TRT) has been widely used to investigate the issue of Local Item Dependence (LID) in the application of Item Response Theory (IRT). While most studies focus on passage dependence as the source of LID, a recent study (Baghaei & Aryadoust, 2015) finds empirical evidence that non-ignorable LID also exists among items sharing the same response format. When items in a test can be classified into different item bundles based on passage prompt or response format, it is expected that dual LID exist. One example would be the IELTS listening comprehension test, in which there are several audio prompts and items under the same audio prompt have different response formats such as map labeling, multiple choice, or sentence completion. In such a scenario, TRT models may be considered mis-specified due to their failure to account for both sources of LID. To address such dual LID this study proposes a Dual TRT (D-TRT) model. Simulation studies will be conducted to explore parameter estimation of the D-TRT model based on Markov chain Monte Carlo (MCMC) algorithm, and the consequences of fitting a TRT model when the D-TRT model is the true model will also be investigated. In addition, an empirical analysis will be conducted to demonstrate the utility of the D-TRT model.

H6-5 The logistic testlet framework for within-item multidimensional testlet-effect

Peida Zhan, *Beijing Normal University, China*
Xiaomin Li, *The Hong Kong Institute of Education, China*
Wen-Chung Wang, *The Hong Kong Institute of Education, China*
Yufang Bian, *Beijing Normal University, China*

A testlet is a cluster of items that share a common stimulus, and the possible local dependence among items within a testlet is called testlet-effect. Under the framework of Item Response Theory (IRT), various Testlet Response Models (TRM) have been developed to take into account such testlet-effect (e.g., Bradlow, Wainer, & Wang, 1999; Wang & Wilson, 2005; Li, Blot, & Fu, 2006). However, these existing TRM all assume that an item is affected by only one single testlet-effect (Zhan, Wang, Wang, & Li, 2013). Therefore, they are essentially within-item unidimensional testlet-effect models. In practice, multiple testlet effects may simultaneously affect item responses in a testlet. For example, in addition to common stimulus, items can be grouped according to their domains, knowledge units, or item format, such that multiple testlet effects are involved. In essence, an item measures multiple latent traits, in addition to the target latent trait(s) that the test was designed to measure. Existing TRM become inapplicable when multiple testlet effects are involved. To solve this problem, we develop a logistic testlet framework that specifically account for the within-item multidimensional testlet-effect. Results of a series of simulations demonstrated that the parameters of the new models were recovered fairly well by using WinBUGS; and ignoring any one of the multiple testlet effects resulted in a biased estimation of item parameters. Additionally, it did little harm on parameter estimation to fit a more complicated model to data with a simple structure.

H1 Estimation Methods

H1-1 Estimating multi-level models in data streams

Lianne Ippel, *Tilburg University, The Netherlands*
Maurits C. Kaptein, *Radboud University, The Netherlands*
Jeroen K. Vermunt, *Tilburg University, The Netherlands*

Multi-level models have found numerous applications in the social sciences for the analysis of grouped data. Mostly, these models are fit to a static dataset. Recent technological advances in the measurement of social phenomena have however led to data arriving in data streams, i.e. the data enter the data set a point at a time. Traditional methods of fitting multi-level models are ill-suited for the analysis of data streams because of their computational complexity. These methods require all data in memory, which in the case of data streams becomes infeasible quickly, and iteratively go through the data set to estimate the model parameters. In this presentation, we introduce a novel method for estimating random intercept models in data streams and big data. Our fully online method enables computationally efficient estimation of multi-level models. This method does not require all data points in memory and is computationally more efficient as it merely updates summary statistics to estimate the model parameters. We present the results of two simulation studies which show that our method is competitive with traditional methods in terms of the estimation of model parameters. We also present an application in which we applied our novel method to real data regarding individuals' happiness measured in a data stream.

H1-2 Evaluating t-distribution and sandwich robust methods for level-1 outliers in hierarchical linear models

Zhenqiu (Laura) Lu, *University of Georgia, USA*
Jue Wang, *University of Georgia, USA*

The Hierarchical Linear Model (HLM) has become popular and widely used in various educational studies recently. However, violations of normality assumptions can have a non-ignorable impact on the model estimation. Both point estimates and standard errors for fixed and random effects can be sensitive to outliers at all levels. This study aims to investigate and compare two main robust methods to deal with outliers at level 1 in HLM: the sandwich estimator and t-distribution. Previous work indicated that the classical sandwich estimator tends to underestimate the variance of fixed effect estimates. This study examined the robustness of bias-corrected sandwich methods in estimating these variances. Also, by applying a t-distribution to level-1 responses, this study evaluated the performance of a t robust method in bias correction. Simulation studies were conducted using the SAS software for a 2-level HLM. We examined the differences in estimates from both robust methods under different conditions including two types of outliers (3 sd away and 5 sd away), various numbers of outliers (from .04% to 10%), and three levels of sample sizes (200, 1250, 5000). Preliminary results showed that as the bias of parameter estimates increased with the growing magnitudes of outliers, both sandwich robust method and t-distribution method performed well in correcting bias. They can be used interchangeably under certain circumstances. Further, the two approaches also provide unique complementary solutions to each other. A hybrid solution of sandwich robust and t robust methods could be a promising approach.

H1-3 Growth curve modeling for non-normal data: A two-stage robust approach versus a semiparametric Bayesian approach

Xin Tong, *University of Virginia, USA*
Zijun Ke, *Sun Yat-sen University, China*

Growth curve models are often used to investigate growth and change phenomena in the social, behavioral, and educational sciences and are one of the fundamental tools for dealing with longitudinal data. Many studies have demonstrated that normally distributed data in practice are the exception rather than the rule, especially when data are collected longitudinally. Estimating a model without considering the non-normality of data may lead to inefficient or even incorrect parameter estimates. Therefore, robust methods become very important in growth curve modeling. Among the existing robust methods, the two-stage robust approach (Yuan & Zhang, 2012) from the frequentist perspective and the semiparametric Bayesian approach (Tong, 2014) from the Bayesian perspective are promising. The purpose of this study is to compare the performance of the two approaches through a Monte Carlo simulation study for a linear growth curve model, by varying conditions of sample size, number of measurement occasions, population distribution, existence of outliers, covariance between the latent intercept and slope, and variance of measurement errors. Simulation results show that both approaches provide more accurate and precise parameter estimates than traditional growth curve modeling when the normal assumption is violated. The semiparametric Bayesian approach performs better when data come from a mixture of normal distributions. If data are normal, the two approaches estimate the model as well as traditional growth curve modeling. A real-data example based on the analysis of a dataset from the National Longitudinal Survey of Youth 1997 cohort is also provided to illustrate the application of the two robust approaches.

H1-4 Robust moderation analysis using a two-level regression model

Miao Yang, *University of Notre Dame, USA*
Ke-Hai Yuan, *University of Notre Dame, USA*

Moderation analysis has many applications in social and behavioral sciences. Classical estimation methods typically assume that errors are normally distributed and homoscedastic. When these assumptions are not met, the conclusions from a classical moderation analysis can be misleading. A two-level regression model has been proposed to allow for heteroscedasticity. However, existing studies of the two-level regression model have been limited to Normal-distribution based Maximum Likelihood (NML). For more accurate parameter estimates and more powerful tests with real data when conducting moderation analysis using the two-level regression model, this article proposes two robust methods. One is based on maximum likelihood with Student's t distribution and the other is based on M-estimators with Huber-type weights. An algorithm for estimating the parameters of the two-level regression model based on the robust approaches is developed. Formula-based SEs of the parameter estimates are provided. The robust approaches as well as NML are applied to the analysis of a data set from the General Social Survey 2004 to study moderating effects on the relationship between education and income. Results show that the robust approaches identify that gender is a significant moderator while NML does not. The robust approaches are further compared against NML with respect to bias and power through a simulation study. Results confirm that the robust approaches outperform NML under various conditions, including data with outliers, contamination and heavy tails.

H1-5 Some aspects of selecting polychoric instrumental variables for a confirmatory factor analysis model with ordinal data

Shaobo Jin, *Uppsala University, Sweden*

The Polychoric Instrumental Variable (PIV) approach is a recently proposed method to fit a confirmatory factor analysis model with ordinal data. In this paper, we first investigate the effects of using different numbers of Instrumental Variables (IVs). Second, we examine the small sample properties of the specification tests for testing the validity of IVs. Our results show that PIV with fewer IVs produces a lower bias but a higher variation than PIV with all available IVs. Specification tests are extremely oversized at all sample sizes in the study. Possible adjustments are discussed.

H5 Response Processes/Styles

H5-1 Using the asymmetry of item characteristic curves (ICCs) to learn about underlying item response processes

Daniel Bolt, *University of Wisconsin - Madison, USA*
Sora Lee, *University of Wisconsin - Madison, USA*

Several recent unidimensional item response models introduce asymmetric Item Characteristic Curves (ICCs). In this paper we illustrate how different underlying item response processes (e.g., conjunctively versus disjunctively interacting item subprocesses) can provide an explanation for such asymmetries. We show how the application of asymmetric models can in turn be used to inform about the nature of the response process underlying an item. First, we examine the reality of asymmetric ICCs using real item response data for Grades 3-8 from the mathematics sections of the Wisconsin state assessment by applying a 2PNO model with residual heteroscedasticity (Molenaar, 2014). Second, we use simulation analyses to demonstrate the connection between item response process and the form of asymmetry that emerges when fitting such models. We generate responses to items of five different types: (1) five conjunctive subprocess items, (2) two conjunctive subprocess items, (3) single subprocess items, (4) two disjunctive subprocess items, and (5) five disjunctive subprocess items, while controlling for differences in difficulty and discrimination across item type.

We then demonstrate relationships between item type and the estimated δ_1 parameter of Molenaar's model when the model is applied to the data. The relationship highlights both the potential in using asymmetric models in learning more about test items, as well as test conditions under which the occurrence of asymmetric ICCs may be expected.

H5-2 Bayesian estimation in hybrid multidimensional item response model

Kensuke Okada, *Senshu University, Japan*
Shin-ichi Mayekawa, *Tokyo Institute of Technology, Japan*

The hybrid multidimensional item response model is an extension of ordinary multidimensional item response models in which the item response function is represented as a weighted sum of compensatory and noncompensatory item response functions. The existing compensatory and noncompensatory models are now seen as two extremes of a continuum, with numerous variations in between. Using the proposed approach, researchers no longer have to subjectively choose either compensatory or noncompensatory models to apply to data. Instead, for each item, the proportion of each component is estimated from the information provided by the dataset, together with other model parameters. The hybrid multidimensional item response model has three types of parameters: compensatory-part parameters, noncompensatory-part parameters, and the weight parameter. In this study, Bayesian estimation of these model parameters is proposed using the No-U-Turn sampler, which is an adaptive extension of the Hamiltonian Monte Carlo sampler. This algorithm is suited for highly correlated posteriors, which is typically obtained in the estimation of multidimensional item response models. Our simulation study showed that with a suitably large sample size, the proposed method can recover the prespecified model parameters. The performance of model selection is also examined.

H5-3 An investigation of rating scale heterogeneity in ecological momentary assessment (EMA) data

Sien Deng, *University of Wisconsin - Madison, USA*
Daniel Bolt, *University of Wisconsin - Madison, USA*
Stevens Smith, *University of Wisconsin - Madison, USA*
Timothy Baker, *University of Wisconsin - Madison, USA*

In this paper we examine the implications of response style heterogeneity on idiographic analyses applied to Ecological Momentary Assessment (EMA) data. EMA data provide an attractive context in which to consider the effects of response styles due to (1) the greater ability to measure response styles when rating scales are administered repeatedly over time, and (2) the larger anticipated effects of response styles when studying aspects of change within (as opposed to between) individuals. Our paper adapts a multidimensional IRT approach to response style considered by Bolt & Johnson (2009) under the assumption that response style tendencies are constant over time, but where the substantive constructs of interest may change. The effects of response style on several types of idiographic analysis are examined by comparing within-person effects observed with and without response style control. The new model extension is applied to measures of positive and negative affect collected over time among smokers following a quit attempt.

H5-4 Validation and optimization of a new IRT approach for measuring response styles using external variables

Lale Khorramdel-Ameri, *Educational Testing Service, USA*
Matthias von Davier, *Educational Testing Service, USA*

The measurement of noncognitive constructs using rating or Likert-type scales in international large-scale assessment gained in importance but comes not without problems. Response Styles (RS) can occur and harm the validity and comparability of the rating data, especially in low stakes assessments where test-taking motivation might not be high. A new IRT approach (Böckenholt, 2012) and its multidimensional extension (Khorramdel & von Davier, 2014; von Davier & Khorramdel, 2013) seem to be promising in the measurement and correction of RS and have already been tested on personality and large scale assessment data. The current study aims to optimize and validate this extended approach using external variables such as cognitive scores and timing information, as well as mixture IRT models. The examined rating data come from a background questionnaire of an international large scale assessment. The responses to selected questionnaire scales using a 5-point rating scale are decomposed into multiple response sub-processes and modeled through unidimensional and multidimensional IRT models. The advantages and challenges of a unidimensional measurement of RS will be discussed together with the attempt to use external variables for optimization and validation of the IRT approach.

H5-5 Implication of thresholds noninvariance in cross-cultural studies

Yiyun Shou, *The Australian National University, Australia*

Martin Sellbom, *The Australian National University, Australia*
Michael Smithson, *The Australian National University, Australia*
Jin Han, *The Australian National University, Australia*

Differences in intercepts and thresholds between groups imply that the latent means of the factors may not be comparable across the groups. Variability in the thresholds observed in cross-cultural studies, however, may also reveal meaningful cross-cultural differences in response patterns. In this paper, we examined the measurement invariance of two common self-report measures of psychopathy, the Triarchic Psychopathy Measure (TriPM; Patrick, 2010) and the Levenson's Self-report Psychopathy Scale (LSRP; Levenson, Kiehl, & Fitzpatrick, 1995) across Chinese participants and United States participants. In the first study, confirmatory factor analysis with robust WLS estimation revealed that the thresholds of the LSRP (4-point likert scale, from strongly disagree to strongly agree) differed between the Chinese sample and a United States sample. Subsequent mixed ordinal regressions revealed that the Chinese participants were less likely to endorse the responses of extreme opinions than the US participants. In a second study, we compared the response patterns of a Chinese sample and a US sample for the TriPM (4-point likert scale, from True to False). Mixed logistic regressions also revealed that Chinese participants had a strong tendency to avoid endorsing the end points of the response scale. The differences in the response patterns revealed essential cultural differences in thinking styles. We discuss implications for cross-cultural measurement invariance analysis.

XH Cognitive Diagnosis Models III

XH-1 A procedure for assessing the completeness of the Q-matrices of cognitively diagnostic tests

Hans-Friedrich Köhn, *University of Illinois at Urbana-Champaign, USA*

The Q-matrix of a cognitive diagnostic test is said to be complete if it allows for the identification of all possible proficiency classes among examinees. Completeness of the Q-matrix is therefore a key requirement for any cognitively diagnostic test. However, completeness of the Q-matrix is often difficult to establish, especially, for tests with a large number of items involving multiple attributes. As an additional complication, completeness is not an intrinsic property of the Q-matrix, but can only be assessed in reference to a specific Cognitive Diagnosis Model (CDM) supposed to underlie the data, i.e., the Q-matrix of a given test can be complete for one model but incomplete for another. A method is presented for assessing whether a given Q-matrix is complete. The proposed test relies on the theoretical framework of general CDMs and is therefore legitimate for any CDM that can be reparameterized as a general CDM.

XH-2 Validating attribute structures in the attribute hierarchy method for making diagnostic inferences

Ren Liu, *University of Florida, USA*
Anne Corinne Huggins-Manley, *University of Florida, USA*

Classifying examinees at the skill level is a test outcome desired by many educational practitioners, and recent developments adding hierarchical structures to Cognitive Diagnostic Models (CDM) bring about psychometric approaches that provide this type of outcome. While introducing attribute hierarchies in CDMs reduces the size of the Q-matrix and improves the classification accuracy, the misspecification of hierarchy structures could largely compromise the accuracy of diagnosis, which in turn inhibits our ability to provide educational practitioners with the valid information they need at the skill level. This paper introduces the Validated Hierarchy Method (VHM), which improves the Attribute Hierarchy Method (AHM) by detecting and validating the attribute structures specified a priori. In doing so, attributes are specified by comparing full and reduced models using statistical tests before they are analyzed in the AHM. Unlike the original AHM where attribute structures cannot be falsified, VHM provides a way to objectively evaluate structures that may be present in the test. A simulation study is conducted to test if the VHM is able to detect misspecified attribute hierarchy structures to subsequently help frame proper structures in AHM. VHM is also compared with the Hierarchical Diagnostic Classification Models (HDCM) and AHM to evaluate its performance. Results show that the VHM produce more accurate classification results than HDCM and the original AHM. Therefore, the VHM improves both the validity and reliability of classification results from a diagnostic test.

XH-3 Bayesian inferences of the Q-matrix with an unknown number of attributes

Xiang Liu, *Columbia University, USA*
Fei Zhan, *University of North Carolina at Greensboro, USA*

When applying a cognitive diagnosis model, the Q-matrix is typically specified by expert judgement and treated

as fixed. Thus, the elements of the Q-Matrix and the number of attributes included in the Q-matrix are subjective and can be uncertain. In this paper, we propose a fully exploratory approach for DINA (Deterministic Input Noisy "And" Gate) model under a Bayesian framework to perform joint estimation of item parameters and Q-matrix whose dimension is uncertain. The elements of Q-matrix are treated as random variables, while the trans-dimensional estimation applies reversible jump Markov chain Monte Carlo. The effectiveness of the proposed approach is demonstrated through simulations and application to the fraction-subtraction data.

XH-4 Effects of low latent class base rate and high item CTT difficulties on CD-CAT classification accuracy and efficiency under the DINO model

Susu Zhang, *University of Illinois at Urbana-Champaign, USA*

Extensive research (e.g., Xu et al., 2003; Cheng, 2009) has been done on combining Computerized Adaptive Testing (CAT) and Cognitive Diagnosis (CD) to accurately and efficiently identify the attributes/skills individuals have mastered. Templin & Hensen (2006) proposed the Deterministic Input, Noisy "Or" Gate (DINO) model, a CD model applicable to psychopathological assessments and diagnosis. De la Torre (2011) showed that the DINO model belongs to the generalized Deterministic Input, Noisy "And" Gate (DINA) models (Haertel, 1989; Junker & Sijtsma, 2001) that are commonly used in educational assessments, enabling CD-CAT item selection algorithms for educational assessments to be used under the DINO model (Kaplan, de la Torre, & Barrada, 2014). However, little research has looked into the performance of CD-CAT item selection methods in typical psychopathological screenings, where the base rates for pathological latent classes may be low. Most items may have a small proportion of people responding "1" (high CTT difficulties) and Type II error (misclassifying a pathological individual to a nonpathological class) tends to be more severe. Using simulation studies, the current project purports to investigate the effects of low latent class base rates and high CTT item difficulties on the performance of several frequently used CD-CAT item selection algorithms under the DINO model. Proportions of different types of incorrect classifications (e.g., Type I and Type II errors) will be used to examine the classification accuracy at various fixed test length levels. Results from the study will provide additional information for developing more efficient psychological diagnosis using diagnostic classification CAT.

XH-5 Power analysis of item-level interactions in a general diagnostic classification model framework

Yu Bao, *University of Georgia, USA*
Laine Bradshaw, *University of Georgia, USA*

The purpose of educational assessment is often to reliably and efficiently determine what students do and do not understand. Diagnostic Classification Models (DCMs) are a newer class of statistical tools well-suited to fulfill this purpose. DCMs classify students according to mastery levels of latent knowledge components, called attributes. Diagnostic tests are carefully designed where each item elicits one or more attributes. A key feature of the general DCM framework based on the Log-linear Cognitive Diagnosis Model (LCDM) is that attribute behavior can vary across items within the same diagnostic test. The Wald test is a statistical hypothesis test that helps identify appropriate item-level specifications, but its properties have not yet been examined. Using a simulation study, we investigated the statistical power and Type I error rates of the Wald Test under a wide-range of testing conditions. We manipulated four main factors: sample size, effect size, nominal Type I error rate, and test design. For all conditions, we specified the number of attributes measured to be 3, the correlation among each attribute pair to be .70, and the base-rate of mastery for each attribute, which is the proportion of students who are masters of the attribute, constant at .50. Through 500 replications for each condition, our result shows the power to detect non-zero two-way and three-way interactions increases as the effect size of the interaction and sample size increases. Compared to two-way interaction terms, three-way interaction terms require substantially greater sample sizes and stronger effect sizes to yield similar power.

X3 Mediation & Causality

X3-1 Bayesian dynamic mediation analysis

Ying Yuan, *University of Texas MD Anderson Cancer Center, USA*

Most existing methods for mediation analysis assume that mediation is a static, time-invariant process, which overlooks the inherently dynamic nature of many human psychological processes and behavioral activities. In this article, we consider mediation as a dynamic process that continuously changes over time. We propose Bayesian multilevel time-varying coefficient models to describe and estimate such dynamic mediated effects. By taking the nonparametric penalized spline approach, the proposed method is flexible and able to accommodate any shape of the relationship between time and mediated effects. Simulation studies show that the proposed method works well and faithfully reflects the true nature of the mediation process. Our method provides a valuable tool to help researchers obtain a more complete understanding of the dynamic nature of the mediation process underlying psychological and behavioral phenomena.

X3-2 The analysis of mediated moderation using VS

Wai Chan, *The Chinese University of Hong Kong, Hong Kong*
Joyce L.-Y. Kwan, *The Hong Kong Institute of Education, Hong Kong*
Cherry Y.-T. Choi, *The Chinese University of Hong Kong, Hong Kong*

The analysis of moderation and mediation effects has long been an important methodological question in psychological research. Generally speaking, moderation occurs when the relationship between the independent variable (X) and the dependent variable (Y) varies as a function of a third variable (moderator). Mediation, on the other hand, occurs when the effect of X on Y is transmitted through an intervening variable (mediator). In reality, however, the relationship among the variables is often more complicated and a simple moderation and mediation model might fail to describe the underlying psychological process adequately. Consequently, psychologists have started to consider a more general theoretical framework that combines both moderation and mediation into a single model, which is commonly regarded as moderated mediation (moME) or mediated moderation (meMO). In the literature, moME has been well studied and there are specific computer programs for handling it, such as PROCESS (Hayes, 2013). By contrast, meMO has received much less attention due to both statistical and conceptual reasons. Statistically, it is difficult to differentiate moME from meMO because these two processes basically share the same underlying mathematical model. Second, it is difficult to interpret the process meaningfully because meMO concerns the mechanism of a product term, which usually does not carry any substantive meaning. The purpose of the present study, therefore, is to propose an alternative formulation for meMO. This new meMO is intuitively meaningful and it is statistically unique so that one will not get it confused with the traditional moME. Computationally, a new structural equation modeling based statistical tool, VS, is introduced for analyzing this model. Real examples are given to demonstrate how VS can be used to examine meMO, moME, and other models that involve conditional process in general.

X3-3 Estimating the confidence interval of the traditional mediation effect size using the delta method

Baojuan Ye, *Jiangxi Normal University, China*
Qing Zheng, *Jiangxi Normal University, China*
Zhonglin Wen, *South China Normal University, China*

In psychology and some other disciplines of social science, mediation models have been applied in a large number of empirical studies. As with other statistical methods, we need to know the effect size after testing a mediation effect. Since Preacher & Kelley (2011) proposed kappa-squared as a mediation effect size measure, it has become popular in mediation analyses. Unfortunately, the recent work of Wen & Fan (2015) invalidated the kappa-squared as a mediation effect size measure, and should put an end to its use in mediation analysis. Although the traditional mediation effect size P_M (the ratio of the indirect effect to the total effect) is not clear enough to reflect the mediation effect by itself, it is meaningful for a basic mediation model where the indirect effect and the direct effect have the same sign. As is well known, the point estimate contains limited information about a population parameter and does not inform how far it could be from the population parameter. The confidence interval of the parameter provides more information. We derived a formula by using Delta method for computing the standard error of P_M. Based on the standard error, the confidence interval can be obtained easily. We used an example to illustrate how to calculate P_M and its confidence interval by using the proposed Delta method. We also illustrated how to obtain the same result with the Mplus software that automatically calculates the standard error and confidence interval using the Delta method.

X3-4 Using the multilevel comparative interrupted time series method to evaluate the impacts of U.S. national No Child Left Behind policy on school-level decision-making

Jiangang Xia, *University of Nebraska - Lincoln, USA*
Jianping Shen, *Western Michigan University, USA*
Xingyuan Gao, *Western Michigan University, USA*

Although the U.S. federal No Child Left Behind (NCLB) Act has been implemented since 2002, its impacts on school performance and practice have rarely been examined empirically. This lack of empirical analysis was in part because of the difficulty of isolating the effect of NCLB. Based on the Comparative Interrupted Time Series (CITS) method, several recent studies have examined the impacts of NCLB on student achievement, teacher job satisfaction, and school operations. This study extended the CITS method by combining it with the hierarchical linear modeling method. Thus it becomes a multilevel CITS method. Further, this study complements the current studies by examining NCLB's impacts on states', districts', principals', and teachers' influences in school decisions. Using the U.S. nationally representative and longitudinal Schools and Staffing Survey (SASS) data, this study takes advantage of (a) the multilevel CITS method, (b) the richness of the SASS data, and (c) the differences in the presence and strength of prior state accountability systems to isolate NCLB effects. This study will add methodological improvement and empirical evidence to the knowledge base regarding the examination of educational policy's impacts (on school level decision-making). The findings will have important implications for international policy makers, legislators, and educational researchers.

X3-5 Interplay of achievements in mathematics and Chinese language: A six-year longitudinal study

Mo Ching Magdalena Mok, *The Hong Kong Institute of Education, Hong Kong*
Cecilia Law, *Education Bureau, Hong Kong*
Anthony Or, *Education Bureau, Hong Kong*
Jinxin Zhu, *The Hong Kong Institute of Education, Hong Kong*
Jacob Xu, *The Hong Kong Institute of Education, Hong Kong*

The present study aimed to apply multilevel longitudinal cross-lagged analysis to examine reciprocal relationship between achievements in mathematics and Chinese language of students over six years from Primary 3 to Secondary 3. Students' previous achievements in Chinese language and in mathematics were used to predict their achievements in these subjects three and six years later. The project made use of longitudinal data collected by the Education Bureau in 2004 (students in Primary 3), 2007 (students in Primary 6) and 2010 (students in Secondary 3) in the Territory System-wide Assessment (TSA). Participants comprised of 49,526 Primary 3 students in Hong Kong when data were first collected in 2004. These students were tracked for six years, with achievement data collected every three years. Multilevel longitudinal cross-lagged panel analysis of the data allowed the exploration of bidirectional causal relationships between students' achievements in mathematics and Chinese language, while controlling for school contextual variables.

X7 Computer-Based Assessment

X7-1 Item analysis for non-traditional item types in scenario-based assessments

Usama Ali, *Educational Testing Service, USA*
Peter van Rijn, *ETS Global, The Netherlands*

Scenario-based tasks are a relatively new type of educational assessment tasks. Such tasks can include multiple (including traditional and non-traditional) item formats which are linked through a unifying scenario. Scenario-based assessments are considered a next-generation assessment and are intended to strike a balance between authenticity and psychometric quality. Among the methodological considerations for non-traditional item types in these tasks are the choice of statistical models and the unit of analysis. The purpose of our research is to develop a set of procedures for the analysis of non-traditional item types in the context of scenario-based assessments. The main goal of such analyses is to evaluate and enhance the quality of these items. We can distinguish at least two relevant types of analyses given the different item formats. The first type concerns detailed analysis for composite selected-response items. This includes distractor analysis for non-traditional selected-response items for which there are multiple correct responses or items for which students have to select multiple parts. This analysis considers the response accuracy. We also consider a second type of analysis that focuses on kernel smoothing of distractor-level response time. Numerical and graphical summaries will be provided.

X7-2 A preliminary result of collaborative problem solving assessment in Taiwan

Chen-Huei Liao, *National Taichung University of Education, Taiwan*
Bor-Chen Kuo, *National Taichung University of Education, Taiwan*
Shu-Chuan Shih, *National Taichung University of Education, Taiwan*
Cheng-Hsuan Li, *National Taichung University of Education, Taiwan*
Kai-Chih Pai, *National Taichung University of Education, Taiwan*

The purpose of this study is to develop a Collaborative Problem Solving (CPS) assessment to assess students' CPS skills on the Internet. This CPS assessment was designed based on PISA 2015 CPS framework including three major collaborative competencies, four problem solving process and 12 CPS skills. Five CPS units, including four domain contents: reading, science, and math, were constructed and test time was about 100 minutes. In the proposed CPS assessment, students should communicate and co-work with one or more computer agents, and attempt to solve a problem by sharing the understanding and make the correct decision and solution. Participants are 82,716 grade 9 and 10 students (44,448 boys and 38,268 girls) in Taiwan. The preliminary results of CPS assessment indicated that students showed higher level of performance in establishing and maintaining shared understanding, and the lower level of performance i) in taking appropriate action to solve the problem, ii) in planning and executing, and iii) in monitoring and reflecting.

X7-3 Sequential learning detection with response times

Sangbeak Ye, *University of Illinois at Urbana-Champaign, USA*

The problem of detecting learning in latent class modeling for cognitive diagnosis is considered. Our objective is to give an examinee a sequence of items requiring one or more unlearned attributes until sufficient evidence of learning has been displayed. This can be viewed as a sequential change point detection problem and we seek to minimize the delay between when learning takes place and when learning has been determined. The intended applications are in intelligent tutoring systems in which many attributes are acquired by the end of the assessments. Computerized adaptive tests afford the chance to record response times and we utilize these response times along with response accuracy for timely detection of learning. A joint item response and response time model is proposed with a latent attribute vector for skill mastery, and a latent speed parameter for each person. The attribute vector is assumed to change as an examinee learns, but the speed parameter remains constant, though responses are assumed to become faster as an examinee learns due to terms in the response time distribution related to knowledge states. Simulations are conducted to examine the properties of the CUSUM statistic for learning detection, both with and without response times. Methods for estimating the time-point of learning are proposed and we examine the validity of these methods in assessing individual learning rates.

Early Career Award Speaker: Sy-Miin Chow

YH-1 Fitting and evaluating dynamical systems models: Current techniques and unresolved challenges

Sy-Miin Chow, *Pennsylvania State University, USA*

Dynamical systems are systems that change over time such that their current states are somehow dependent upon their previous states. Change concepts described in most dynamical systems models are by no means novel to social and behavioral scientists, but most applications of dynamic modeling techniques in these disciplines are grounded on linear theories of change. In this talk, I will illustrate the potential utility of nonlinear dynamical systems models using empirical examples drawn from the behavioral sciences. Current frequentist and Bayesian techniques for performing exploration, estimation and diagnostics of nonlinear dynamical systems models, as well as challenges and unresolved issues in utilizing these techniques, will be reviewed.

Presidential Address: Sophia Rabe-Hesketh

YH-1 Ignoring non-ignorable missingness

Sophia Rabe-Hesketh, *University of California, Berkeley, USA*

In longitudinal data, maximum likelihood estimators of mixed-effects model parameters are consistent if missingness depends only on the covariates. Missingness can also depend on observed outcomes, under correct specification of the covariance structure, but this result is useful only under monotone missingness. When missingness depends on unobserved outcomes or on the random effects, it is said to be Not Missing At Random (NMAR). For such NMAR missingness, joint modeling of the outcomes and missingness has been advocated, but these approaches are known to rely on unverifiable assumptions. In this talk, I will consider methods that ignore NMAR missingness but are consistent for (some of) the parameters of interest. An example of such "protective" estimators are conditional maximum likelihood estimators, also known as fixed-effects approaches. For binary data, such approaches can be used to obtain consistent estimators of regression coefficients under a wide range of NMAR mechanisms (Skrondal & Rabe-Hesketh, 2014, *Biometrika* 101, 175-188). Other protective estimators for binary and continuous outcomes will be considered in this talk.

Author Index