

Contributing Presentations

AAP1

#1

AAP

Howard Wainer
Harris L. Zwerling

*National Board of Medical Examiners
Pennsylvania State Education Association*

Logical and empirical evidence that smaller schools do not improve student achievement

Proponents of the "Smaller Schools Movement" have cited extensive evidence to support the inference that small schools improve the social atmosphere and academic performance for students. One aspect of the evidence cited is what appears to be the overabundance of small schools among the highest achieving schools on academic achievement tests. We present evidence that such inferences are incorrect, yielded by faulty logic and the omission of important statistical information. We support our technical argument with test data from Pennsylvania's public schools.

#16

AAP

Angela Oktavia Suryani

Atma Jaya Catholic University

Validity and Reliability testing for academic potential test at Atma Jaya Catholic University, Jakarta, Indonesia

It is a common that universities, colleges, and other higher educational institutions hold an entrance test to select prospective students. In order to get qualified students, we need to be sure that the instruments used in the test are valid and reliable. The aim of this research was to investigate the validity and reliability of the academic potential test used by Atma Jaya Catholic University, Jakarta, Indonesia. There were three kind of instruments with two parallels for each kind (total 6 instruments) with 1941 number of samples. Item Response Theory (IRT) - 2 parameter and factor analysis were used to examine the construct validity of the test. Multiple regression was used to predict the student's GPA in second semester, t-test was used to examine whether the item difficulty and item discrimination were equal for each parallel (alternate form reliability). The result showed that the instruments were content with easy and moderate item difficulty, valid to measure the construct (numeric & verbal reasoning, and basic english), and equal for each parallel form. The predictions analysis showed different result when tested at different faculty.

#28

AAP

Delphine Gross
Michael Eid
Fridtjof Nussbeck
Christian Geiser
David Cole

*University of Geneva
University of Geneva
University of Geneva
University of Geneva
Vanderbilt University*

How to account for several methods of measurement in latent state-trait models:
Interpretation of an empirical application

Multitrait-multimethod models (MTMM) and latent state-trait models (LSTM) are widely applied multidimensional models to represent different sources of individual differences. Whereas MTMM models have been developed to separate trait from method-specific influences, LSTM have been defined to separate trait from occasion-specific influences in longitudinal studies. The integration of the two approaches is still missing. It is shown how an MTMM-LST model can be defined that allows analyzing the convergent and discriminant validity on the trait as well as on the occasion-specific level. This model also allows the quantification of stable and variable method effects as well as the stability and change of states and traits. This model, which is based on the CTC (M-1) modeling idea, will be illustrated by an empirical example analyzing children's depression and anxiety measured by self- and teacher-ratings on four waves of measurement.

AAP1

#51

AAP

Samantha Bouwmeester

*Erasmus University*Modern techniques for modelling latent variables of cognitive processes

Wohlwill discussed the questions to be asked and the methods to be used when studying cognitive development. He explained that methods used were suited primarily for analyzing data collected in an experimental context and often inappropriate for studying developmental change. According to Wohlwill, developmental psychology requires a differential approach in which changes in behavior are described using response patterns with the emphasis on individual differences. Modern test theory or item response theory has grown substantially over the past decades, now offering appropriate and sophisticated analysis methods to handle differential questions. In this presentation I compare the use of classical and modern techniques in the context of transitive reasoning. Brainerd and Kingma elaborated fuzzy trace theory but used an experimental design to test different aspects of the theory. The availability of new and advanced statistical methods enabled us to analyze response patterns and predict the responses processes on different kinds of transitive reasoning tasks. The recently developed multi-level latent class model is a sophisticated and powerful tool for testing the hypothesized structure of the theoretical model and for describing the development of transitive reasoning at a detailed level of analysis.

AAP2

#56

AAP

Anke Weekers

*University of Twente*The use of item response theory in personality assessment

Scale construction and revision in the field of personality assessment relies heavily on classical test theory. Item response theory (IRT) offers new solutions that classical test theory does not provide. For the application of IRT in the personality domain the psychometric structure of personality scales has to be clear. Therefore, it is important to investigate which IRT models describe personality data best. Most studies done so far used dominance IRT models like the 2-, 3- parameter logistic models to describe personality data. A recent development is to analyze personality data using unimodal IRT models with a single peaked response curve. Reviewing the literature, however, it is unclear when dominance and single peaked IRT models can best be used to describe personality data. In the present study, the fit of the response curves will be investigated for a broad range of IRT models, including both parametric and nonparametric dominance and single peaked models using data from unidimensional Dutch personality scales. The ultimate aim is to obtain more insight into the structure of personality data.

#70

AAP

Jang Schiltz

*University of Luxembourg*Developing and Evaluating a Computerized Adaptive Mathematical Test

The theoretical basis underlying the construction of a CAT is summed up and the item response theory explained. In comparison with a classical paper-and-pencil test, the characteristics and the advantages of a CAT are exposed. After performing a thorough analysis of the data, we developed a CAT based on the responses of 3590 children to a standardized mathematical test. The software presents a great amount of flexibility and allows treating multiple-choice items, as well as items with free response options and graphical items. Some results of an empirical evaluating study are presented. We showed that qualities of power, flexibility and economy of time of the CAT do not imply the loss of the classical psychometric properties of a paper-and-pencil test; we analyze the carry-over effect between the computerized and the paper-and-pencil test situation and as well as the relations between mathematical skills and general cognitive abilities.

AAP2

#95

AAP

Andreas Frey
 Claus H. Carstensen
 Johannes Hartig

Leibniz Institute for Science Education (IPN)
Leibniz Institute for Science Education (IPN)
German Institute for International Educational Research (DIPF)

BIB-Designs in large scale assessments

Typically, in large scale assessments several hundred items are used to measure one or more subject areas. Administering all items to each participant would mostly be too time consuming. Therefore, participants are given booklets containing a small number of item clusters. Each participant is presented a subset of the items. If clusters are sufficiently linked within the booklets, item difficulties and person abilities can still be estimated within IRT models. Typically, the assignment of clusters to booklets is done manually. However, under complex conditions (many subject areas with unequal numbers of clusters, position effects etc.) this may not be feasible. This paper demonstrates the use of balanced incomplete block designs (BIBDs) for systematic booklet composition. Depending on the model used to generate a BIBD, the resulting design can be optimized for different criteria. The procedure is illustrated by the results of a simulation study. BIBDs based on different models allow the detection of item misfit as well as the estimation of covariances between subject areas even if systematic position effects and autocorrelated errors within booklets are introduced. It is discussed how the advantages of BIBDs can be extended if single items instead of clusters are systematically assigned to booklets.

#97

AAP

Marc A. Tomiuk
 Ursula Hess
 Chankon Kim
 Heungsun Hwang
 Daniel Durand

HEC Montréal
Université du Québec à Montréal (UQAM)
St-Mary's University
HEC Montréal
HEC Montréal

Development of a Preliminary Measure of the Intrapersonal Aspects of the Food Enjoyment Construct

Clearly, food consumption involves instrumental and hedonic components; and the enjoyment of food can involve much more than ingestion and the activation of sensory receptors. In fact, based on a review of the literature and qualitative interviews, the domain of the construct *Food Enjoyment* is vast and it seems to effectively be captured by six second-order dimensions. In turn, each of these "big" dimensions is comprised of first-order factors or facets. An empirical study was conducted. Its objectives were to: (1) provide a focus on the second-order dimension labeled *Intrapersonal Aspects of Food Enjoyment* along with its underlying facets and (2) develop a preliminary measure of each facet of this dimension. Measure development and purification procedures essentially involved the steps prescribed in Churchill. Exploratory factor analyses revealed the following eight facets: (a) anticipation of food intake; (b) health outcome and food safety assurances; (c) food as a source of gratification; (d) solitary food consumption; (e) food as a means of affect regulation; (f) epicurean delight; (g) feasting/overeating; and (h) discovery of exotic foods. The final structure accounted for 81.9% of the total variance in the data. In the case of each composite, the coefficient Alpha estimate was greater than .9.

AAP3

#44

AAP, OTR

Gad Saad
Richard Sejean

Concordia University
Concordia University

A Process-Tracing Interface for Studying Attribute-Based and Alternative-Based Sequential Sampling

Within the behavioral literature, sequential sampling has been studied in two distinct manners. In attribute-based sequential-sampling, individuals iteratively acquire information on the two competing alternatives until a desired level of cumulative discrimination is achieved to identify a winning option. In alternative-based sequential-sampling (e. g. , the Secretary Problem), individuals iteratively sample one alternative at a time until a satisfactory option is identified. In other words, while in the former form of sampling, the iterative counter is over the number of attributes acquired with the number of options set at two, in the latter case, the counter is over the number of options that will be holistically evaluated. Some scholars have suggested that these two processes can be combined, yielding one grand sequential sampling framework. As part of a Master's thesis exploring the genetic underpinnings of decision-making styles, a process-tracing interface was developed to investigate this "mega" sequential-sampling framework. In the current presentation, we shall introduce the framework, provide a demonstration of the interface, and if time permitting briefly discuss some collected data.

#130

AAP

Verena D. Schmittmann
Conor V. Dolan
Han L. J. van der Maas
Maartje E. J. Raijmakers

University of Amsterdam
University of Amsterdam
University of Amsterdam
University of Amsterdam

Latent Markov models with covariates applied to learning processes

Mathematical learning models with time-constant and time-varying covariates are used to investigate the dynamics and the development of category-learning in a cross-sectional sample of children (6-16 years). The learning process is modeled with a latent Markov model. A mixture approach allows the distinction of different learning processes or strategies. The inclusion of time dependent covariates allows to model changes during the learning process of an individual, while time-constant covariates may model the relation of the learning process with cognitive resources such as working memory and learning efficiency. The complete model is specified as a Bayesian hierarchical model. The models are fitted using Markov Chain Monte Carlo simulation in WinBUGS to simulated data, and to the trial-by-trial data of children performing a discrimination learning experiment.

#144

AAP

Jimmy de la Torre

Rutgers, The State University of New Jersey

Skills Profile Comparisons at the State Level:
An Application and Extension of Cognitive Diagnosis Modeling in NAEP

Traditional analyses of NAEP have relied on item response models that utilize a single continuous and broadly defined latent trait. By measuring a unidimensional trait, analyses using these models aggregate the multiple correlated traits measured by the tests into a single dominant trait. Although the unidimensional latent trait of standard item response models accounts for much of the dependence in the items, it also ignores the rich information in the tests that may have diagnostic value. To take advantage of the available information in tests, this paper uses a cognitive diagnosis approach in analyzing NAEP to unravel the inherent multidimensionality of NAEP tests. This more efficient utilization of test information allows for skills to be reported with finer granularity. The paper analyzes in detail the data collected from Texas, Michigan, and Maine for the 2003 NAEP Grade 8 Mathematics assessment using the higher-order DINA model. The three states are compared in terms of their skill profiles, and their relative strengths and weaknesses with respect to the skills measured by the test. In addition to the response data, background information about the students collected during the test administration is also utilized in estimating the model parameters. The Q -matrix developed by DeBello et al. for the same data is used in the analysis.

AAP3

#169

APP

Rien van der Leeden

*Leiden University*Studying immediate and long-term treatment effects in therapy evaluation research using multilevel models

In clinical research, treatments intended to reduce, for instance, depression or hypochondriacal complaints are often studied using experimental designs. Commonly subjects are assigned to different treatment groups which are repeatedly measured over time. Usually a pre-treatment measure is obtained, and after the experimental treatment, for instance, a period of using medication or a therapy program, a post-treatment measure is used to assess the immediate effect on the dependent variable. Long-term effects are often studied by a (usually short) series of measurements over time. The analysis of this kind of data imposes a few problems, including an increasing number of missing observations over time, a small number of time points unequally spaced in time, a large "drop" in the dependent variable due to the immediate effect of the treatment, and a relatively "flat" developmental trajectory related to the long-term effect. Due to these data characteristics, traditional ANOVA repeated measures analysis does not seem to be the appropriate method. Instead, data of this kind call for multilevel analysis. In this paper, we study and compare several ways of dealing with the aforementioned problems using longitudinal applications of multilevel analysis.

BSI1

#7

BSI

Paula Fariña
 Fernando Quintana
 Ernesto San Martín
 Yuri Goebeur

Pontificia Universidad Católica de Chile
Pontificia Universidad Católica de Chile
Pontificia Universidad Católica de Chile
University of Southern Denmark

A Non Parametric Bayesian Analysis of a large Educational Data Set

The Chilean government annually implements a system to measure the quality of national education called SIMCE. It consists of a national test applied to the totality of students in a certain grade. Data from the 2001 SIMCE Mathematics test are available. The exam was applied to 191. 441 10th grade students. Each examinee had to answered 48 multiple choice items. One of three different test forms were randomly assigned to each student. Some categorical covariates are also available such as gender, type of school (private or public), educational level of parents, number of books at home, possession of PC (yes-no), access to internet and socioeconomic group. The test was completely answered by 79 % of the students. A Rasch model for binary outcomes is introduced to estimate examinees traits. We adopt a Non-Parametric Bayesian perspective assuming traits to be modelled as Mixtures of Dirichlet processes. This approach allows us more flexibility in the estimation of abilities. The novelty in this work is the introduction of covariates in order to analyze how individual's socioeconomic information impacts traits.

#59

BSI

Daniel M. Rice

*Maritz Research*Logit Regression with a Very Small Sample Size and a Very Large Number of Correlated Independent Variables

Multicollinearity is one of the most fundamental problems with psychometric data in Marketing Science. At last year's Psychometric Society meeting, I presented preliminary data on a new technique that appears to reduce the effects of multicollinearity in Logit Regression dramatically. This year's talk will present new findings over the past year. This new technique modifies Standard Maximum Likelihood Logit Regression to include constraints that embody the expected properties of sampling error. This approach is different from more standard regularization because it does not employ smoothing, so it gives significant attribute-level resolution. In highly multicollinear psychometric datasets to date, the solutions are valid and reliable even when independent variables significantly exceed responses in number. Such results will be presented and comparisons will be made with more standard predictive modeling approaches. All of my results to date support the view that misestimates due to multicollinearity are simply the manifestation of sampling error; the removal of sampling error seems to solve the problem. This new approach has some formal similarities to Zellner's Bayesian Method of Moments (BMOM), although these sampling error constraints and the applications to Logit Regression are new.

BSI1

#118

BSI

Ru Lu
Chin-fang Weng
Tiandong Li

University of Maryland at College Park
University of Maryland at College Park
WESTAT, Inc

A Bayesian Approach to the Calibration of New Test Tasks in the
Latent Class Model Design with Correlated Latent Variables

Within the framework of probability-based evidential reasoning, new test items can be calibrated into the same scale as the old test items. The focus of this study is on inference about new tasks in assessment programs with different "domain-behavior" structures in a latent class model. We examine and compare the precision of estimation on item parameters. Monte Carlo Markov Chain (MCMC) is applied to estimate the posterior distribution in inferences. The study simulates 2000 replicates under a binary latent class model with 4 independent latent variables and 16 old test items. The results show that the number of old items is more important than the "Domain-behavior" structure in calibrating new items. During calibration, the items with higher discrimination produce better results. Our ongoing study furthers the research by correlating the latent class variables to make the results of the study more applicable.

#141

BSI

Ed Merkle

Wichita State University

Bayesian methods for incomplete, multivariate data

Many general-purpose missing data methods are comprised of the following steps: first obtaining complete data in some way (for example, by deletion or imputation) and then estimating the model of interest via a standard method. These two steps can be combined via a Bayesian sampling algorithm, making the "data completion" and "model estimation" steps interdependent on one another. This interdependence is advantageous because, at a given iteration in the algorithm, the sampling of missing data can be influenced by the sampled model parameters (and vice versa). As compared to methods in which a model is estimated using a small number of completed data sets, this sampling technique results in more precise parameter estimates. In this talk, two Bayesian methods for missing data are described. Based on the concepts of data augmentation and multiple imputation, the methods differ in whether or not the statistical model of interest plays a role in completing the data. The methods will be applied to confirmatory factor analysis, and a data example will be presented to illustrate and compare the methods.

BSI2

#129

BSI, OTR

Kazuo Shigemasu
Takahiro Hoshino
Kerry L. Jang

The University of Tokyo
The University of Tokyo
University of British Columbia

Bayesian general multivariate behavioral analysis without additive polygenic hypothesis

Multivariate Behavioral Genetic analysis is a useful tool to estimate separately the degree to which multiple phenotypes share a common etiology. When the interest is in the latent constructs such as intelligence, psychological disorders, factor analytic multivariate analysis of twin data is desirable, and Bayesian approach has some advantages over traditional methods in that it enables precise inferences about parameters and the heredity index (function of parameters) by means of their posterior distributions. In this report, we propose a general behavioral genetic model in which the multiple addition hypotheses of genes is discarded by introducing a new parameter which expresses the influence of interaction of plural genes. Also, we show the results of the simulation study and real data analysis using the MCMC technique.

BSI2

#137

BSI, IRT

Eric Loken
Kelly Rulison*The Pennsylvania State University*
The Pennsylvania State University"Low-stakes" Learning and Assessment with IRT

Computer adaptive methods are routinely employed for "high-stakes" testing applications such as admissions and licensing tests. The same methods can be used to provide practice questions for students in less controlled settings. We show how questions can be deployed to provide adaptive practice for students while also yielding accurate assessment. Under "low-stakes" conditions, the joint estimation of item and person parameters may be more challenging as some students may only answer a few questions, and as the questions may be delivered over a narrow range of ability. Using Markov chain Monte Carlo methods to simulate the posterior distribution for item parameters after an initial pilot phase, accurate estimates for the three-parameter logistic model are possible. We also introduce a fourth parameter to the model to account for student carelessness in a low-stakes setting. The model works well in a simulation of students answering practice items within an academic domain. We then apply the model to data gathered on students preparing to take the new SAT which began in March 2005.

#157

BSI

Shin-ichi Mayekawa

*Tokyo Institute of Technology*Educational Tools for Introductory Bayesian Statistics

A system of tools for teaching introductory Bayesian statistics using computer was developed. The main body of the tools is the distribution identifier, which receives an expression and a random variable name as the input and identifies the distribution of the variable whose density function is proportional to the input expression. When the input distribution is identified, it returns the name of the distribution, the parameters of the distribution, and the result of the integration of the expression with respect to the variable. Therefore, if we input the product of the likelihood function and the prior density, it identifies the posterior distribution of the specified random variable, and returns the marginal density of the rest of the variables. The tools cover the standard Bayesian conjugate analysis including binomial-beta, Poisson-gamma and univariate and multivariate normal distributions.

#160

BSI, OTR

Anzalee Khan
Jean-Pierre Lindenmayer*Fordham University*
New York University,
Nathan Kline Institute for Psychiatric Research,
& Manhattan Psychiatric Center
Nathan Kline Institute for Psychiatric Research

Mohan Parak

A Bayesian State-space model for the rate of nonresponse to treatment that lead to continued hospitalization:
Analysis of variables of the Positive and Negative Symptoms Scale (PANSS) and treatment regimen

PANSS psychometric properties have been demonstrated in trials of schizophrenics to estimate changes in symptomatology. The interrelated response (symptoms, symptom domain) to discharge from a psychiatric facility is complex and considered in the state-space models. Initial work on developing a robust method for predicting continued hospitalization based on symptomatology is presented. Our state-space model demonstrates execution using Bayesian approach that affords efficient computation and prior information. We consider continued hospitalization to be predicted at a series of times $t = 1, 2, \dots$ to T , with error giving rise to a measurement equation: $O_t = N_t(\hat{u}_t, ct)$ where O_t represents the predicted next course of symptomatology and \hat{u}_t is the actual next symptom that prevents discharge. If predicted discharge decision of a patient depends upon the actual symptom exhibited, process noise, \hat{A}_t (i. e. , mutually independent of \hat{u}_t), and a vector of parameters describing the course of symptomatology, then $\hat{u}_t = f_t(\hat{u}_{t-1}, \hat{A}_t; \lambda)$. The state space model estimates the future outcomes, \hat{u}_t , using proposed equations and estimates parameters describing patient's course. A Bayesian meta-analytic tool is used for merging information from multiple observations of patient paths to facilitate broad inference to probable paths of future patients.

BSI2

#174

BSI

Geoffrey J. Iverson

*University of California, Irvine*Prep, p-values and Bayesian Inference

Psychological Science, the flagship journal of the Association for Psychological Science, recently advised contributors to accompany estimates of experimental effects by Prep, a measure of replicability of the sign of an effect, and to do so at the expense of the traditional p-value. However Prep is merely a proxy for the usual p-value, in its construction and in its suggested use as a decision statistic. A more telling criticism of Prep is that, being conditional on the size of an observed effect, it cannot possibly deliver on its promise to predict effect sizes obtained in replicate substantive experiments.

There is good news in all of this. Once distractions like Prep are ignored, attention can be more usefully focused on the deficiencies of evidentiary measures such as the traditional p-value and what might replace them. Bayesian methods seem especially promising in this last respect. It is curious that Prep results from a Bayesian calculation, though not one that most Bayesians would usually contemplate.

CCC

#61

CCC

Brandon Vaughn
Qiu Wang*Florida State University*
*Florida State University*Classification Based on Hierarchical Linear Models

Many areas in educational and psychological research involve the use of classification statistical analysis. For example, school districts might be interested in attaining variables that provide optimal prediction of school dropouts. In psychology, a researcher might be interested in the classification of a subject into a particular psychological construct. The purpose of this study is to investigate alternative procedures to classification other than the use of discriminant and logistic regression analysis. A classification rule utilizing hierarchical linear modeling (HLM) will be derived and examined, with a following example which will show the benefit for using such an approach by comparing the hit rates to those of a logistic regression analysis. A Monte Carlo simulation study is considered to compare a multi-level classification procedure to typical methods. Finally, application of the rule to a real data set will be conducted.

#153

CCC

Kensuke ISOMURA
Shin-ichi MAYEKAWA
Takahiro HOSHINO*Tokyo Institute of Technology*
Tokyo Institute of Technology
*The University of Tokyo*Latent Class Analysis of Three Mode Data

Consider the case where N subjects rate n stimuli on p attributes scales. The resulting $N(n \times p)$ matrices (three way matrix) can be analyzed in many ways including three mode factor analysis. In this study, we assume that the subjects can be classified into q latent classes and that the individual differences are expressed in terms of the differences among the factor patterns and the unique variances which characterize the latent classes. The simulation study showed that the method can recover the true configuration, and the marketing application showed that the most parsimonious models were chosen among several hierarchical factor analytic models such as the invariant factor models and the PARAFAC model.

CCC

#154

ccc

Tomoya Okubo
Shin-ichi Mayekawa

Tokyo Institute of Technology
Tokyo Institute of Technology

Some extensions of MDPREF by Latent Class Modelling

Okubo and Mayekawa proposed MXPREF-1 which is a latent class extension of MDPREF for the paired comparison data where the subjects are treated as the observations to be classified. In this study, we developed MXPREF-2, where the preference judgments for each pair are treated as the observations to be classified. The implication of this model is that the subjects have several uni-dimensional preference scales (view points) and that they use different view point for the paired comparisons. Simulation study showed that our method can recover the true configuration, and the analysis of the real data confirms that the subjects actually shift the view points during the paired comparison session.

#162

ccc

Michael J. Brusco
Hans-Friedrich Köhn

Florida State University
University of Illinois at Urbana-Champaign

Optimal Partitioning of a Data Set Based on the p-Median Model

The p-median problem represents an NP-complete clustering problem, requiring the selection of p objects to serve as cluster centers such that the sum of the Euclidean distances (or some other dissimilarity measure) of the remaining objects assigned to each center is minimized. Lagrangian relaxation can provide tight bounds for the p-median problem; when used in conjunction with heuristics and/or branch-and-bound methods, globally-optimal solutions for clustering problems of non-trivial size can be obtained. Computational results for the p-median problem are pervasive in the operations research literature. However, these findings mostly focus on problem structures not necessarily generalizable to partition tasks involving pairwise dissimilarity values. We evaluate p-median problems involving dissimilarity values based either on Euclidean or squared Euclidean distances (with some data sets measured in more than two dimensions). Computational results are reported for 12 published data sets, ranging in size from 59 to 724 objects. Optimal p-median solutions were obtained for all 12 data sets for p equal two to ten. In addition, we present an application of the p-median model to a 'real' data set (N=617) on substance abuse.

CDA

#20

CDA, CCC

Marian Hickendorff
Willem J. Heiser
Cornelis M. Van Putten
Norman D. Verhelst

Leiden University
Leiden University
Leiden University
CITO

Clustering Nominal Data with Equivalent Categories

Two techniques are discussed for clustering data of nominal measurement level, where the categories of the variables are equivalent (the variables are replications). One technique is GROUPALS, an algorithm for the simultaneous scaling (by multiple correspondence analysis) and clustering of categorical variables. To account for equivalent categories, equality restrictions on the category quantifications were incorporated in the algorithm, resulting in a new technique. The second technique is latent class analysis (LCA) with equality restrictions on the conditional probabilities. In a simulation study, the clustering performance of restricted GROUPALS and restricted LCA was studied by assessing the recovery of true cluster membership of the objects, by means of the adjusted Rand index. The performance of both techniques ranged from acceptable to excellent, dependent on several systematically varied data features.

CDA

#37

CDA

Maarten Cruyff
 Peter G. M. van der Heijden
 Ardo van den Hout
 Ulf Böckenholt

Utrecht University
Utrecht University
MRC
McGill University

A Log-Linear Model to Estimate Cheating in Randomized-Response

Randomized response (RR) is an interview technique designed to eliminate response bias when sensitive questions are asked. In RR the answer depend to certain degree on to the outcome of a randomizing device. Although RR elicits more honest answers than direct questions, the method is susceptible to cheating, in the sense that respondents do not answer in accordance with the outcome of the randomizing device. In this paper we present a log-linear randomized-response model that accounts for cheating. The main results of this model are (1) an estimate of the probability of cheating; (2) log-linear parameters estimates describing the associations between RR variables and; (3) prevalence estimates of the sensitive behavior that are corrected for cheating. We illustrate the model with two examples from a Dutch survey measuring non-compliance with social welfare rules.

#87

CDA, GLM

Ayrin Calachan Molefe

University of Central Arkansas

Regions of Significance and the Point of Intersection in Comparative Studies using Logistic Regression

The Johnson-Neyman technique has been widely applied in educational research as an alternative to ANCOVA when regression slopes are not homogenous. It produces a region of significance which is a range of predictor values for which significant differences in the mean criterion score exist between the groups under comparison. This region of significance has been shown to be equivalent to the confidence interval for the abscissa of the point of intersection of the regression lines. Since it was originally formulated as an alternative to ANCOVA, the Johnson-Neyman technique is appropriate only for normally distributed criterion scores. However, in many comparative studies, it is not uncommon to have an outcome that is dichotomously scored. For example, examination results may be reported as either pass or fail rather than as a raw score (as in the nursing licensure examination or NCLEX). A modification of the Johnson-Neyman technique to dichotomous outcomes is presented and the resulting significance region is shown to be asymptotically equivalent to the intersection point confidence interval. The proposed extension is illustrated using real data and its small-sample performance is assessed through a simulation study.

#156

AAP, CDA

Elena Eroshva
 Emily Walton
 David Takeuchi

University of Washington
University of Washington
University of Washington

Using propensity score matching to investigate differential item functioning

The 5-category health status scale is used in a wide range of surveys across many countries. Some research suggests that self-rated health response categories may be biased for certain social groups, for example, for those who associate themselves with cultures that value modesty and restraint. In this case study, we investigate differential functioning of the self-rated health scale for Asian Americans with respect to their nativity status. We hypothesize that foreign-born Asian Americans are less likely to report extreme categories than their native-born counterparts of similar health status. We use propensity score matching to derive groups who share similar demographic and health characteristics. Each native-born person is matched to a foreign-born of the same ethnicity by nearest available Mahalanobis metric within a caliper defined by the propensity score. Propensity score framework allows us to make descriptive comparisons of self-rated health by nativity status, controlling for background characteristics. We find that nativity is not associated with higher likelihood of reporting the extremes on the health status scale. In addition, we find no evidence of imbalances in endorsement of any particular category between the two groups. Finally, we compare our approach to regression-type and anchoring vignettes approaches for investigating differential item functioning.

Timothy R. Johnson

University of Idaho

Multinomial Logit and Probit Models for Ordinal Response Variables:
A Reexamination and Generalization of the Stereotype Model

In this talk I present a class of hierarchical multinomial models for ordinal response variables based on a generalization of the stereotype model. A reexamination of the stereotype model suggests that it can be motivated by assuming that respondents choose among ordered response categories on the basis of underlying ordered utility/membership functions corresponding to the response categories. The stereotype model can be usefully generalized by relaxing the assumption that the latent responses are independent to account for individual differences in response style, and by considering a hierarchical model to account for dependence among responses within respondents. This talk concerns the motivation, specification, identification, and estimation of generalized stereotype models. Two applications are provided for illustration.

Carolyn J. Anderson
Zhushan Li
Jeroen K. Vermunt

University of Illinois at Urbana-Champaign
University of Illinois at Urbana-Champaign
Tilburg University

Pseudo-Likelihood Estimation of Rasch Models as Linear by Linear Association Models

Linear by linear association models for manifest probabilities can be derived from models in the Rasch family of models, including unidimensional and multidimensional models for dichotomous or polytomous items with or without covariates. The major problem with using linear by linear association models is that the current estimation methods are limited to small number of items. Pseudo-likelihood estimation is proposed as a solution to this problem. We show that the estimates are consistent and asymptotically normal. We also propose using Jackknife and Bootstrap to estimate standard errors and confidence intervals for parameters. The results of our simulation studies indicate that the pseudo-likelihood parameter estimates are nearly identical to the maximum likelihood ones with a negligible loss in efficiency. We use pseudo-likelihood estimation to fit multidimensional Rasch models as log-multiplicative association models to a large midwestern state assessment where we use Q-matrix developed for the test to specify the underlying multidimensional structure.

William H. Batchelder
Jared Smith

University of California at Irvine
University of California at Irvine

Modeling Subject and Item differences in Multinomial Processing Tree Models

Typically the data structure for both IRT models and cognitive memory models is the observation of a subject by item random matrix with categorical responses. Despite this common data structure, there has been little collaborative work between test theorists and cognitive modelers. IRT models have parameters associated with subjects or items, and memory models have parameters for specific cognitive processing steps like memory storage, organization, inference, and retrieval. Usually IRT models do not have parameters for latent cognitive acts, and memory models typically lack parameters explicitly associated with subjects or items. Multinomial processing tree (MPT) models are a popular class of cognitive models for memory and other areas of cognition. Recent work by ourselves and others has developed random effects versions of MPT models to capture individual differences in subjects and/or items. In this paper we use some of the psychometric ideas of Spearman and Rasch to model each latent MPT parameter as an additive logit in 'cognitive intelligence' and 'item difficulty' parameters. We think the work is leading to useful tools for psychological assessment of specific cognitive skills in special subject populations in areas such as aging, clinical, and cognitive neuroscience.

CDA / IRT

#138

CDA, IRT, GLM

Matthew S. Johnson

*Baruch College, CUNY*Constrained Quasi-symmetry Models for Positively Related Responses

The Rasch model is a generalized linear mixed effects model designed for the analysis of positively related dichotomous item responses. Several authors have demonstrated that the model is equivalent to a quasi-symmetry model for the 2^J contingency table, where the symmetry parameters can be related to the moments of a positive random variable. In this presentation I will discuss this and other less restrictive constrained quasi-symmetry models. I will introduce an estimation algorithm that guarantees the estimated symmetry parameters are contained within the constrained parameter space for the various classes of constrained quasi-symmetry models. Methods for model selection will be discussed and all techniques will be demonstrated with a real data set.

CTT

#3

CTT

Shun-Wen Chang

*National Taiwan Normal University*Comparisons of Score Transformation Methods for the BCTEST Using Real and Simulated Data

This study evaluated the effects of employing the linear, normalizing, arcsine, and log-odds transformation methods for constructing scale scores on the BCTEST, a nationwide standardized test that is used for high school admission in Taiwan. Tests in five subject areas (Chinese, English, Mathematics, Natural Science, and Social Studies) were studied using both the BCTEST real data and the simulated data. The resulting scale scores for each of the five tests were examined with respect to the raw-to-scale score conversions and measurement properties calculated based on the strong true score model. The effects of adjustments in rounding and truncating and the gaps resulting from the score conversions were investigated. The findings indicated that for all transformation procedures, the results produced by using the real and simulated data were very similar. For all tests, employing the arcsine transformation stabilized the error variability along much of the entire scale. But, the linear transformation yielded the most satisfactory results regarding the size of the gaps, and the normalizing approach created similar distributional characteristics among the tests. This research has offered useful information about the properties of scales based on different transformation methods.

#6

CTT, VCA

Sandip Sinharay
Shelby Haberman
Gautam Puhan

Educational Testing Service
Educational Testing Service
Educational Testing Service

Subscores for Institutions

Institutions often desire aggregate information on performance of their examinees on test subscores. Except for Longford, little research has considered the added value of such information given summary reports of the performance of examinees on total scores. We show that measures conceptually similar to those used for reliability analysis in classical test theory can be applied to assessment of the value of institutional summaries of examinee subscores. Subscores at institutional level are considered to have added value over the total scores when the true institutional mean is more accurately predicted by the mean total subscore of examinees from the institution than by the corresponding mean total score. Analyses of two operational data sets provide little support in favor of reporting subscores for either examinees or institutions.

CTT

#17

CTT, OTR

Jean-Paul Fox

*University of Twente*Classical Test Theory For Randomized Responses

In classical test theory, it is usually assumed that an observed test score can be treated as a continuous variable. Further, it is assumed that the test score has a distribution with a finite mean and variance over (hypothetical) repeated measurements of the same test to the same subject. The observed test score can also be treated as a discrete variable. The unweighted sum of the item scores for dichotomously scored items can be assumed to have a discrete distribution leading to a strong true-score theory. This way a true score model is defined for randomized response data. It is shown that the observed score of randomized responses on a test are distributed according to the Binomial error model where the success probability is a linear transformation of the proportion-correct true score. A proportion-correct true score can be inferred from the nonlinear regression of proportion-correct true score on the observed randomized score. Bayesian estimators are constructed for estimating individual proportion-correct true scores given randomized responses. A reliability statistic, related to the KR-21 statistic, is developed. Results are presented from a study measuring cheating behavior of Dutch university students where a randomized response sampling design was used.

FAC1

#5

FAC

Guangjian Zhang
Michael W. Browne*The Ohio State University*
*The Ohio State University*Dynamic Factor Analysis of Polychoric Autocorrelation Matrices

Most dynamic factor analyses have been applied to adjective rating data on Likert scales. These are usually treated as if they were continuous and product moment autocorrelation matrices are computed in the usual manner. Because the data are discrete it is natural to assume that the factor analysis model is satisfied by continuous unobservable variables which underlie the discrete data. The dynamic factor analysis model is therefore fitted to polychoric autocorrelation matrices estimated from the discrete data. Details are given of a simulation study which compares results obtained from product moment autocorrelation matrices with those obtained from polychoric autocorrelation matrices.

#11

FAC

Kamel El Hedhli
Jean-Charles Chebat*HEC Montréal*
*HEC Montréal*Proposing, Developing and Validating a Psychometric Shopper-Based Mall Equity Measure

This paper, based on Keller's consumer-based brand equity paradigm and store equity conceptualization of Hartman and Spiro, introduces a new marketing concept that of shopper-based mall equity (SBME). SBME is defined as the differential effect of mall knowledge on shoppers' responses to the mall's marketing activities. The authors report the results of a study undertaken in two Canadian malls with 905 shoppers who were administered a questionnaire. A third-order confirmatory factor analysis shows that SBME is a bidimensional construct, composed of two sub-constructs, namely mall awareness and mall image. The mall image component is in itself a multidimensional construct that can be captured by four dimensions. Further psychometric tests show a parsimonious SBME scale with support for convergent and discriminant validity, and appropriate reliability. Multigroup latent mean structures show that the SBME measure is able to discriminate shoppers that globally evaluate high a mall from those who globally evaluate low a mall, lending support for the sensitivity of the SBME measure to different situations. The authors provide a SBME index likely to aid mall managers in assessing and managing the equity of their own mall. The SBME measure provides a means for retailing researchers to examine potential outcomes of SBME.

FAC1

#12

FAC

Brian F. French
Holmes Finch

*Purdue University
Ball State University*

Locating the Invariant Referent in Multi-Group Confirmatory Factor Analysis

The use of multi-group confirmatory factor analysis (MCFA) has become a popular method for the examination of measurement invariance and specifically, factor invariance. Additionally, recent research has begun to focus on using MCFA to detect invariance for items on tests. MCFA requires certain parameters (e. g. , factor loadings) to be constrained for the purpose of model identification, and are assumed to be invariant across groups and act as referent variables. When this invariance assumption is violated, location of the parameters that actually differ across groups becomes quite difficult. The factor ratio test and the stepwise partitioning procedure in combination have been suggested as methods to locate invariant referents, and have performed favorably with real data. However, the procedures have not been evaluated through simulations where the extent and magnitude of invariance is known. This study examines these methods in terms of accuracy of identifying invariant referent variables. Data were simulated with known model differences across two groups under varying conditions of sample size, number of factors and indicators per factor, and percent of non-invariant indicators. Replications (N = 1000) for each combination of conditions ensured stable results. Findings suggest that Type I error and power are acceptable under limited conditions.

#14

FAC

Longjuan Liang
Michael W. Browne

*The Ohio State University
The Ohio State University*

An Extension Procedure to Circumplex: How two circumplexes are Related?

Circumplex model is used for the data whose relationships of several variables are shown to be in a circular order. It's been a useful model in personality and clinical area, especially for interpersonal personality scales. Among the many models and methods for circumplex data, Browne suggested a circular stochastic process model with Fourier Series function for the correlations among common scores of different tests. This model can handle negative correlations as well as positive ones. The corresponding program CIRCUM produces estimates of the polar angles for the variables and some other useful indices as output. Based on this model, we developed an extension procedure to project a new test or a new circumplex structure onto the existing one by analyzing the correlation matrix between the new variables and the existing variables. We also report a communality index as a measure of the extent to which the external variable is related to the existing variables. A practical example will be presented to illustrate this extension procedure.

FAC2

#27

FAC

Michael W. Browne
Guangjian Zhang

*The Ohio State University
The Ohio State University*

Rotation in Dynamic Factor Analysis

A critical difference between the classical factor analysis (FA) model and the process model for dynamic factor analysis is that in classical FA the factors are exogenous and in process FA they are endogenous. Consequently, inter-factor correlations are free parameters in classical FA and are functions of time series parameters in process FA. Rotation in process FA affects the time series weight matrices and the shock covariance matrix as well as the usual factor matrix. Furthermore, interpretable patterns are required for the time series weight matrices as well as for the factor matrix. As a result, methodology for rotation in classical FA must be modified for process FA. Rotation criteria suitable for process FA will be suggested and the Jennrich-Sampson algorithm for oblique rotation will be adapted for process FA. A practical example will be presented to illustrate the use of this methodology.

FAC2

#45

FAC

Roger E. Millsap
Soyoung Lee

Arizona State University
Arizona State University

Stochastic ordering of the expected factor score by the sum of the indicators

Recent work in item response theory (IRT) has established some useful results on the stochastic ordering properties of various IRT models. These results apply to either dichotomous or polytomous items, with some properties found to hold only partially in the polytomous case. Similar results do not appear to be available for the common factor model with continuous indicators. Here we will address the stochastic ordering of the expected factor score by the sum of the indicators in a single-factor model. We will describe conditions under which this stochastic ordering will hold, and conditions under which it will not hold. Implications of these results are discussed.

#82

FAC

Robert I. Jennrich

University of California, Los Angeles

What Do Factor Analysis Loadings and Their Standard Errors Estimate: The Signed Permutation Problem

The signed permutation problem in exploratory factor analysis arises from using rotation criteria that are invariant under column sign changes and permutations. A quick fix is to choose an optimizing loading matrix that has positive column sums and columns of decreasing length. While this works in theory, it can fail dramatically in practice. For typical sized samples estimates can have very non-normal distributions of disturbing width. A fix is to change the definition of what is being estimated. Rather than viewing an estimate as an estimate of a population loading matrix, view it as an estimate of the population matrix aligned with the estimate. Error distributions using alignment can be much narrower and much more normally distributed than without alignment. Moreover, aligned versions of jackknife and bootstrap standard errors can be much smaller than ordinary jackknife and bootstrap standard errors and can provide reasonable estimates of aligned root mean squared errors. This is also true using Browne's asymptotically distribution free standard errors that do not require alignment. Coverage probabilities for confidence intervals using these standard errors are obtained. They are surprisingly similar and accurate. Asymptotically distribution free standard errors are computationally much less expensive than jackknife and bootstrap standard errors.

FAC3

#93

FAC

Duan Zhang
Victor L. Willson

University of Denver
Texas A & M University

Evaluating Factor Structure Stability from Summary Data: A Monte Carlo Investigation

Factor analysis has been widely applied in behavioral sciences to investigate the psychometric properties of measurement scales. Common practice is to conduct an exploratory factor analysis (EFA) with one sample to identify an initial factor structure. A confirmatory factor analysis (CFA) would then follow to test the structure with another sample. When the CFA model fails to confirm the EFA structure, usually the researchers would want to draw additional samples to get more evidence, which could not always be accomplished easily or in an affordable manner. Two published articles using EFA and CFA were taken as examples. For one the EFA structure was replicated in CFA while for the other it failed. This study utilized the available EFA data together with descriptive statistics to simulate the raw data matrices which were then used to fit both the correct and misspecified CFA models. The focus was to explore the influence of sample size and number of replications on the model fit. The results would help researchers using EFA and CFA in the above way, especially those with unfavorable outcome, to get more evidence from replications.

FAC3

#111

FAC

Gilles Raïche
 Martin Riopel
 Jean-Guy Blais

Université du Québec à Montréal
Université du Québec à Montréal
Université de Montréal

Non Graphical Solutions for the Cattell's Scree Test

Several strategies have been proposed to determine the number of components to retain following a principal component analysis of a correlation matrix. Most of these rely on the analysis of the eigenvalues of the correlation matrix and on numerical solutions. For example, Kaiser's eigenvalue greater than unity rule, parallel analysis and significance tests, like the Bartlett test, allow to make use of numerical criterions of comparison or statistical significance criterions. Apart of these numerical solutions, Cattell proposed the scree test, a graphical strategy to determine the number of components to retain. With the Kaiser's rule, this test is probably one of the most used strategy and is integrated in almost all statistical software dealing with principal components analysis. Unfortunately, it is generally recognized that the graphical nature of the scree test doesn't favour agreement in the number of components to retain. To palliate this problem, some numerical solutions are proposed. A first family of numerical solutions deals with the acceleration of the plot of the eigenvalues, while a second family, more in Cattell's spirit, deals formally with the scree part of this plot.

#165

FAC

Kensuke Okada
 Hitomi Tahara
 Takahiro Hoshino
 Kazuo Shigemasu

The University of Tokyo
Senshu University
The University of Tokyo
The University of Tokyo

Maximizing Simplicity: A new factor rotation method

In factor analysis, a factor rotation is performed aiming to obtain simple structure. Recently, Lorenzo-Seva proposed a new index of factor simplicity (called LS), which is based on the idea that the communality of each variable should be related to few factors. Although LS index has attractive properties such as high power and insensitivity to the scale of the factors compared to existing ones, the main disadvantage of LS is the maximization, because its derivative is difficult to derive. We propose using the derivative free gradient projection algorithm of Jennrich to numerically obtain the simplest factor loadings. The method is a modification of gradient projection rotation algorithm, and gives almost precisely the same result as using exact gradients. The proposed rotation is compared with other existing rotation methods Monte Carlo simulation.

#172

FAC

Ab Mooijaart

Leiden University

An Estimation Procedure in Independent Component Analysis and Factor Analysis

In this paper a comparison will be made between independent component analysis (ICA) and factor analysis (FA). FA is an important model for almost a century in mainly Psychology. ICA is popular since the eighties of the last century in disciplines like signal processing neural and cognitive sciences, information theory and engineering. The main resemblance between both models is the assumption of the existence of latent variables, i. e. the factors. The most important difference between both models is the assumption of independence of the latent variables in ICA. This assumption is stronger than the assumption of uncorrelatedness, which is often made in FA. Another difference between ICA and FA is that in the latter model it is assumed that the observed variables are measured with a measuring error. This assumption is not made in ICA. In this paper a method for estimating the mixing matrix in ICA (the matrix of factor loadings in FA) will be proposed. This method is based on fitting the first and second order moments and the fourth order cumulants. An empirical study will show how well our estimation procedure behaves.

Joseph R. Rausch

*University of Notre Dame*Methods for Investigating Change in Intraindividual Factor Structure Over Time

"Structural change is a prominent thread in the fabric of developmental theory and, to the extent that it is meaningfully represented as changes in factor-loading patterns, its appearance should not be summarily precluded" (Nesselroade, 1983, p. 63). Thus, while changes in factor structure are often seen as hindrances to scientific research, such as in the context of factorial invariance, the investigation of factor structure changes can provide new opportunities for the development of theory in psychology. The model proposed in the present work to investigate such structural changes over time, the Time-varying Individual Factor Structure (TIFS) model, is an extension of P-technique factor analysis used to examine changes in the factor loadings at the individual level. The method of estimation utilized relies on a two stage approach: (1) windowed estimation to estimate the factor loadings at a particular time point, termed "local P-technique estimation", followed by (2) fitting trajectories over time to the factor loadings to model reliable changes in factor structure. The potential scientific impact of the TIFS model in psychology is illustrated via an application to data on the self-regulation of mental distress for bereaved widows.

Andrew Dean Ho

*University of Iowa*Describing the Pliability of Effect-Size-Based Trend and Gap Statistics Under Monotone Transformations

If the equal-interval property of a test score scale is suspect, monotone transformations of the scale may be reasonable and change the magnitudes of effect-size-based trend and gap statistics. Previous work has shown that trends and gaps can undergo sign-reversal under transformations when the Cumulative Distribution Functions (CDFs) of the two distributions cross. However, crossed CDFs are neither necessary nor sufficient to raise concerns about the validity of trend and gap interpretations; sign-reversal may only occur under unrealistically violent transformations, and transformation-induced pliability may be a concern even if sign-reversal is impossible. This paper develops a mathematical framework for describing the pliability of effect-size-based trend and gap statistics under "plausible transformations." Three different families of plausible transformations are presented; these can be located on a continuum between defensible and questionable equal-interval scale properties. The mathematical framework extends principles from the Receiver Operating Characteristic (ROC) curve literature to give a "plausible range" of effect sizes under different families of transformations. Plausible ranges are calculated for commonly cited trends and gaps from the National Assessment of Educational Progress (NAEP) for discussion. When the stakes on trend and gap interpretations are high, these interpretations should be invariant to plausible transformations of scale.

Denis Alamargot
Gilles Caporossi
David Chesnet

University of Poitiers
GERAD & HEC Montréal
University of Poitiers

Assessing the visual strategy of the writer composing from sources

Writing from sources is a frequent task at the workplace and its importance increases with the use of computers. This task necessitates a double competence in both reading-extracting information and inventing-composing text. The 'Eye and Pen' system was designed to study these skills. Based on a synchronous recording of eye movements (via an eye-tracking system) and pen movement (via a digitizing tablet), the dispositif provides a fine-grained description of the visual strategies used by the writer while composing. To study these strategies, we recorded the graphomotor and eye activity of 25 adults while they were composing a procedural text by referring to documentary sources. Multivariate analyses confirms the existence of contrasted visual patterns, depending on the writer capacity. From the visual patterns, some statistics are computed. One of them is the transition matrix indicating the frequency with which the eye moves from zone *i* to zone *j*. Other statistics such as the average number of different informations explored when the eye leaves the writing zone are also considered. From these various statistics, a measure of dissimilarity between respondents is computed and clustering is used to find groups of respondents with similar exploration patterns. The results of this study will be presented and discussed during the talk.

FAC / EDA

#133

FAC, OTR

Takashi MURAKAMI

*Nagoya University*Oblique Procrustes transformation of principal components by orthonormal transformations of the matrix of weights

We propose a simple method of oblique Procrustes rotation of principal components. The matrix of weights for component scores consisting of standardized eigenvectors corresponding to m largest eigenvalues of a given matrix of correlation coefficients is orthogonally rotated to attain least-squares approximation of a hypothetical target matrix whose columns are vectors of length one. The solution obtained by the method is free from the indeterminacy due to the arbitrariness of scale of the target, and each column of the obtained pattern matrix is proportional to the corresponding column of the rotated matrix of weights as is in the case of the Independent Cluster Rotation by Harris & Kaiser. The least-squares criterion is equivalent to the maximization of the sum of the coefficients of congruence between matched columns of the pattern and of the target. By using general orthonormal transformation, the procedure is extended to the cases where the number of columns of the matrix of weights and that of the target are different.

#54

FAC

Soonmook Lee

*Sungkyunkwan University*Factor Analysis on Data of Multiple Testlets

In a testlet-based test, items are locally dependent due to the context created by the text in each testlet, causing difficulty in estimating factor structure. We propose to estimate a factor structure from raw data disregarding differential context effects perturbed in item scores and use the factor structure as a hypothesis in a subsequent confirmatory factor analysis with uniquenesses correlated within each testlet. This approach enables us to keep the communality in the data and take into consideration that item scores are not locally independent. We performed a Monte-Carlo study to examine the effectiveness of this approach. We had a factorial design of 3 levels of communality (.16, .36, .64), 3 levels of factor correlation (.1, .3, .5), and 3 levels of uniqueness correlation (.1, .3, .5). Sample size in each cell was fifteen which was needed to obtain statistical power of .80 with medium effect size. The result from oblique rotation was input to confirmatory factor analysis with uniquenesses correlated in each testlet. Analyses in many cells showed convergence and good fit to population correlation matrix. However there are some cases that do not converge. Details will be examined and discussed.

GRM / CCC

#104

CCC, IRT, LDA

Moon-ho Ringo Ho
Hernando Ombao*Nanyang Technological University & McGill University
University of Illinois at Urbana-Champaign*Discrimination and Classification of Non-stationary Psycho-physiological Signals Using the SLEX Model

There has been extensive research for the classification and discrimination problem of stationary signals in physical sciences and engineering problems. However, many psycho-physiological signals are non-stationary in nature. For example, brain waves recorded during an epileptic seizure have waveforms whose amplitude and oscillatory behavior change over time. Thus, the classical Fourier-based methods assuming stationarity are seriously limited for general use in neuro-scientific research. In this talk, a computationally efficient discriminant-classification scheme that can extract local features of the time series will be introduced. This scheme is based on the SLEX (Smooth Localized Complex Exponential) library, which forms a collection of bases that are orthogonal and localized in time and frequency domains. The first step involves finding a basis from the SLEX library that can best illuminate the difference between known classes of time series. A discriminant criterion that is related to the likelihood ratio between the SLEX spectra of the different classes is proposed. The second step is to extract diagnostic features from the best basis obtained from the first step. These features can be used to classify time series with unknown class membership. Applications of the proposed method to EEG signal classification in brain-computer interface research will be discussed.

Frank Rijmen
 Jiří Vomlel

*VU University Medical Center
 Academy of Sciences of the Czech Republic*

Assessing the performance of variational methods for mixed logistic regression models

Variational techniques encompass a variety of approximation techniques. The common denominator is to simplify a complex optimisation problem by the introduction of additional, variational, parameters. For a fixed set of values for the variational parameters, the transformed problem has a simpler (e. g. closed-form) solution, providing an approximate solution to the original problem. The variational parameters are optimised in a separate step. Estimation is performed by alternating these two steps in order to obtain a sequence of approximations of increasing accuracy. Examples of the use of variational lower bounds to the log-likelihood function can be found in the context of missing data for Markovian models, and in the context of graphical models. We will present a scheme for estimating the parameters of a mixed logistic regression model based on the lower bound variational approximation of the logistic function proposed by Jaakkola and Jordan. In addition, we assess the performance of the variational technique for the Rasch model, comparing it to established estimation techniques such as Gaussian quadrature, PQL, and the Laplace approximation.

Christopher W. T. Chiu

*Accu Measurement & Testing (AMT),
 National Center for Educational Statistics (NCES),
 & American Statistical Association (ASA)*

An Overview of Data Visualization for Psychometrics and Educational Measurement

It is a challenge to display large-scale categorical, longitudinal, and repeated-measure data, such as data collected in educational measurement and psychometrics. The challenges lie on factors such as the number of categories, the number of cases, the number of variables, and the amounts of missing information that appears in these types of data. Design problem is another concern when displaying complex information (e. g. , representation of abstract information, layering, color coding, details-on-demand association, orientation of objects, relationship understanding, and event representation). Since abstract information does not have a physical shape like geographical landmarks, designing cognitively-oriented, user-friendly, expressive, and effective displays becomes difficult in any information visualization techniques for psychometrics and educational measurement. To this end, a number of widely used methods have been developed and modified to overcome these limitations. The methods include, for example, GAP, Linked-Micromap, Mosaic, SEER, and Trellis. We first summarize frameworks of data visualization (e. g. , Scalar vector graphics, SEER, pixel-based methods, and Mackinlay's evaluation framework) in the context of psychometrics and educational measurement; we then illustrate successful examples showing how to visualize complex information in the context of psychometrics, survey, and assessments.

István Hidegkuti
 Paul De Boeck

*K. U. Leuven
 K. U. Leuven*

A comparison of two approaches to investigate the categorical vs. dimensional nature of psychological phenomena

To find out the latent structure (i. e. , the categorical or dimensional nature) of different phenomena there are several methods exist of which taxometrics is the most well-known and most widely used, mainly in the field of psychopathology. Taxometrics stands for a series of methods that are based on Coherent Cut Kinetics. A recent development to address the continuity vs. discontinuity controversy is the Dimension/Category (DIMCAT) framework that is based on item response theory. A simulation study will be presented to compare the performance of both methods under various circumstances, including different data types, category main differences, within-category heterogeneity, etc. Of the taxometric methods MAXCOV and L-Mode will be used for the comparison.

IRT1

#22

IRT

Dipendra Raj Subedi

*Michigan State University*The Optimal Design of Two Stage Test

Numerous years of research in Computerized Adaptive Testing (CAT) has greatly improved the use of item pools and selection of test items to match examinee ability. However, it is my contention that the innovative use of two-stage testing will provide even better improvements. Much past research on multi-stage testing has focused either on performance of multi-stage design or on comparison of linear fixed, multi-stage, and adaptive test designs. Nevertheless, research on design of two-stage test that provides the most accurate estimates of the trait level has received limited attention. This study extends a prior research by Reckase, and investigates for the optimal combination of first-stage (routing) test and second-stage (measurement) test that will give the most accurate estimates of trait level. Another aspect of this study is to determine the best location for the cut-scores in such a way that the measurement accuracy is maximized. That to say, this study also focuses on the precision with which examinee ability is measured in the first-stage or routing test so that the examinees are accurately routed to the appropriate measurement test matching their ability.

#23

IRT

Kelly Rulison
Eric Loken*The Pennsylvania State University*
*The Pennsylvania State University*Reconsidering the need for an upper asymptote (< 1) in computerized adaptive testing

One concern with the standard three-parameter IRT model is that for high ability students, missing even one easy question may negatively bias estimates due to the assumption that $P(X=1|\theta;a,b,c)$ is effectively equal to 1 for such items. Mislevy and Bock considered this problem, and Barton & Lord added an upper asymptote, δ , re-estimating student scores on several tests. They concluded that adding a fixed $\delta = .98$ did not impact ability estimates and was not worth the effort. In computer adaptive testing, however, the 3PLM's asymmetry may be more serious than in fixed-length tests as estimates are continually updated after each response. Through simulations we show that early incorrect answers bias final theta estimates, particularly for above average students ($\theta \geq +1$). Early misses create a steep drop in the estimates, followed by a slow ascent such that even after 50 items high ability students are not seeing items matched to their true ability. We revisit Barton and Lord's four-parameter model and show that setting $\delta = .98$ can indeed lower bias and RSME for these misestimated cases. The CAT algorithm ascends more quickly after initial underperformance, allowing for less biased estimates for high ability respondents.

#30

IRT

Cheryl D. Hill

*RTI Health Solutions*A Model for Longitudinal Item Response Data

Questionnaires are sometimes administered to the same sample of examinees on more than one occasion. Even when longitudinal data are available, researchers employing item response theory (IRT) often use data only from the first administration for item calibration because there is likely a lack of conditional independence between responses by the same individual. However, in many longitudinal study designs, the sample size at one occasion is too small for reliable item calibration. Thus, a longitudinal IRT model for use with repeated measures study designs is desirable. This research developed two distinct approaches to longitudinal IRT. One of these models is based on latent class analysis, while the other is based on full-information bi-factor analysis. Both account for the local dependence among items that are administered twice by introducing parameters that describe how the repeated nature of each item affects the response (separately from the effect of the latent trait). The models include parameters that describe the latent trait distribution at the second administration relative to the standardized distribution at time one and the correlation between the latent traits at two time points. The addition of these model components allows item parameters to be calibrated using available data from two occasions.

IRT1

#31

IRT

Michelle M. Langer

*University of North Carolina at Chapel Hill*Linking in Developmental Scales

Developmental scales permit achievement test performance to be compared across grade levels, and individual growth to be assessed in terms of changes in average performance and variability from grade to grade. The use of developmental scales to summarize student achievement has attracted more interest with recent legislation. However, there remain a surprising number of unanswered methodological questions. Through a simulation study, this research examines the effects of the number and difficulty level of linking items, the inclusion of noncommon items, the size of the mean difference between groups, and the sample size in accurately linking a score scale across groups. All data are simulated for binary items, without taking guessing into consideration, based on parameters from the 1998 National Assessment of Educational Progress Reading Assessment. This study uses item response theory concurrent linking for 1,000 replications each under more than 100 conditions. These results are evaluated using the average difference between the true mean difference between groups and the recovered mean difference between groups, as well as root mean squared error.

IRT2

#39

IRT

Han Bao

University of Maryland at College Park

Amelia Wenk Gotwals

University of Michigan

Nancy Butler Songer

University of Michigan

Robert J. Mislevy

*University of Maryland at College Park*The Structured Item Response Theory Models for Change in a Pretest-Posttest Setting

Pretest-posttest designs are widely used in educational and psychological research, primarily for the purpose of measuring change resulting from experimental treatments. A number of psychometric models for measurement of change have been developed within Item Response Theory. The focus of this paper is to present a set of multidimensional structured item response theory models for measuring change and learning based on research by Fischer, Spada, and others. As special cases of the multidimensional random coefficient multinomial logit model (MRCMLM), our models can take into account the differences in item difficulties between two time points by reflecting the change in the parameters of the multidimensional linear logistic test model. Assuming that the basic item parameters are the same for pretest and posttest, the variation in item difficulty between the two time points is due to learning and is expressed by a series of changing effects related to the experimental treatments. The interaction of the changing and treatment effects can also be taken into consideration. With such models, it is possible to study the effects of instruction on distinguishable affects of knowledge, given that tasks can be created to target that knowledge in isolated or in known combinations. The practical applications of the models are illustrated in the framework of the University of Michigan's BioKIDS 2002-2003 Fall Assessment.

IRT2

#53

IRT

Mark D. Reckase

*Michigan State University*The Effect of Inter-trait Correlations on MIRT Calibrations

All item response theory models need to address scale indeterminacy issues when model parameters are estimated. For multidimensional item response theory (MIRT) models, three types of indeterminacy need to be addressed: origin of the space; units for each coordinate axis; and rotational orientation of the axes. Estimation programs account for these indeterminacies by setting the mean proficiency vector to the 0-vector, the units to the standard deviation of the estimated proficiencies, and fixing the rotation based on characteristics of the item set. Sometimes the assumptions of the estimation procedure do not match underlying data structure. This is most obvious when simulation studies use MIRT models. When data are generated using correlated dimensions and parameters are estimated from the data, the estimates will not match the generating parameters because the way indeterminacy is handled does not match the underlying structure. This paper evaluates parameter recovery using TESTFACT when dimensions are correlated and uses the analyses to clarify the distinction between correlated dimensions and correlated coordinates. Several inconsistent interpretations of correlated dimensions from published articles are used to emphasize the importance of clear understandings of the constraints placed on MIRT estimation procedures.

#57

IRT

Maomi Ueno

*University of Electro-Communications*A new IRT parameters estimation method using a joint binary probability distribution

This paper shows that the 2-parameters logistic IRT model can be analytically derived from a joint binary probability distribution under a condition about the prior distribution. Using this result, this paper proposes a new IRT parameters estimation method without using any iteration process such as a Newton method. The unique features of this method are as follows: 1) The estimation method uses only counting the number of item response patterns. 2) The estimation method can estimate the parameters from the data including all correct responses or all wrong responses. Furthermore, this paper provides some Monte-Carlo simulation experiments to show the effectiveness of the proposed method. The results show that this estimation method is effective and efficient.

#58

IRT

Ernesto San Martín
Jean-Marie Rolin*Pontificia Universidad Católica de Chile*
*Université Catholique de Louvain*Are the 1PL, 2PL and 3PL models with random effect identified?

Psychometricians know that the 3PL with random effect is difficult to fit. In practice, a lot of items should be eliminated to obtain a "reasonable" solution, but far from satisfactory. This phenomenon could be explained by the lack of identifiability of the 3PL with random effect. Nevertheless, up to the best of our knowledge, there does not exist a published proof on its identifiability (or lack of it). A natural strategy to consider this problem is to start by the identifiability of the 1PL with random effect. However, a formal proof (not a proof confusing the existence of a consistent estimate with identifiability) is known by someone. In this paper we want to answer the following questions: (1) Is identified a Rasch model when the ICC function is replaced by other one and when the normal distribution of the random effect is replaced by a general one? (2) Suppose the random effect of the Rasch is distributed according to a general probability distribution G; under which conditions, G and the difficulty parameters are identified? (3) Is the 3PL with random effect and the discrimination parameters equal to 1 identified?

IRT3

#64

IRT

Fumiko Samejima

*University of Tennessee*Some Considerations in Mathematical Model Selection for Graded Responses

A researcher tends to select a mathematical model for his/her data rather casually, without considering whether the model fits the nature of the data or not. This could lead to a serious mistake, because once a model is set, it provides some outcomes, whether they describe or distort the psychological reality behind the data. In the present paper, several important considerations in selecting a mathematical model for graded responses are considered. Truly scientific principle suggests for us to start from non-parametric estimation of the operating characteristics of graded responses without assuming any mathematical form, and then fit a carefully selected mathematical model. Goodness of fit of a model to the outcomes of the nonparametric estimation is not a sufficient condition for the model validation. Thus out of all mathematical models that have reasonably high levels of fit selection of a model should be considered as to whether the principle (s) behind it fits the nature of the data, whether the responses are discrete in nature or could be recategorized holding additivity or even tend to continuous responses in the limiting situation. Also for the benefit of practioners, it is exemplified that failure of selecting a right model can be ameliorated by switching to another more appropriate model without reanalyzing the data.

#65

IRT

Yue Zhao

*University of Massachusetts Amherst*Does Choice of IRT Model Make Differences in Equating?

Various methods and approaches have been suggested for detecting misfit in IRT models. While no general agreement has ever been reached on the best methods or approaches to use, perhaps the more important comment based upon the research findings is that rarely does the research evaluate item misfit by focusing on the consequences of using misfitting items and their less than accurate item statistics. For example, the one-parameter model and the Partial Credit Model may not fit a data set very well in some instances, but what are the implications for identifying equating tests. Are tests improperly equated, and if so, by how much? How far off are the ability estimates? Ultimately, it is the consequences of the misfit that should be considered in deciding on the merits of a particular model for use in particular situations. The current study attempted to evaluate the nature of item level misfit and its consequences in equating. Both binary-scored and polytomously-scored realistic data were analyzed.

#66

IRT

Cheongtag Kim
Heungchang Kwon*Seoul National University*
*Seoul National University*Applying Kernel Method for Scaling of Latent Trait

The present study proposes a kernel method to scale a latent trait from correlated binary items. The kernel method embeds the lower dimensional data into the higher dimensional space to reveal the nonlinear structure of the data. To scale a latent trait, a three-step-procedure is applied. First, each binary variable is regressed on the other binary variables in the higher dimensional space and the predicted value is used to estimate the continuous latent variable producing the binary response. Second, the continuous latent variables are then mapped into higher dimensional space and the principal axis is found on which the scale of the latent trait is located (Majoring kernel method). Third, the higher dimension is restored to the original dimension and the value of latent trait is estimated. Simulation studies were conducted to compare the majoring kernel method with the other scaling methods such as IRT, factor analysis, and Bayesian. The results showed that the present kernel method was better than the other methods in recovering the true values of latent trait. These results suggest that the proposed kernel method may be a promising candidate for a method to scale a latent trait from binary items.

IRT3

#72

IRT

Margo G. H. Jansen

*University of Groningen*Scoring Rules for Time-Limit Tests

Many tests intended to be power tests are actually more or less speeded because a considerable number of people are unable to attempt every question. The strategy selected for scoring missing item responses may have considerable impact on the interpretation of the results both on individual and on group-level. Even simple scoring rules, such as number correct or proportion correct, might favor different response strategies. A number of authors have argued that scoring rules should allow for separating the speed and accuracy (or precision) components. In this paper we compare simple scoring rules such as counting the number of correct with alternative scoring rules based on an IRT model, in particular with respect to the predictive validity of the test scores. For this purpose we used two subsets of a non-verbal intelligence test battery, while standardized achievement tests were used as external criteria. None of the different scoring rules proved to be clearly superior over the others. A simulation study with specifications chosen to resemble the empirical example was performed to gain more insight in these results.

#179

IRT

Negar Sharifi Yeganeh

*National Organization for Educational Testing*Assessing the IRT Item Parameter Invariance Property for Mathematics Test of University Entrance Exam

An increasing growth in the application of test in different situations, make testing and test analysis become more important. There are two approaches for test and item analysis: classical test theory and Item response theory (IRT).

IRT has several desirable features. One of them is that the item parameters are not dependent upon the ability level of the examinees responding to the item and also ability estimates are not dependent upon the test. In other words, in IRT, item and ability parameters are said to be invariant. The purpose of this study was to investigate the property of item parameter invariance for mathematics test of university entrance exam for math and physics. Invariance of item parameters across different gender was assessed. Data have been drawn from the 2003 administration of university entrance exam. For investigation of item parameter invariance, four random samples of size 1500 were drawn from examinees population. Gender groups were sampled using random sampling procedures. After checking the assumptions of model, item parameter estimate was performed using BILOG MG3. Then the degree of linear relationship between item parameter estimations in examinee subgroups were examined. Results suggest that item parameters are invariance with respect to gender.

IRT4

#74

IRT

Yuan-Horng Lin
Seock-Ho Kim
Allan S. Cohen*National Taichung University
The University of Georgia
The University of Georgia*Local Dependence Indices and Detection Investigation for Polytomous Items

The assumption of local independence under item response theory (IRT) is important if the benefits of IRT are to extend to the data being calibrated. If local independence is violated, the unidimensional IRT models may not fit the data. Two common situations are considered in which local independence may be violated. One is due to a type of item format known as a testlet in which a common stimulus is used for a subset of items. A second situation is that in which tests are speeded likewise, thereby, violating the local independence assumption for some examinees. Four local dependence indices implemented in the computer program LDIP are used in this study. LDIP is intended to assess local dependence in tests containing polytomous items under IRT. The four indices are the Pearson chi-square statistic, X^2 , the likelihood-ratio chi-square statistic, G^2 , Yen's index of local dependence, Q_3 , and the Fisher-transformed correlation difference statistic, Z_d . A simulation study using the LDIP program indicates that statistical power of the four indices is acceptable. All four indices appear to be sensitive to violations of local independence. For an item K response categories, X^2 and G^2 indices have distributions similar to a χ^2 distribution with $(K - 1)^2$ degree of freedom.

IRT4

#76

IRT

Jay Verkuilen

*University of Illinois at Urbana-Champaign*The Fisher Information Function in Unfolding Item Response Models

The Fisher Information Function (FIF) is one of the essential quantities generated by an IRT model. Indeed, arguably the parameter estimates are really a means to getting it and the ICC. It allows users of IRT to design tests with desired levels of precision in the locations of the latent trait where measurement precision is required. One peculiar feature of many existing unfolding IRT models is the fact that the FIF is bimodal, with the FIF identically equalling 0 when the subject is given an item that exactly matches her ideal point. A bimodal FIF is particularly undesirable for the construction of a computerized adaptive test (CAT) because the decision strategy behind the algorithm becomes complicated given the fact that there are usually two plausible directions in which to move. Rigorous explanations for the phenomenon have, unfortunately, been lacking. This article provides necessary and sufficient conditions for the FIF of unfolding IRT models to be unimodal. The theorems are general and apply to a broad class of unfolding models.

#78

IRT

Eduardo Rodríguez

Pontificia Universidad Católica de Chile

Paul De Boeck

K. U. Leuven

Ernesto San Martín

*Pontificia Universidad Católica de Chile*An Extension of the SSB-formulation of the Rasch model to the LLTM and to the PCM

Mellenbergh and Vijn formulated the Rasch model as a sum score based (SSB) model of the log-linear type, in which items and person sum scores are used as factors. Building on this idea Del Pino, San Martín, González and De Boeck examine the relationships between the maximum likelihood estimation (MLE) and their standard errors of the two different estimation procedures: the joint maximum likelihood (JML) and the SSB procedure both applied to the Rasch model. In this paper we present an extension of the SSB-formulation of the Rasch model to the LLTM and the Partial Credit Model. The bias corrections proposed by Haberman and Baker will also be applied to these models and simulation results will be reported.

#81

IRT

David Thissen

University of North Carolina at Chapel Hill

Li Cai

University of North Carolina at Chapel Hill

R. Darrell Bock

*University of Illinois at Chicago*A new parameterization of the nominal item response model

The nominal item response model is increasingly used as a model for patterns of item responses when individual items are combined to create testlets, and to investigate empirically the order of item responses on questionnaires (as in the personality and health domains) when nominal alternatives may have no clear *a priori* order. There are a number of alternative parameterizations for the nominal model (and for its strictly graded specializations, like the generalized partial credit model, the partial credit model, and the rating scale model). However, none of these parameterizations invite a practical extension of the nominal model to become a multidimensional item response model. In this presentation, we describe a new parameterization of the nominal model that is a step toward item factor analysis for strictly nominal responses, in the sense that the order of the responses (in the direction of greatest discrimination in the multidimensional space) is not specified *a priori*, but rather is determined as part of the analysis. We describe the relation of the new parameterization to many of the previous representations of the model, and present preliminary results on maximum likelihood estimation in this form.

IRT5

#90

IRT

Yanyan Sheng

*Southern Illinois University*Bayesian IRT models incorporating general and specific abilities

As item response models gain increased popularity in large scale educational and measurement testing situations, many studies have been conducted on the development and applications of unidimensional and multidimensional models. However, to date, no study has yet looked at IRT-based models with an overall ability dimension underlying all test items and several ability dimensions specific for each subtest. The purpose of the study is to propose such a model under the Bayesian framework so that both general and specific abilities can be estimated with one implementation. The proposed model is further compared with the conventional IRT models using Bayesian model choice techniques. The results from simulation studies as well as actual data suggest: 1) models with general and specific abilities can be developed, 2) fully Bayesian method is proved to be more accurate and efficient in parameter estimation compared with the usual marginal maximum likelihood method, 3) compared with the conventional IRT models, the proposed model describes the actual data conceivably better. Therefore, the proposed model offers a better way to represent the test situations not realized in existing models. The model specifications for the proposed model also give rise to implications for test developers on test designing.

#94

IRT

Johannes Hartig
Andreas Frey*German Institute for International Educational Research (DIPF)*
*Leibniz Institute for Science Education (IPN)*Using Plausible Values from Multidimensional IRT Models to Estimate Change in Large Scale Assessments

In matrix sampling designs employed in large scale assessments, each student answers only part of the items used to assess a certain subject area, resulting in sparse response patterns. Instead of estimating individual point estimates for students' proficiencies, conditional plausible values can be estimated to directly obtain unbiased estimates for population characteristics on IRT scales. If plausible values are estimated from multidimensional IRT models, they also provide unbiased estimations of correlations between latent ability dimensions. The present paper investigates the use of multidimensional conditional plausible values for the assessment of change in large scale studies. Student proficiencies at different time points are treated as different latent dimensions within a multidimensional Rasch model, while the item parameters are constrained equal across time points. Change scores are obtained as differences between plausible values for the separate dimensions. The performance of this method was examined in a simulation study. Compared with other ability estimates, multidimensional conditional plausible values performed best in the estimation of the overall change as well as of interactions between group variables and time, i. e. differential change. The method is illustrated further by an application to data from a German large scale assessment of language competencies.

#98

IRT

Irin Moustaki
Martin Knott*Athens University of Economics and Business*
*London School of Economics*Identifying extreme response patterns: a latent variable modeling approach

The paper investigates the problem of over-represented response patterns in a sample. For example, giving positive/negative answers to all items under analysis can be the result of guessing or avoidance to think properly each question. We propose the use of a latent variable model that accounts for extreme response patterns in the estimation of the model parameters. The methodology proposed distinguishes between cases where guessing is completely at random and cases where the guessing response mechanism depends on the latent variables measured by the observed items. The proposed methodology will be illustrated with real examples.

IRT5

#114

IRT

Nilufer Kahraman
Paul De Boeck
Rianne Janssen

K. U. Leuven
K. U. Leuven
K. U. Leuven

Modeling DIF in Complex Response Data Using Test Design Strategies

This study introduces an approach for modeling multidimensional response data with construct-relevant complexity factors. The item level parameter estimation process is extended to incorporate the refined effects of test dimension and group factors. Differences in item performances over groups are evaluated distinguishing two levels of differential item functioning (DIF): a domain level and an item level. An illustration is presented using dichotomously scored responses of 535 3rd and 4th grade students completing a spelling proficiency scale. A set of IRT Models was estimated using an adaptation of the logistic regression approach. The model with domain specific item-by-grade interactions or DIF performed better than the other models neglecting domain or grade differences. The method appears to be promising in that explicit domain factors can be implemented into model estimation procedure to better understand why items function differently over subpopulations. The goodness of fit is also expected to improve by the use of additional or auxiliary information from complexity factors, which may enhance DIF evidence if unaccounted.

IRT6

#124

IRT

Sun-Joo Cho
Allan S. Cohen
Seock-Ho Kim

The University of Georgia
The University of Georgia
The University of Georgia

An Investigation of Priors on the Probabilities of Mixtures in the Mixture Rasch Model

Mixture item response theory (IRT) models have been used for detecting latent groups which fit one IRT model better than another. The simplest of these combines a continuous IRT model, a Rasch model, with a latent class model. Estimation of simple models is done usually with maximum likelihood algorithms. As models become increasingly complex, however, Bayesian methods become more appealing. Prior distributions are known to be important factors in Bayesian applications. Improperly specifying a prior can affect the accuracy of the resulting estimates. Of particular concern for mixture IRT modeling is the effects of priors on estimates of model parameters, particularly for the mixture probabilities. In this study, we implement a Bayesian approach using a Markov chain Monte Carlo (MCMC) algorithm for a mixture Rasch model and compare three priors for the mixture probabilities used in previous research with finite mixture modeling: a Dirichlet prior, a Dirichlet process with stick-breaking prior, and a multinomial logistic regression prior with a covariate. Results of a small simulation study suggest all three priors performed well at detection of latent group membership. Recovery of item parameters, however, appeared to differ among priors. A simulation study will be presented in which the impact of the three priors is assessed on model estimates.

#132

IRT

Sayaka Arai
Shin-ichi Mayekawa

Tokyo Institute of Technology
Tokyo Institute of Technology

A comparison of equating methods for developing an item pool under item response theory

In practical applications of the item response theory (IRT), the construction of item pool is the most difficult one. Developments of a calibrated item pool include two procedures: the data collection procedure, which links multiple test forms to each other, and the item parameter calibration procedure, which place item parameter estimates on a common scale. This study considered five types of linking design and four methods for item parameter calibration: characteristic curve method, moment method (Mean/Sigma), concurrent calibration and fixed common item parameter (FCIP) calibration. Simulated data were generated assuming that a small-scale calibrated item pool has already been developed and new items were to be added to the item pool. Using this data, we examined the robustness of these calibration methods in each linking design.

IRT6

#134

IRT

Wulfert P. van den Brink
Gideon J. Mellenbergh
Rudy Ligtoet

University of Amsterdam
University of Amsterdam
Tilburg University

Item response models within a restricted probabilistic domain

Many parametric unidimensional item response theory models have been proposed for modeling binary items scores. Among these models, the choice of a logistic shape, relating the ability of subjects to the probability of a 'correct' response, seems most popular. Less frequently applied shapes of functions are a polynomial function, and a sinus function. A reason for these latter shapes to be less popular might result from their distinct characteristics of being probabilistic over a restricted domain of the ability scale, making the estimation of the parameters more complicated, and as a result, less practical. It will be illustrated how these models are mathematically related. Also results of Bayesian parameter estimation will be presented.

IRT7

#128

IRT, GLM, EDA

Katalin Balázs
Paul De Boeck

K. U. Leuven
K. U. Leuven

Detecting individual differences based on hidden item properties

In several domains of social sciences, such as for example cognitive psychology, educational measurement, item properties (item features, item covariates) are in the focus of research. For binary data, the combination of marginal modeling (GEE1) with logistic regression, the so-called alternating logistic regression (ALR), provides a suitable method to reveal individual differences based on item properties (covariate based random effects). However, often there is no a priori information about item properties. In order to overcome this problem, a cluster extraction algorithm can be used as a preliminary step for the ALR to determine the relevant item properties. The aim of the present study is to investigate whether the latest version of the ADCLUS software applied on the marginal log-odds ratios can be successfully combined with ALR for detecting heterogeneity in data with hidden (latent) item properties. ADCLUS is a model for overlapping clusters based on properties with an additive contribution to associations (such as log-odds ratios), so that the properties in question may overlap. We will report on a set of simulation studies with data generated with different sizes and cluster structures. The efficiency of ADCLUS for this data structure and the diagnostic performance of ADCLUS combined with ALR will be investigated.

#149

IRT

Ying Liu
Jeffrey A. Douglas

University of Illinois at Urbana-Champaign
University of Illinois at Urbana-Champaign

Testing Person Fit of Cognitive Diagnosis Model

In cognitive diagnosis situation, the test-taking behavior of some examinees may be idiosyncratic so that their test scores may not reflect their true cognitive abilities as much as that of more typical examinees. Statistical tests are developed, based on the single-strategy DINA model, to recognize: (1) non-masters of the required attributes who correctly answer the item ("guessing"); and (2) masters of the attributes who fail to correctly answer the item ("slipping"). For a certain person, non-zero probability of aberrant behavior is tested as the alternative hypothesis, against normal behavior as the null hypothesis. The generalized likelihood ratio test statistic has an asymptotic distribution of a 50:50 mixture of a chi-square with 0 df and a chi-square with 1 df. Simulation results are used to investigate: (1) how accurate the statistical tests are to identify normal / aberrant behaviors; (2) how the power of the statistical tests depends on the length of the cognitive test; and (3) how sensitive the statistical tests are to inaccurate estimate of model parameters.

IRT7

#159

IRT

Haniza Yon
Mark D. Reckase

*Educational Testing Service
Michigan State University*

A Comparison Between Two Multidimensional Item Response Theory Approaches to Vertical Scaling

Different algorithms exist for carrying out vertical scaling according to the logic of multidimensional item response theory (MIRT). This study considered the performance of two such algorithms, the Test Characteristic Function (TCF) and Non-Orthogonal Procrustes (NOP) methods, with respect to both real and simulated data. Linking was carried out at two levels, between grades 7 and 8 and between grades 6 and 7. Simulated data were generated for both approximate simple (APSS) and mixed (MS) test structures. The performance of the two methods was compared in terms of linking error on both the discrimination and difficulty parameters, as measured by bias, correlation, and root-mean-square-error (RMSE) metrics. Repeated measures ANOVA analysis showed that APSS data can generally be linked with greater accuracy and that differences between the two methods with respect to item discrimination parameters are minimal. In the case of item difficulty, however, the NOP method yields estimates that are more biased but show higher correlation and lower RMSE values. As a whole, these results imply that vertical scaling is likely to be less accurate in the case of MS data, but that the problem can be alleviated to some degree by using the NOP rather than the TCF method.

#170

IRT

Johan Braeken
Francis Tuerlinckx

*K. U. Leuven
K. U. Leuven*

Flexible latent variable models for teratology

A Generalized Linear Mixed Model (GLMM) for multivariate binary data will be presented that is more flexible towards two common basic assumptions, normality of the latent variable and conditional independency, and this by means of finite mixture distributions and copula functions respectively. Violations of both assumptions may lead to biased inferences.

While the normality assumption puts a constraint on the shape of the distribution of the latent variable, finite mixtures avoid the specification of a parametric form for the latent variable and are capable of fitting a large variety of distributions. The conditional independence assumption states that conditional upon the latent variable (and in addition to the covariates) the outcomes are independent realizations, and thus that the dependence in the data is solely ascribed to this latent variable. Introducing copula functions into the latent variable model can help to provide a more appropriate dependency structure than the one imposed by the conditional independence assumption, while still preserving the model of the univariate marginals.

The model will be applied to a teratology study and will be shown a valuable tool for diagnosis and etiology research in the study of birth defects.

IRT / BSI

#18

IRT, FAC, BSI

Michael C. Edwards

The Ohio State University

A Markov chain Monte Carlo approach to confirmatory item factor analysis

Item factor analysis has a rich tradition in both the structural equation modeling and item response theory frameworks. While great strides have been made in the past three decades in parameter estimation for these types of models, significant limitations remain. The goal of this paper is to examine the feasibility of using Markov chain Monte Carlo (MCMC) estimation methods to estimate parameters of a wide variety of confirmatory item factor analysis models. After providing a brief overview of item factor analysis and MCMC, results from a small feasibility study will be discussed. Where possible comparisons are made to estimates from "gold-standard" estimators such as WLS and MML/EM, but the bulk of the examples focus on models that are problematic for these estimators. In all cases MCMC provided reasonable parameter estimates. Future directions, including software development, model fit, and model extensions, will also be addressed.

IRT / BSI

#96

AAP, IRT

Edward H. Ip
Yuri Goegebeur
Paul De Boeck
Geert Molenberghs

*Wake Forest University
South Denmark University
K. U. Leuven
University Hasselt*

All Unidimensional Models are Wrong, But Some are Useful:
Functional Unidimensionality and Methods of Estimation

Thurstone argued that a universal characteristic of all measurements, be they psychological, social, or whatever, is that they should only measure one attribute. Scientists in psychology and other fields often strive to create instruments that focus on measuring a unidimensional construct. In reality, operational characteristics of even well-developed instruments suggest that the unidimensional and the closely related local independence assumptions are often questionable. The deviation from these assumptions may arise from complexity in the design of the test, as modern tests continue to become more sophisticated. One example is the use of item blocks in educational assessment. Bejar argued that even when there is a dominant dimension within a test, unidimensionality will hold as long as items within the test function in unison. Motivated by this notion of unidimensionality and building upon previous work from this group and others, we propose a functional unidimensional approach that aims to increase the flexibility of IRT models for handling possible deviation from unidimensionality and local independence. Estimation methods including GEE for functional unidimensional models will be discussed.

#136

AAP, BSI

Feifei Ye

University of Pittsburgh

MCMC Methods for Item Response Models in Longitudinal Diagnostic Assessment

Patz and Junker describe a general Markov chain Monte Carlo (MCMC) strategy for Bayesian inference in complex item response theory (IRT) settings to address issues such as non-response, designed missingness, multiple raters, guessing behavior and polytomous test items. This paper extends their strategy to address issues in multidimensional tests with testlets in a longitudinal design of diagnostic assessment. A random testlet effect model is used to account for possible dependence between observables within an assessment task. Growth curve model is used to estimate individual learning curves. The methodology is applied to one example of diagnostic assessment of student learning a content unit of patterns in a 9-week instruction period of pre-algebra, which presents a complex longitudinal assessment setting with designed missingness, polytomous test items in testlets, and multiple time points of assessment. The IRT model is the multidimensional partial credit model embedded in a Bayesian hierarchical framework and inferences are obtained using MCMC methods. Applications of MCMC methods in diagnostic assessment are discussed where the assessment model is complex and the ratio of examinees to parameters is small.

#148

AAP, IRT

Insu Paek
Mark Wilson

*Educational Testing Service
University of California at Berkeley*

Formulation of the Rasch DIF model and its comparison with SIBTEST

Two types of Rasch DIF models are introduced. The first Rasch DIF model is constructed under the hierarchical or multilevel modeling perspective, which is called population regression DIF model, and the second is constructed under the logistic latent trait model with linear constraints perspective, which is called within-logit-mean DIF model. It is shown that these two DIF models are equivalent, being special cases of the random coefficient multinomial logit model. Montecarlo evaluation and real data comparison of the population regression DIF model with SIBTEST were also conducted for their statistical power and Type I error rate. In the 3-parameter logistic item response function (IRF) with a long test (31 items) condition, an interaction between item slope and population mean difference produced inflated Type I error and less power in the population regression DIF model. SIBTEST showed a dependency on item slope in its power. In the Rasch IRF with short test lengths (15- or 8-item test) condition, the population regression DIF model stuck to more closely the nominal Type I error rate than SIBTEST, and interestingly for an extreme case of a single matching item, the population regression DIF model performed well.

Li Cai

*University of North Carolina at Chapel Hill*SEM of Another Flavor: Two New Applications of the Supplemented EM Algorithm

The Supplemented EM (SEM) algorithm augments the original EM algorithm by computing the observed information matrix. In this paper, the SEM algorithm is applied to address two problems in psychometrics. The first problem involves computing the information matrix for item parameters in item response theory (IRT) models fitted using Bock and Aitkin's EM algorithm. This matrix is important for limited-information goodness-of-fit testing, and it can also be used to compute standard errors for the item parameter estimates. For the second problem, it is shown that the SEM algorithm provides a convenient computational procedure that leads to an asymptotically chi-squared goodness-of-fit statistic for the "two-stage EM" procedure of estimating covariance structure models in the presence of missing data. Both simulated and real data are used to illustrate the proposed procedures.

Marika Polak
 Willem J. Heiser
 Mark De Rooij

Leiden University
Leiden University
Leiden University

Testing Single-Peakedness of Item Responses

Several researchers have developed models for single-peaked data. However, for practical researchers it is often difficult to decide whether or not their data are single-peaked. We developed a method for item analysis of single-peaked items based on the criterion of irrelevance by Thurstone and Chave. This is a graphical method for evaluating the "relevance" of a given dichotomous attitude item a , where scale values of all items are plotted against the conditional probability of endorsing another item given that a subject endorses item a . The more the diagram shows a peaked pattern with the peak located at the scale value of item a , the more "relevant" item a is. We generalized this method to polytomous items and quantified the "relevance" by fitting a normal curve. The resulting goodness of fit was used as a test for single-peakedness. Furthermore, a measure of fit for the scale as a whole is suggested. The properties of this method were explored using data generated with a well-developed unfolding model called GGUM. We varied sample size, number of items, item discrimination, and category thresholds. Evidence is presented that shows this method distinguishes single-peaked items from monotonic or cumulative items.

Martijn G. de Jong
 Jan-Benedict E. M. Steenkamp
 Jean-Paul Fox

Tilburg University
Tilburg University
Twente University

Relaxing Measurement Invariance in Cross-National Consumer Research Using a Hierarchical IRT Model

With the growing interest of consumer researchers to test measures and theories in an international context, the cross-national invariance of instruments designed to measure consumer behavior constructs has become an important issue. Consumer researchers now routinely test for measurement invariance using confirmatory factor analytic (CFA) techniques before comparing countries on substantive issues. Yet at least two issues still need to be addressed. First, the ordinal nature of the rating scale is ignored, which has recently been shown to have deleterious effects on the validity of cross-national comparisons. Second, when few, or no items in CFA exhibit metric and scalar invariance across all countries, comparison of results across countries is difficult, if not impossible. We propose to solve these problems using a hierarchical item response theory measurement model. The model takes differential item functioning, including scale usage differences into account. Countries can be substantively compared, even in case of absence of cross-national measurement invariance. An empirical application is provided for the consumer susceptibility to normative influence scale, using a sample of 5,484 respondents from 11 countries on four continents.

IRT / SEM

#150

AAP, IRT, SEM

John T. Willse
Josh Goodman

University of North Carolina at Greensboro
University of North Carolina at Greensboro

Comparison of MIMIC and IRT Based Analyses of Subgroup Differences

This research provides a direct comparison of structural equation modeling (SEM) to item response theory (IRT) in the context of examining subgroup differences on a latent trait. Differences between the SEM and IRT approaches will be highlighted under a variety of data conditions (test lengths, magnitude of group differences, IRT models). A multiple-indicators, multiple-causes (MIMIC) framework will be used for the SEM analysis. MIMIC analyses are based on the regression of latent factors onto group variables. Previous studies have demonstrated that SEM can address traditionally IRT related topics. There are (however) no direct comparisons of MIMIC and IRT analyses for detecting subgroup differences and effect sizes. The current study uses simulated data to compare the relative effectiveness of these two methods of describing subgroup differences. The study is replicated under two IRT model conditions: (1) data generated to fit the two-parameter logistic IRT model (2PL) and (2) data generated to fit the three-parameter logistic IRT model (3PL). The 2PL data condition allows comparison of the SEM and IRT methods where the two approaches have their greatest correspondence. The 3pl data condition allows examination of the impact of a pseudo-guessing parameter on the SEM MIMIC model's performance.

LDA

#38

LDA, BSI

Zhiyong Zhang
Lijuan Wang
John R. Nesselroade

University of Virginia
University of Virginia
University of Virginia

Growth Rate Models and Bayesian Estimation

Generally, *rates of growth* in growth curve models are represented as random slope coefficients analyzed in relatively simple, linear growth functions. Because of the significance of *rates of growth* in understanding dynamic processes, we propose modeling them with a stronger, more versatile approach. We define the *rate of growth* as the first order derivative of the growth function. For example, the rate of growth for the quadratic growth model (first order derivative) is a linear combination of the coefficients of the linear and quadratic terms of the quadratic growth function. If the rate of growth includes two or more growth function coefficients in some combination, they are difficult to estimate using traditional methods, so alternative estimation procedures must be found. Similarly, if the growth rate includes unknown parameters, appropriate estimation procedures are needed. We present a Bayesian method, using Gibbs sampling, for estimating models with such complex growth rates. This method is illustrated by fitting linear, quadratic, and exponential growth functions to simulated and real data. Results support the plausibility and potential value of the general approach. Some implications for understanding growth processes are discussed.

#55

LDA, IRT

Claus H. Carstensen

Leibniz Institute for Science Education (IPN)

Measurement of true change in the IRT framework

The measurement of change typically challenges the researcher with the question of measurement error cumulation and the dependence between a pre-test and a post-test measurement obtained from the same subjects. In the framework on Structural Equation Modelling (SEM), numerous models on latent variables have been introduced for the measurement of change problems. For example, the common variance in measurements from different time points may be modelled by introducing the according paths into the model. As well, a latent variable which directly captures the change to be measured may be defined. In order to employ such models in the framework of Item Response Theory, multiple latent variables with some restrictions have to be estimated. In the presentation, an item response model for true change between two points in time will be introduced. The model is a submodel of the MRCML model and the software ConQuest is used to estimate parameters. In addition, by defining a latent regression model, true correlations between the measure of change and context variables, i. e. treatments or conditions that apply to the respondents, may be modelled. The presentation will be illustrated by some results from a longitudinal study which is part of the extension to the PISA 2003 study in Germany.

LDA

#60

LDA, MVA

Heungsun Hwang
Yoshio Takane
Wayne S. DeSarbo

*HEC Montréal
McGill University
The Pennsylvania State University*

Fuzzy Clusterwise Growth Curve Models via Generalized Estimating Equations

The growth curve model has been a useful tool for the analysis of longitudinal data. It helps investigate the average change in trajectories of repeated responses and its relations to time-invariant explanatory variables. However, the growth curve model is designed only for an aggregate-sample analysis based on the assumption that all individuals come from a single homogenous population. Thus, it is not suitable for investigating the existence of heterogeneous subgroups in the population, which involve qualitatively distinct patterns of trajectories. In this paper, the growth curve model is generalized to a fuzzy clustering framework so as to take into account such group-level heterogeneity in trajectories of change over time. Moreover, the proposed method estimates parameters based on generalized estimating equations thereby enabling to specify a more variety of covariance structures among repeated responses than the traditional growth curve model. An empirical application concerning the antisocial behavior of a sample of children is presented to illustrate the usefulness of the proposed method.

#119

LDA, SEM

Lijuan Wang
Zhiyong Zhang
John J. McArdle

*University of Virginia
University of Virginia
University of Southern California*

Alternative Structural Models for Ceiling Effects in Longitudinal Data Analysis

Score limitation at the top of a scale is commonly termed "ceiling effect". Ceiling effects can lead to serious artifactual estimates in the parameters in most forms of data analysis. For example, in longitudinal mixed models, there is substantial estimation bias in the mean vector and covariance matrix of the random-effects parameters. The current study examines different alternative methods including the Bayesian hierarchical Tobit model to deal with ceiling effects in the longitudinal modeling framework. Data was simulated based on a latent growth curve model with T=5 occasions. The proportion of ceiling data was manipulated by using different thresholds $C=[10\%-40\%]$, and estimated parameters were examined for R=100 replications. The results showed that the hierarchical Tobit model performed very well, recovering the true population parameters even in cases where the percentage of ceiling data in the last occasion is larger than 40% and the sample size is as small as 30. The results of applying the Hierarchical Tobit model into an empirical aging study and some other related issues were also discussed.

MDS

#68

MDS

Akinori Okada
Tadashi Imaizumi

*Rikkyo (St. Paul's) University
Tama University*

Two-Mode Three-Way Asymmetric Multidimensional Scaling with Dominance Points

The purpose of the present study is to introduce (a) a model of multidimensional scaling to analyze two-mode three-way asymmetric proximities (object X object X source) where proximities among objects for each source are not necessarily symmetric, and (b) an associated nonmetric algorithm to fit the model. Several models for analyzing two-mode three-way asymmetric proximities have been introduced. In most of the models already had been introduced by the authors, each object is represented by a point and a circle (sphere hypersphere) or by a point and an ellipse (ellipsoid, hyperellipsoid) in a multidimensional Euclidean space. The radius of the circle or the length of the axis of the ellipse represents the asymmetry of the corresponding object in the proximity relationships. In the present model, each object is represented as a point in a multidimensional Euclidean space, and does not have a circle or an ellipse, and each source is also represented as a point in the same multidimensional space. The asymmetry in the proximity relationships among objects for a source are represented by a point, called a reference point, representing the source. An application of the present asymmetric multidimensional scaling will be shown.

MDS

#77

MDS

Willem J. Heiser
Laurence E. Frank

*Leiden University
Leiden University*

Network Representations of City-Block Models

City-block models for similarity always allow network representations that reproduce the same distances as the unique coordinate representation. A rule to construct such networks is given, based on additivity of city-block distances across sequences of intermediate points along monotonic trajectories in space. The paper also defines the concept of internal node, which helps in reducing the complexity of networks and in making them better interpretable. The general graph construction rule and definition of internal nodes also apply to the distinctive features model, the common features model (additive clustering), as well as to hierarchical trees, additive trees, and extended trees. Additivity is the key property that makes the city-block metric so versatile and causes a basic unity of dimensional, hierarchical and featural representations of similarity.

#83

MDS

Marie-Ève Provost
Gilles Caporossi
Mehran Ebrahimi
Anne-Laure Saives

*HEC Montréal
GEIRSO & HEC Montréal
GEIRSO & HEC Montréal
GEIRSO & HEC Montréal*

Study of the perception of drugs by health professionals

Therapeutic observance is defined by the action to take drugs as prescribed by health professional. Unfortunately, non-observance of the prescriptions is frequent and the problems and consequences surrounding it should not be neglected. A difference between the perceptions of health professionals could have a negative effect on therapeutic observance. A contradiction between the perceptions of different health professionals, principally if reflected in the indications given to the patients, can easily confuse the patients or affect the credibility of the professionals and thus lead to non observance of the treatment. The aim of this study is to compare the perceptions of health professionals on different drugs. We are principally interested by the perceptions of health professionals that are directly in contact with the patients: physicians, pharmacists and nurses. Therefore, the data were collected from these three groups of professionals. Their perception was analyzed by using multidimensional scaling. Nine drugs, analgesics and nonsteroidal anti-inflammatory, were used in our study. Methods for the analysis of several matrices were used and the results were analysed in order to highlight the similarities and differences between the various groups.

#84

MDS

Adil Allillat
Gilles Caporossi
Sihem Taboubi

*HEC Montréal
GEIRSO & HEC Montréal
GEIRSO & HEC Montréal*

Evaluating the parameters used in MDS

The MDS would enable us to represent a corpus of stimuli in a space of low dimensions. Depending on the distance measure used or the error function (stress), the perceptual map drawn changes significantly. This study aims at helping the researcher who wants to find the best parameter and, consequently, improving the perceptual map. In order to reduce the number of attributes that are not shared by the different products used in this study, it is based upon 8 brands of acetaminophen caplets 325mg; technically, all the products are similar. For this study, 50 respondents are involved (mostly students). Furthermore, in this study we have decided to test the influence of the distance and the error function parameters. We have opted for the choice of 2 forms of distances and error functions. At the end of the questionnaire 4 perceptual maps are presented to the respondent (each corresponding to a given set of parameters) that will have to say which one best fits his perception. In the presentation, we will describe the study and discuss its results.

MVA1

#10

MVA

Denis Larocque
Jaakko Nevalainen
Hannu Oja

*HEC Montréal
University of Tampere
University of Tampere*

A Weighted Multivariate Sign Test for Cluster Correlated Data

We consider the multivariate location problem with cluster correlated data. A family of multivariate weighted sign tests is introduced for which observations from different clusters can receive different weights. Under weak assumptions, the test statistic is asymptotically distributed as a chi-squared random variable as the number of clusters goes to infinity. The asymptotic distribution of the test statistic is also given for a local alternative model under multivariate normality. Optimal weights maximizing Pitman asymptotic efficiency are provided. These weights depend on the cluster sizes and on the intracluster correlation. Several approaches for estimating these weights are presented. Using Pitman asymptotic efficiency, it is shown that appropriate weighting can increase substantially the efficiency compared to a test that gives the same weight to each cluster.

#33

MVA

Haruhiko Ogasawara

Otaru University of Commerce

Approximations to the distribution of the sample coefficient alpha under nonnormality

Approximate distributions of the sample coefficient alpha under nonnormality as well as normality are derived by using the single- and two-term Edgeworth expansions up to the term of order $1/n$. The case of the standardized coefficient alpha including the weights for the components of a test is also considered. From the numerical illustration with simulation using the normal and typical nonnormal distributions with different types/degrees of nonnormality, it is shown that the variances of the sample coefficient alpha under nonnormality can be grossly different from those under normality. The corresponding biases and skewnesses are shown to be negative under various conditions. The method of developing confidence intervals of the population coefficient alpha using the Cornish-Fisher expansion with sample cumulants is presented.

#42

MVA

Yoshio Takane
Sunho Jung

*McGill University
McGill University*

Regularized Partial and/or Constrained Redundancy Analysis

This paper discusses methods of incorporating a ridge type of regularization into partial and/or constrained redundancy analysis (PRA, CRA, and PCRA). The usefulness of ridge estimator in reducing MSE (mean square error) has been demonstrated when columns of X , matrix of predictor variables, are nearly collinear, and consequently the ordinary least squares estimator is poorly determined. In this paper, the procedure of ridge estimation is extended to PRA, CRA, and PCRA, where a ridge estimator for reduced rank parameters is obtained by minimizing the ridge least squares criterion. An optimal value of a regularization parameter is found by K -fold cross validation. Illustrative examples are given to demonstrate the usefulness of the method in practical data analysis contexts.

#43

MVA

Yoshio Takane
Haruo Yanai

*McGill University
National Center for University Entrance Examinations*

On Ridge Operators

Let X be an n by p matrix, and define $R_X(\lambda) = X(X'X + \lambda P_X)^{-1}X'$ which we call ridge operator, where λ is a nonnegative constant (called ridge parameter), and $P_X = X'(XX')^{-1}X$. In this talk we discuss various properties of $R_X(\lambda)$ and derive additive decompositions of this matrix similar to those of $P_X \equiv R_X(0) = X(X'X)^{-1}X'$, orthogonal projector onto the range space of X . These properties and decompositions are useful, especially in ridge estimation of reduced rank regression analysis, and multiple-set canonical correlation analysis, as demonstrated in this talk.

MVA2

#25

MVA, OTR

Pieter M. Kroonenberg

Leiden University

Multiple imputation in three-mode analysis: A research program

One of the ways to handle missing data is via multiple imputation, i. e. generating several data sets with differently imputed values for the missing data. The basic idea propose in the paper is to use generalised procrustes analysis to evaluate the stability of the parameters. Attention is paid to the possibilities and impossibilities of using multiple imputation in three-mode analysis. Much of the research is still in its infancy and primarily ideas are presented rather than firm results.

#112

MVA, FAC

Kohei Adachi

Osaka University

Joint Procrustes Analysis of Semantic Differential Data

A set of semantic differential data is expressed as $\{\mathbf{X}_k; k = 1, \dots, N\}$, where an $n \times m$ matrix \mathbf{X}_k contains the ratings of subject k on m adjective scales for n concepts. We consider approximating \mathbf{X}_k as $\mathbf{X}_k \cong \mathbf{F}_k \mathbf{A}_k'$, where \mathbf{F}_k ($n \times p$) and \mathbf{A}_k ($m \times p$) represent the perceptual structure and the semantic structure, respectively, which subject k has in mind. The approximation $\mathbf{X}_k \cong \mathbf{G}_k \mathbf{B}_k' = \mathbf{G}_k \mathbf{S}_k \mathbf{S}_k^{-1} \mathbf{B}_k'$ is attained by principal component analysis, and \mathbf{F}_k and \mathbf{A}_k are thus given by $\mathbf{F}_k = \mathbf{G}_k \mathbf{S}_k$ and $\mathbf{A}_k = \mathbf{B}_k \mathbf{S}_k^{-1}$, with \mathbf{S}_k a suitable nonsingular matrix. In order to identify \mathbf{S}_k , we assume that \mathbf{F}_k 's are similar among k and \mathbf{A}_k 's are also similar, and minimize

$$\phi(\mathbf{S}, \mathbf{F}, \mathbf{A}) = \frac{1}{n} \sum_{k=1}^N \|\mathbf{G}_k \mathbf{S}_k - \mathbf{F}\|^2 + \frac{1}{m} \sum_{k=1}^N \|\mathbf{B}_k \mathbf{S}_k^{-1} - \mathbf{A}\|^2$$

over $\mathbf{S} = [\mathbf{S}_1, \dots, \mathbf{S}_N]$, \mathbf{F} and \mathbf{A} , with \mathbf{F} and \mathbf{A} representing the perceptual structure and the semantic structure common over subjects. The above minimization, for which we reparameterize \mathbf{S}_k by its singular value decomposition, is coined as joint Procrustes analysis (JPA). Using the ratios

$$\frac{\sum_{k=1}^N \|\mathbf{F}_k - \mathbf{F}\|^2}{\sum_{k=1}^N \|\mathbf{F}_k\|^2} \quad \text{and} \quad \frac{\sum_{k=1}^N \|\mathbf{A}_k - \mathbf{A}\|^2}{\sum_{k=1}^N \|\mathbf{A}_k\|^2}$$

obtained from the result of JPA for double-centered \mathbf{X}_k , we can compare the largeness of individual differences in perceptual structure against that in semantic structure.

#151

OTR, FAC, CCC

Sungjin Hong

University of Illinois at Urbana-Champaign

Fuzzy Hidden Parafac2

Parafac is a three-way component model whose component weights are uniquely identifiable under minor conditions. Parafac2, as a variant of Parafac (sometimes called Parafac1 for distinction), relaxes the invariance condition of component weights for one of three data modes, still yielding a unique decomposition of three-way data. What is invariant on the relaxed mode is angles between components instead of the component weights themselves. Parafac2 is suited well to multiple-groups factor analytic problems in that the relaxed invariance condition allows for fitting multiple groups of distinctive observational units by reweighted invariant component loadings. Multiple-groups component/factor analysis typically assumes known grouping of observation units (e. g. , ethnic backgrounds, age cohorts, etc.). However, it would be very desirable in some cases to consider the grouping itself as unknown model parameters to optimize (as in k-means clustering). Unlike separate modeling of data reduction followed by clustering in a tandem analysis (e. g. , k-means clustering on reduced-dimensional subject space), Fuzzy Hidden Parafac2 (FUZHIP2) simultaneously estimates component weights and (non-crisp) grouping memberships, minimizing a unified fit function. From the mixture-modeling perspective, FUZHIP2 can be considered as a constrained component mixture modeling of two-way data by a (reweighted) invariant component loading pattern and hidden membership. An alternating least squares algorithm will be presented for parameter estimation in which group membership is iteratively estimated by a fuzzy k-means step, given all other model parameters. Simulated data are used to assess performance of the algorithm.

MVA2

#173

MVA

Joachim Vandekerckhove
Francis Tuerlinckx

K. U. Leuven
K. U. Leuven

Fitting the diffusion model to experimental data: Methods and tools

The diffusion model can be used for the simultaneous analysis of RT and accuracy data. By implementing design matrices, substantive restrictions can be imposed on the model's parameters across conditions, in a flexible way. This way, it becomes possible to fit a model in which parameters are regressed onto predictors. We also demonstrate a method to handle outliers, which consists of a preprocessing stage (to remove fast guesses) and a mixture model with the diffusion model as one component and a distribution representing the contaminant data points as another component. The talk includes a brief demonstration of an easy-to-use graphical software tool that should allow any user to perform diffusion model analyses. The basic MATLAB algorithm makes use of several techniques which boost its efficiency, some of which we will briefly discuss. The program also lets users construct queues of models in order to compare the fit of different models.

MVA/VCA

#32

MVA, VCA, LDA

Peter Ebbes
Ulf Böckenholt

The Pennsylvania State University
McGill University

A new approach to account for general regressor-error dependencies in two-level models with an application to estimating the effect of student and school characteristics on test scores

Recently several studies have appeared in the psychometric literature that stress the importance of investigating regressor-error dependencies in multilevel models. Much of this work has focused on traditional approaches to account for such dependencies, such as fixed effects estimation, the Hausman test, and instrumental variables approaches. However, these methods are not without limitations. In this paper we propose a new method to test and solve for general regressor-error dependencies in two-level models. Extending the latent instrumental variables approach proposed by Ebbes et al. , this new method has the important advantage that regressor-error dependencies, at both level-1 and level-2, can be taken into account without requiring additional (instrumental) variables. We illustrate our method on simulated and real data. The application focuses on effects of student and school characteristics on test scores. We find that the available explanatory variables are correlated with the model errors, a result that cannot be accounted for by such traditional methods as the Hausman test and fixed effects estimation. We show that such dependencies lead to biased estimates and erroneous decision making and that these problems can be avoided by applying the proposed approach.

#52

VCA

Wen Luo
Oiman Kwok

Texas A & M University
Texas A & M University

Impacts of Ignoring a Crossed Factor in Analyzing Cross-classified Multilevel Data:

A Monte Carlo Study

In the past two decades, many efforts have been made in modeling cross-classified data. However, in real educational researches, cross-classified models are rarely used. A quick look into ERIC database from 2004 to 2005 shows that around 60 studies used multilevel models but only one study employed cross-classified models. Many researchers ignore the cross-classified structure of their data and just use hierarchical linear models because the cross-classified model could be very complex. This Monte Carlo study examines the impact of ignoring a crossed factor on variance component estimates, fixed parameter estimates, and the estimated standard errors of fixed parameters. Preliminary results show that the variance of the ignored crossed factor will be redistributed among the other levels, causing the variance component estimates to be biased. The degree of biasness is related to the magnitude of the variance and the average cluster size of the ignored crossed factor, and the correlation of the two crossed factors. The fixed parameter estimates will not be biased, but the related standard errors will be biased downwards, causing inflated type I error rates.

MVA/VCA

#73

MVA, VCA

Maria-Pia Victoria-Feser
Samuel Copt

University of Geneva
University of Sydney

A robust alternative to the F-test in ANOVA and mixed linear models

The F-test is not robust to model deviations! Indeed, classical inference in ANOVA or mixed linear models (including MLE, REML t-test, F-test) is based on the assumption that the data are generated from a multivariate normal distribution with a constrained covariance matrix. When this is not the case, like for example when the model is a mixture between two different multivariate normal distributions, classical inference fails. The consequence of such model deviations is that the classical estimators become biased and for testing procedure, that the actual level of the test can be really far from the nominal one (5%). We propose here an alternative estimation and testing procedure for ANOVA and mixed linear models that is robust to (small) model deviations of any type. The estimator is taken in the class of *S*-estimators adapted to the case of constrained covariance matrices. We also use it as a diagnostic tool to detect outlying observations. We also consider robust inference for multivariate hypotheses as an alternative to the classical *F*-test by using a robust score type test statistic based on the *S*-estimator for constrained covariance matrices and study its properties by means of simulations and the analysis of a real dataset.

ODS / MDS

#85

MDS

Gilles Caporossi
Sihem Taboubi

GERAD & HEC Montréal
GERAD & HEC Montréal

A Variable Neighborhood Search Algorithm for Multidimensional Scaling in Arbitrary Norm

The Variable Neighborhood Search (VNS) metaheuristic was introduced less than 10 years ago and was already successfully applied to a wide variety of problems of combinatorial or global optimization. By its structure, VNS is well suited to solve continuous optimization problems such as the Multidimensional Scaling. The VNS principle is to alternate local search and increasing magnitude perturbations. For our study, local search is handled by a gradient descent. We then propose methods to get out of local optima by the use of two levels of perturbation. One level aims at improving the solution by small strategic perturbations and may be seen as an improved version of the local search. The second aims at diversifying the search. Two methods are proposed for each level of perturbation. All the 4 combinations are compared together and to the multistart approach. The implementation of these strategies leads to a new algorithm for the multidimensional scaling that may handle arbitrary norm for distance measure as well as various error measures. The algorithm was tested on a classical benchmark for MDS (the morse code dataset) and the results are described and discussed.

#142

ODS, OTR

Anli Lin
Don Meagher
Eugene Bowles
Christina P. Stellato

Harcourt Assessment, Inc.
Harcourt Assessment, Inc.
Harcourt Assessment, Inc.
Harcourt Assessment, Inc.

Using the Kernel Method for the Essay Test Equating of the Pharmacy College Admission Test

The Kernel equating method is a powerful, accurate and flexible equating method. It is a more accurate method than other methods and is a more consistent technique over different equating designs. In this study, we compared the Kernel equating method with different methods, e. g. , linear equating, equipercentile equating and mean / mean equating. Also, we discussed how to choose suitable equating method based on data distributions. Generally, the result of linear equating is very close to that of Kernel equating with small differences in low and high scores. Generalization validity test shows the kernel equating is stable over different data.

ODS / MDS

#143

ODS

Daniel R. Lawrence

*Rochester Institute of Technology*Getting a Dichotomous Incidence Item to Determine the First Solution
in a Dual Scaling Analysis of Paired-Comparison Data

Psycho-physical experiments often involve paired comparisons. For example, in imaging science, studies sometimes require "judges" to do paired comparisons of images with a subsequent dual scaling analysis of these data intended to reveal dimensions along which the images are subconsciously evaluated (e. g. , contrast, sharpness, and/or overall quality). If respondents comprise different demographic groups, such as "trained" and "untrained" judges, one could detect differences between the two groups in the multiple dimensions by simply applying a two-sample "t-test" to the subject scores generated in the analysis. But to find out which stimuli (images) do most to elicit different responses from the judges based on group affiliation—that is, which stimuli project most strongly onto an axis determined by the incidence variable—the dichotomous "trained-untrained" item must be made to define that axis. To that end, the incidence data must be included in the analysis of the dominance data. A relatively simple procedure for restructuring the matrix of paired-comparison data to include the incidence variable will result in weights being assigned to the stimuli that enable the investigator to identify those stimuli or items whose responses are most affected by group affiliation of subjects.

#163

MDS

Jesse Spencer-Smith

*University of Illinois at Urbana-Champaign*Detecting and Characterizing the Presence of Curvature in Scaling Data

Most applications of multidimensional scaling (MDS) assume that proximities are related to distances in a flat geometry (e. g. Euclidean or city block). Some data sets, although seemingly well-fit by traditional MDS, actually are properly modeled as residing in curved spaces. I demonstrate how traditional MDS absorbs and obscures curvature. I review new tests which indicate the presence of curvature, and describe a tool which graphically details curvature in data.

OTR: Applied

#146

OTR

Robert A. Henson
Jonathan L. Templin*University of North Carolina at Greensboro
University of Kansas*The DINO: A Disjunctive Model for Skills Assessment

Cognitive diagnosis models (CDMs) are a set of constrained latent class models in which classes are determined by mastery or non-mastery of a set of dichotomous latent variables, or skills. Because cognitive diagnosis models define groups based on mastery/nonmastery of skills, a profile containing the chances (i. e. , posterior probabilities) that each of the skills have been mastered can be estimated. Because of its simplicity of item parameterization in addition to its easy interpretation, one of the more common CDMs is the DINA model (Deterministic Input; Noisy "And"). The DINA model assumes an examinee will only have a high chance of correctly responding to an item when all required skills measured by that item have been mastered (i. e. , a conjunctive model). Although such a model works well in educational assessment, when measuring psychological constructs this model may be unreasonable. We present a new model, the DINO (Deterministic Input; Noisy "Or"), and its estimation as a disjunctive adaptation of the DINA. Then, we provide an example of an analysis using the DINO for diagnosis of a psychological disorder, pathological gambling.

OTR: Applied

#161

OTR

Chiu-Hsia Huang

*National Ping-Tung Education University*The WISC-III analysis and the Effectiveness of TWLEST for LD in Taiwan, R. O. C.

This aim of this study was to investigate the Wechsler Intelligence Scale for Children -Third Edition (WISC-III) analysis for children with learning disabilities (LDs) in Taiwan, R. O. C. ; in addition, to explore how Twelve-Week-Life-Education-Story-Telling-Treatment (TWLEST) to remediate and even ameliorate LDs' academic performances, learning attitudes, motivation, and social-emotional behaviors. The major findings of this investigation were as follows:

The average WISC-III scale of 32 participants was: FIQ (78. 10), VIQ (80. 76), PIQ (80. 10), VCI (82. 35), POI (83. 12), FDI (82),and PSI (85. 73). The result was indicated that there was no significant difference for all participants, even though in both individual groups including the experimental group and control group. In addition, FIQ was 70-85 was 41. 4%, 86-110 (31%), below 70 (27. 6%), and none was above 110. The discrepancy between VIQ and PIQ ≥ 20 for 32 participants was 27. 6%. However, the discrepancy between VIQ and PIQ ≥ 15 was 41. 4%. VIQ>PIQ ≥ 20 was 17. 24%; VIQ<PIQ ≥ 20 was 10. 34%. WDI>0. 20 was 30. 8%. After TWLEST, 59% experimental participants agreed, even 12. 5% strongly agreed, with their significant progresses not only in their academic performances, learning attitudes, motivation, social-emotional behaviors, and other perspectives. In addition, 62. 85% resource teachers agreed with TWLEST impact LDs' learning outcomes.

#168

OTR

Xiang Bo Wang

*The College Board*Investigating SAT Essay Rating Consistency: An Overview from Six SAT Administrations in 2005

As of December 30, 2005, six administrations of the SAT Reasoning Test™ had taken place, in March, May, June, October, November and December of 2005, respectively. How consistent had the rating of SAT essay readers been across five categories of essay prompts throughout six SAT administrations? Six important conclusions can be drawn. First, the distributions of five categories of essay prompts were found to be highly similar across the six SAT administrations in 2005, indicating sound test spiraling and management. Second, consistently high levels of exact matches and adjacent agreements between readers 1 and 2 were found across all essay prompts throughout the six administrations. Third, on the average, only about 6. 3% of essays needed adjudications. Fourth, no particular category of essay prompts was found to cause any unusually high rates of adjudication, which signifies highly similar quality of essay prompts and/or comparable scoring standards and practices across different prompts. Fifth, adjudications occurred approximately equally between female and male examinees' essays at both the national and essay prompt levels. Finally, based on the tendencies of adjudicated scores in relation to first two readers' scores, most adjudicated essays seemed to have been well placed, serving the function of mediating significant differences in the first two essay scores just as designed.

OTR: Applied IRT

#26

OTR

Wolfgang Viechtbauer

*University of Maastricht*A new conceptualization of moderator tests in meta-analysis and its implications

One of the greatest benefits of a meta-analysis is the potential to examine how the effect sizes are influenced by the studies' characteristics, methods, and procedures. Whether one should use fixed- or random/mixed-effects models for such moderator analyses has been discussed extensively in the literature. Simulation studies have revealed that the Type I error of moderator tests can become severely inflated when using fixed-effects methods. However, a review of published meta-analyses indicates that practitioners seem as uncertain as ever in their choice of model. To bring some clarification to this ongoing debate, the Type I error of moderator tests was derived analytically and shown to depend on the true underlying model generating the effect sizes and the type of inference desired (the population of studies to which we want to generalize the results). Most importantly, the necessity to use random/mixed-effects models becomes imminent if we acknowledge that the values of the moderator variables are not actually determined by the analyst, but should be thought of as random themselves. However, since the moderator variables are typically not normally distributed, this conceptualization leads to a discussion about the robustness of moderator tests when the random-effects distribution is misspecified.

#123

OTR

Ali Ünlü

*University of Graz*Generalized Knowledge Space Theory: An Extension of the Basic Local Independence Model to Incorporate Local Dependence and Covariate Information

Knowledge space theory represents a qualitative type of test theory in which the basic local independence model (BLIM) is fundamental. The BLIM can be viewed as a constrained latent class model and is based on the assumption of local independence. In this paper, we give a generalization of the BLIM which allows for local dependence among the indicators given the knowledge state of an examinee and/or for the incorporation of covariates. This general probit regression latent class model with random effects has a number of important special cases which are also briefly discussed.

#145

OTR

Jonathan L. Templin
Robert A. Henson*University of Kansas*
*University of North Carolina at Greensboro*The Random Effects Reparameterized Unified Model:
A Constrained Finite Mixture Model for Skills Diagnosis and Psychological Assessment

Models for skills diagnosis (also called cognitive diagnosis models) are special cases of latent class models in which classes are determined by mastery or non-mastery of a set of categorical latent variables, or skills. Because such models define groups based on mastery/nonmastery of a set of skills, a profile containing the probability that each of the skills have been mastered can be estimated. In practice, however, many applications of such models suffer from poor fit, potentially due to violations of the conditional independence assumptions imposed by the latent class framework. In an attempt to generalize the modeling framework, and weaken the assumption of local independence, a new model was created. The new model (called the Random Effects Reparameterized Unified Model, RERUM) parameterizes the association between items conditional on skill pattern, enabling users to gain additional insight into model results. In this talk, I provide a description of the RERUM with an emphasis how it differs from other models for skills diagnosis. An application to a real data set is presented to demonstrate the types of information that can be obtained from the RERUM model parameters.

OTR: IRT related

#2

OTR

Todd E. Bodner

*Portland State University*On linear composites of dependent effect sizes

Meta-analysts frequently encounter studies reporting more than one effect size between the two conceptual variables of interest. For example, a study on the effectiveness of two training programs may have two measures of job performance, yielding two estimates of the effect of training type on job performance. In such cases, the meta-analyst may deem the two effect sizes as "conceptual replications" of the same relationship and may wish to include information from both estimates in the meta-analysis. The composite (i. e. , weighted average) effect size method is a suggested procedure for extracting a single effect size from a study with several dependent effect sizes for use in a meta-analysis. This paper formally evaluates the method. After contrasting the difference between two forms of dependence (i. e. , stochastic and structural), mathematical results show that the method 1) implies relationships among the variables never likely to be true and 2) generally will not produce an effect size in the desired metric. The second result implies that meta-analyses of both effect sizes and composites of dependent effect sizes violate the implicit principle that the analyzed effect sizes are "of like kind. " Together, these results seriously question further use of the method.

#19

OTR

Ken Kelley

Indiana University

Sample Size Planning for the Squared Multiple Correlation Coefficient:
Accuracy in Parameter Estimation Via Narrow Confidence Intervals

In the behavioral, educational, and social sciences, one of the most commonly used statistical methods is multiple regression. In order to avoid "embarrassingly large" confidence intervals around the population squared multiple correlation coefficient (SMCC), a sample size (SS) procedure is developed with the goal of achieving accurate parameter estimates by obtaining sufficiently narrow CIs. Narrow CIs lead to accurate estimates by homing in on the corresponding parameter values. One approach developed determines SS so that the expected width of the confidence interval will be sufficiently narrow. A modified procedure is also developed that allows for a desired degree of certainty that the obtained interval will be sufficiently narrow. SS planning using the AIPE approach is more difficult for the SMCC than for many other effect sizes, especially when employing a desired degree of certainty that the obtained confidence interval will be no wider than desired. These difficulties stem from the fact that (a) CIs around the population SMCC require confidence limits that depend on a noncentral F -distribution, (b) the population SMCC is systematically overestimated by its sample value, (c) and there is a nonmonotonic relationship between the size of the effect and the width of the corresponding CI. These difficulties are overcome and algorithms are developed that lead to the necessary SS.

#63

DIF

Sungwon Ngudgratoke
Dipendra Subedi*Michigan State University
Michigan State University*

Assessment of Differential Item Functioning in Testlet-Based Items:
A Generalization to 3PL Testlet Response Model

The objective of this study is to evaluate the power of the likelihood ratio test and Wald statistic test which are newly proposed procedures for detecting differential item functioning (DIF) in testlet-based tests. This simulation study is an extension of the one-parameter model-based DIF approach proposed by Wang and Wilson to investigation of the three-parameter testlet response model developed by Wainer, Bradlow, and Du. As statistics used to test the null hypothesis of no DIF in testlet-based tests, the likelihood ratio test and Wald statistic are computed using the posteriors drawn from Bayesian estimation. While the number of items is fixed (e. g. 36 items, 7 testlets in this study), four independent factors, sample size, magnitude of DIF, the proportion of DIF items, and variance of the random effects for testlets are manipulated. For each experimental condition, the results are summarized to assess the recovery of the DIF items by calculating the power rates and Type I error. Finally, the results are compared across independent factors to evaluate the extent to which each factor contributes for power rates and type I error.

SEM1

#9

SEM

Takahiro HOSHINO
Kazuo SHIGEMASU

The University of Tokyo
The University of Tokyo

Doubly Robust Weighted Estimating Equation for Covariate Adjustment in Structural Equation Modeling

Due to the difficulty achieving a random assignment, quasi-experimental or observational study design is frequently used in the behavioral and social sciences. If the nonrandom assignment depends on covariates, multiple group Structural equation modeling that includes regression function of the dependent variables on the covariates can provide reasonable estimates, under the condition of correct specification of the regression function. However, it is usually difficult to specify the correct regression function, because the dimension of dependent variables and that of covariates are usually large. In this study, we propose an propensity score weighting type estimation method for marginal multiple group Structural Equation Modeling. The proposed estimator is consistent if either, but not both, the regression function or the assignment mechanism is correctly specified. Efficiency of the proposed estimator is also considered.

#40

SEM

Fridtjof W. Nussbeck
Michael Eid
Delphine Gross
Christian Geiser
Tanja Lischetzke

University of Geneva
University of Geneva
University of Geneva
University of Geneva
University of Geneva

Conceptions of Method Factors in CFA-MTMM Models:
A Latent Regression, a Latent Difference, and a Latent Means Approach

Recently it has been shown that the choice of confirmatory factor analytic Multitrait-Multimethod models such as the CTCU, CTUM, CTCM, and CTC (M-1), should be based on theoretical considerations taking method-properties into account. This decision depends on the interchangeability of methods or their structural independence. These theoretical considerations determine the number of method factors, if they are correlated and their interpretation. Yet, method factors may be conceived as a residual of a latent regression (latent regression approach), as a latent difference score (latent difference approach), or as the deviation of a latent means score. This presentation shows how the three conceptualizations can be formulated for the standard MTMM data situation measuring three traits with three methods using multiple indicators. An empirical application will illustrate the three modeling strategies and show how the different method factors can be related to external variables explaining these effects.

#50

SEM

Jorge González
Paul De Boeck
Francis Tuerlinckx

K. U. Leuven
K. U. Leuven
K. U. Leuven

A Double-Structure Structural Equation Model for three-mode data

Individual differences and situational differences are important phenomena in the study of human behavior. The regular structural equation model (SEM) considers latent person variables to account for individual differences, but it does not consider a latent structure for situational differences. In this paper a general version of the SEM called Double-Structure Structural Equation Model (2sSEM) is presented which incorporates a latent structure for situational differences besides of for individual differences. The model is evaluated using a three-mode data set of persons by situations by responses about emotions.

SEM1

#62

SEM

Keith A. Markus

*John Jay College, CUNY*Causation de novo versus causation in sequence:
Implications for potential-response theory applied to linear causal models

Potential response formulations of causal effects analyze causal effects as differences between counterfactual observations that represent hypothetical ideal manipulations of the causal variable. These formulations involve causation de novo, in which causal effects contrast two counterfactual responses. The paradigmatic instance involves an experimental manipulation in which the causal variable has no value prior to entering the experimental task. One standard causal interpretation of linear model effect coefficients involves causation in sequence, where the causal effect represents the change from the previous value of the response variable. The paradigmatic instance involves an intervention designed to change one variable as a means of changing another. These differing formulations of causal effect differ in deductive strength and thus in meaning. In particular, the de novo formulation of the interpretation of a linear effect coefficient as a causal effect involves at least two counterfactuals whereas the in-sequence formulation requires only one counterfactual. If the counterfactuals remain independent, then the two formulations converge. If not, then they can differ. One instance of divergence involves ratcheted causal mechanisms that only permit manipulation in one direction. Such differences have implications for interpreting empirical estimates of linear causal effects in substantive applications.

SEM2

#69

SEM

Eisuke Segawa

University of Illinois at Chicago

Jungwha Lee

*University of Illinois at Chicago*Mediation analysis with variables in different levels: Bayesian and non-Bayesian approaches

Mediation analyses involve a set of three variables, Y (outcome), X (covariate), and M (mediator). The levels of Y, X, and M can be different, e. g. , Y and M are measured in the county but X in the state level. Although Krull and MacKinnon presented such mediation analyses, they solved two multilevel regressions, and then combined two results to compute the direct and indirect effects. Their approach is not ideal if the residuals of the two regressions are correlated. SEM solves the two regressions simultaneously and can model the correlated errors. However, the current software of SEM does not allow regressing Y on a variable in a different level. GLLAMM and WinBUGS can solve the between-level mediation problems by simultaneously solving the two regressions. GLLAMM and WinBUGS are compared to the Krull and MacKinnon approach through both simulated and real data analyses. A number of smoking cessation programs is Y (county level), state expenditure on tobacco cessation is X (state level), and the priority of community leaders on smoking cessation is M (county level) in our Helping Young Smokers Quit study. Y and X are observed but M is a latent variables.

#75

SEM

Akihiro Saito
Hideki Toyoda*Waseda University
Waseda University*A new group AHP model using SEM and its application

The analytic hierarchy process (AHP) is the method of decision making and the group AHP is an expansion of AHP to a group. It can be used in various scenes (e. g. marketing and policymaking). Firstly, the present study proposes a new group AHP model, using the notation of structural equation modeling (SEM). It is shown that commonly available computer programs for SEM, such as CALIS, can be used to estimate the weights of several criteria and alternatives for group AHP models. This method has several important features: (1) Various statistics can be used in order to evaluate these weights. (2) Evaluation of accuracy for weight estimation is possible. Secondly, we apply latent structure analysis to this new model. Through this application, we find that AHP can be used as the method of analysing decision making process and we can find some people has unique decision making process. Therefore, it is concluded that the proposed method has adequate utility.

SEM2

#86

SEM

Donna L. Coffman

*The Pennsylvania State University*Consequences of violating the parameter drift assumption in Covariance Structure Models on power analysis procedures and Type I error rates for the test of close fit

The test of close fit in Covariance Structure Models (CSMs), in which the null hypothesis specifies that $RMSEA \leq .05$, relies on the assumption (known as the parameter drift assumption) that as sample size increases both sampling error and model error (i. e. the degree to which the model is an approximation) decrease. The power analysis procedures proposed by MacCallum et al. for both the test of close fit and the test of exact fit also rely on this assumption. The present study investigated the degree to which violations of this assumption effects the Type I error rate for the test of close fit. Model error was introduced using a procedure proposed by Cudeck and Browne. The empirical power for both the test of close fit and the test of exact fit is compared with the theoretical power computed using the MacCallum et al. Procedure. The results indicated that the test of close fit maintains the nominal Type I error rate under violations of the assumption. The empirical power and theoretical power for both the test of close fit and the test of exact fit are nearly identical under violations of the assumption.

SEM3

#109

SEM

Rolf Steyer
Steffi Pohl*University of Jena
University of Jena*A new Approach to Modeling Method Factors: The ICE-Model

Multitrait-multimethod data, their analysis and interpretation have puzzled many since Campbell and Fiske. Marsh distinguished 20 models dealing with this phenomenon. Among those are: (1) allowing for correlations between the corresponding measurement errors, (2) introducing m *methods factors* that are uncorrelated with the latent traits and uncorrelated among each other. Eid introduced the CTC (M-1) model, in which $m - 1$ *methods factors* represent the systematic differences between a reference method and the method considered. These method factors may correlate among each other, but they are not allowed to correlate with the latent trait variables. In this presentation, we will introduce a new approach of dealing with method factors, which is based on the theory of individual and average causal effects in the tradition of Neyman and Rubin. This approach is similar to the approach proposed by Eid, because it uses only $m - 1$ latent variables representing the method factors. However, it differs from Eid's approach, because: (a) the *ICE*-variables may correlate with the latent traits, (b) they may have non-zero expectations, namely the average causal effects of using the method (item, scale) instead of the reference method, and (c) the structure of the factor loadings differs from Eid's model.

#113

SEM

Libo Li
Peter M. Bentler*University of California, Los Angeles
University of California, Los Angeles*Robust Statistical Tests for Evaluating the Hypothesis of Close Fit of Misspecified Mean and Covariance Structural Models

Model fit is the key issue in the mean and covariance structure analysis. Although many ADF-like tests exist for evaluating model exact fit, few advances have been made in statistical tests for evaluating model close fit. Recently, Yuan, Hayashi and Bentler applied the theory of Vuong to mean and covariance structure analysis and derived an asymptotic distribution of the normal theory based likelihood ratio statistic under the alternative hypothesis. We utilize their results and propose several ADF-like RMSEA tests for evaluating the hypothesis of close fit of misspecified models. Simulation studies show that three of these tests have robust and desirable performance in spite of severe nonnormality across the examples when sample size is as large as 300. A new two-stage procedure which combines model exact fit tests and the proposed RMSEA tests for model close fit is further proposed for overall model fit evaluation in mean and covariance structure analysis.

SEM3

#116

SEM

Tron Foss
Karl G. Jøreskog
Ulf H. Olsson

Norwegian School of Management
Norwegian School of Management
Norwegian School of Management

Testing Structural Equation Models by ADF: The Effect of Kurtosis

Various chi-square statistics are used for testing structural equation models. If the observed variables are non-normal, Satorra & Bentler proposed a chi-square statistic, often called the SB statistic, which is maximum likelihood statistics multiplied by a scale factor which is estimated from the sample and involves an estimate of the asymptotic covariance matrix of the sample variances and covariances. Yet another chi-square statistics, often called the ADF test is used with the weighted least square method proposed by Browne. The two chi-squares SB and ADF are both on the form nc , where n is sample size minus one. If the model does not hold, c converges to a positive constant C when $N \rightarrow \infty$. Foss, Jøreskog & Olsson developed the relationship between SB and the kurtosis and showed that $C_{SB} \rightarrow 0$ when $\gamma_{2s} \rightarrow +\infty$ for all s ; $\gamma_{2s} = \mu_{4s} - 3$ is the univariate excess kurtosis. In this study we develop the relationship between ADF and the kurtosis and show, under some additional assumptions, that also $C_{ADF} \rightarrow 0$ when $\gamma_{2s} \rightarrow +\infty$ for all s . The practical consequence of this is that models that do not hold tend to be accepted by the chi-square test if kurtosis is large. Thus, these tests have low power for detecting misspecified models.

#117

SEM

Walter Herzog
Anne Boomsma

University of St. Gallen
University of Groningen

Finite Sample Corrections for RMSEA Estimation

The root mean square error of approximation, RMSEA, has become a standard statistic in evaluating structural equation models. Recent research identified two problems concerning the use of the regular estimate RMSEA. First, it significantly overestimates the population value when the sample size is small. Second, this overestimation vanishes with increasing misspecification of the investigated model. Therefore, when RMSEA is used to evaluate model fit, it is hard to distinguish models that fit approximately well from models that do not. In applied research, when the sample size is small, it is not clear whether a large value of RMSEA is an indication of a large RMSEA or evidence for a moderate population value plus a bias component. Partly based on earlier work, we propose to use Bartlett, Swain, or Yuan corrections of the maximum likelihood chi-square statistic for the computation of RMSEA. In a simulation study, it is shown that these corrections reduce the two aforementioned problems. The findings are further illustrated by correcting RMSEA in applied structural equation models.

SEM / FAC / AAP

#36

FAC, SEM

David J. Hessen
Conor V. Dolan

Utrecht University
University of Amsterdam

A Heteroscedastic One-Factor Model and Marginal Maximum Likelihood Estimation of the Parameters

In the present paper, a heteroscedastic one-factor model is considered. Heteroscedasticity is explicitly modeled by defining the residual variances of the observed scores as parametric functions of the unidimensional factor score. For the estimation of the parameters a marginal maximum likelihood procedure is proposed. Under the assumption of multivariate normality of the observed scores conditional on a single normally distributed factor score, the procedure yields consistent parameter estimators. Furthermore, a likelihood ratio test is derived, which can be used to test the usual homoscedastic one-factor model against the heteroscedastic model. A simulation study is carried out to investigate the robustness and the power of this likelihood ratio test. Results show that the asymptotic properties of the test statistic hold under both small test length conditions and small sample size conditions. Results also show under which conditions the power to detect different heteroscedasticity parameter values, is small, medium or large. Finally, for illustrative purposes, the marginal maximum likelihood estimation procedure and the likelihood ratio test are applied to real data.

SEM / FAC / AAP

#101

AAP, SEM

Kathleen Bentein
Christian Vandenberghe

*Université du Québec à Montréal (UQAM)
HEC Montréal*

Are the Effects of Increases and Decreases in Commitment on Turnover Intentions Symmetrical?
A Subgroups Latent Growth Modeling Approach

This study develops a dynamic approach of the relationships between organizational commitment dimensions and turnover intentions. To answer this issue with an adequate conceptualization of intra-individual change, we used a Second Order Factor Latent Growth Modeling approach to data collected among 330 employees over three repeated measurements of the dimensions of commitment and intent to quit, spaced at 3-month intervals. The use of a subgroups analysis procedure allowed us to relax the constraint of symmetrical effects on intended turnover, i. e. the 'effect' on intentions of a negative change in commitment is similar in magnitude, though opposite in sign, than a positive change in commitment. There is indeed no reason to believe that individuals who experience a drop in commitment would raise their level of quit intentions in the same proportion than they would reduce them in response to an increase in commitment. Rather, research from other areas suggests that negative events have more impact on individuals than positive events. By identifying subgroups of individuals with different change trajectories in commitment, results confirm that the constraint of symmetrical effects is not valid, showing that declines in commitment, whatever the dimension, are more impactful on intent to quit than increases in commitment.

#155

SEM, FAC

Jamshid Etezadi

Concordia University

Non-linear factor analysis: a means to investigate dimensionality and non-linear structure of multivariate data

Identification of the underlying dimensionality of multivariate data and revealing its structure are two main objectives in exploratory analysis of multivariate data. When the relations among the observed and latent variables are nonlinear, popular techniques such as principal components and factor analyses are not suitable to reveal dimensionality of data. This paper employs Etezadi and McDonald's nonlinear factor analysis method to reveal dimensionality of Schull, Jenkins and Carol data set and demonstrates that the above method can be generalized to model non-linear structures among latent variables.

#166

AAP, SEM

R. Mac Turner

University of the Sciences

Untangling the Complex Origins of Suicide Attempts using Structural Equation Modeling

The objectives of this presentation are to further develop a theoretical model of the complex interactions of cognitive, emotional, behavioral, and social facets of life that activate suicide attempts and to report on a variation of Structural Equation Modeling that combines latent variables with Hierarchical Logistic Regression to estimate risk and protective factors involved in suicide attempts. A sample of 486 persons diagnosed with Major Depression, Bipolar Disorder, Schizophrenia, and Schizoaffective Illness were assessed through structured interview, self-report instrument, and behavioral observations on a wide range of potential risk factor measures. Results of the analysis showed the complex interactive model to provide a good fit to the data. Suicide Intent, Hopelessness, Suicide Ideation, Depression, Anxiety, and a lack of Reasons for Living coalesced into a Misery Construct. High Misery increases the risk of suicide attempts by a factor of 3. 2. Severe symptoms and impaired cognition increased the risk by a factor of 1. 8. An Impulsive construct increased the risk of a suicide attempt by a factor of 2. 3. Unlike previous research, this analysis showed that previous suicide attempts increased risk by a factor of only 1. 5.

Poster Presentations

Poster Session 1

#29

AAP, BSI

Duncan K. H. Fong

*The Pennsylvania State University*Dynamic models for stated preferences in conjoint studies

The collection of repeated measures in psychological research is one of the most common data collection formats employed in survey and experimental research. The behavioral decision theory literature documents the existence of the dynamic evolution of preferences that occur over time and experience due to learning, fatigue, boredom, and so on. We discuss a number of recently developed Bayesian models for capturing such dynamic effects in conjoint applications. Real and simulated data sets are used to illustrate the methodology as well as demonstrate the ability of the models to recover a variety of different sources of dynamics that may surface with preference elicitation over repeated measurement.

#152

AAP, SEM

Robert C. Daniel

*Georgia Institute of Technology*Flying safe: An analysis of a pilot's role and decisions to produce a safe landing

Despite the various threats in the world today, the single most important factor to keep us safe in the sky is the pilot's flying ability. While computers can land a plane under normal conditions, a pilot's ability to quickly and correctly deal with an unexpected event is crucial to a safe flight. Federal regulations require various assessments of the complex task of flying a commercial aircraft, but a full understanding of the many factors that influence a pilot's ability is largely unknown. Using actual flight data, the current study uses structural equation modeling with autoregressive decision making to further our knowledge of a pilot's ability. Originally developed as a task analysis of less than ideal piloting decisions, this study incorporates many facets of a pilot's experience. Among these facets, an investigation is made of the effect of previous decisions and changing environmental factors influence various stages of a flight. The influence of the pilot's experience and personality traits is also assessed to determine which stages of flight are most susceptible or resilient against human error. The outcome of this model will provide for better selection criteria of new pilots and diagnostic abilities to increase training effectiveness before problems occur.

#48

AAP

Carmen Ximénez
Javier Revuelta*Universidad Autonoma de Madrid*
*Universidad Autonoma de Madrid*Optimizing sample size and power in sequential hypothesis testing for one-way ANOVA under group sampling: The CLAST rule

Several studies have demonstrated that the Fixed-sample Stopping Rule (FSR), in which the sample size is determined in advance, is less practical and efficient than sequential stopping rules. The Composite Limited Adaptive Sequential Test (CLAST) is one of such sequential stopping rules. Previous research in the context of the t test of mean differences and the χ^2 independence test for twofold contingency tables has found that CLAST is more efficient in terms of sample size and power than the FSR and other sequential rules. The present work extends previous research on the efficiency of CLAST to multiple group statistical tests. Simulation studies are conducted to test the efficiency of CLAST in one-way ANOVA for fixed effects models. The general test and two linear contrasts of multiple comparisons among treatment means are considered. We also introduce four rules for allocating N observations in J groups under the general null hypothesis, and three allocation rules for the linear contrasts. Results show that CLAST is more efficient than the FSR in terms of sample size and power for one-way ANOVA tests. However, the allocation rules vary in their optimality: selecting an allocation rule depends on the cost of sampling and the intended precision.

Poster Session 1

#110

APP, IRT

Bobby D. Naemi
Daniel J. Beal
Stephanie C. Payne

Rice University
Rice University
Texas A & M University

A Comparison of Proportional and Mixed Rasch Methods of Measuring Extreme Response Style

Extreme response style (ERS) refers to the tendency to disproportionately favor the endpoints or extreme categories of ordinal response or Likert-type scales. This study examined measurement differences in ERS by comparing past approaches of measuring extreme responses (a simple count or proportion of endpoint use) with recent approaches based on item response theory and latent class modeling. A mixed-model IRT approach is used to assign respondents to an extreme class based on category threshold differences, as opposed to simple endpoint counts. In this way, it is possible to distinguish between extreme responses that are a result of response bias, and extreme responses that are a legitimate reflection of high trait levels, a distinction unaccounted for in the ERS literature. Proportional methods of measuring ERS including the Greenleaf scale, ratings of an ambiguous stimulus, and overall endpoint count across a battery of personality questionnaires are thus compared with an approach identifying extreme responders based on the polytomous Mixed Rasch model. Results compare the extent to which the correlations between ERS and personality predictors differ based on the way in which ERS is measured. Measurement invariance between extreme and non-extreme responders is also assessed.

#140

CCC

Jamison D. Fargo
Bruce K. Schefft

Utah State University
University of Cincinnati

Of one or many types? A latent class analysis of psychogenic non-epileptic seizures

The majority of neuropsychological and psychiatric research considers or describes the diagnosis and study of psychogenic (non-epileptic) seizures (PS) as a mostly homogenous psychiatric condition. To empirically test this position, latent class analysis was applied to sociodemographic, neuropsychological, and psychological data collected from a sample of 260 consenting individuals diagnosed with PS from a Midwestern US epilepsy center. A series of one through six latent class measurement models were specified and estimated. Results indicated that a 4-class model best fit the data and made the most theoretical sense (R^2 and entropy = .93, classification error = .03). Classes were identified as: 1) high social and neuropsychological function / moderate psychological disturbance (31%); 2) low social function / poor neuropsychological function / severe psychological disturbance (24%); 3) moderate social and neuropsychological function / minimal psychological disturbance (23%); and 4) low social function / moderate neuropsychological function / severe psychological disturbance (22%). Although individuals with PS may share a common set of symptomological characteristics, results indicate that distinct subclasses of PS exist that differ in terms of sociodemographic characteristics, neuropsychological functioning, and psychological disturbance. PS subclass differences may be useful in understanding the etiology and course of PS.

#99

CTT

Younyoung Choi

University of Maryland at College Park

Unbalanced ANOVA

The purpose of the article is to review literature on analytic approaches to unequal-n (unbalanced) ANOVA and existing methods for analyzing experimental design models with unbalanced data and to relate them to existing computer program. The using the $R(\cdot)$ notation is applied to the sums of squares in the overparameterized linear model and the hypotheses are described in terms of the full-rank cell means model. This article summary the way of a solution by hand such as unweighted mean and weighted mean and fitting contract. Also, this article summarizes theory of each approach about unbalance ANOVA based on David. G. Herr and Jacquelyn Gaebelein's hypotheses and some authors' suggestions about the best type of sum of squares.

Poster Session 1

#121

FAC, IRT

Olesya Falenchuk

University of Toronto

Investigation of Linear Factor Analytic Data Generation Approach
for Comparative Studies of Polytomous IRT Models

Comparative studies of polytomous IRT models are usually based on simulated data. Often the data are simulated using one or more of the IRT models. A few studies, though, have used data generated using linear factor analytic approach with the assertion that this approach minimizes bias in favor of one of the models under comparison as it does not use any IRT model for data generation. However, it has never been investigated whether the factor analytic data generation method favors any specific group of polytomous models. The purpose of this study was to investigate the linear factor analytic data generation approach and examine whether it can be used for unbiased comparison of polytomous IRT models with the three types of item step response function (ISRF): cumulative probability models, adjacent category models, and continuation ratio models. Specifically, the mathematical model used for this data generation method was examined both empirically and analytically.

#13

IRT

Holmes Finch
Brian F. French*Ball State University
Purdue University*

Type I error rate for one type of DIF detection in the presence of the other type: A Monte Carlo study

Differential item functioning (DIF) is an issue of importance to psychometricians and policy makers, given the increasing importance of testing in educational accountability. Uniform DIF occurs when the conditional (on the measured ability) probability of a correct response to an item differs for two groups of interest (reference and focal), while Nonuniform DIF occurs when an item discriminates among students by ability differently for the two groups of interest. There exist several statistical tools for assessing each of type of DIF. An anomalous finding has shown that these methods can exhibit high rates of rejection for one type of DIF when only the other type is present. For example, when only Uniform DIF is present, rates of incorrect Nonuniform DIF detection can be very high, and vice versa. This simulation study focused on such rates of incorrect DIF detection, varying a number of conditions relevant to examinees, as well as qualities of the items such as values of difficulty and discrimination and level of DIF. Results suggest that the more extreme the level of one type of DIF, the greater likelihood of identifying the other type of DIF. Final results will also focus on the impact of the manipulated variables.

#92

IRT

Kristine Braaten
James S. Roberts*University of Maryland at College Park
Georgia Institute of Technology*

The Use of Marginalized Bayesian Item Parameter Estimation in the Generalized Graded Unfolding Model

If binary or graded *disagree-agree* responses result from an ideal point process--a respondent endorses an attitude statement to the extent that the sentiment expressed by the statement matches the respondent's opinion--the responses are best analyzed with an unfolding model. This study focuses on the generalized graded unfolding model (GGUM) and its use of marginalized Bayesian item parameter estimation. While a marginal maximum likelihood (MML) approach is typically used to estimate GGUM item parameters, this study adapts the methodology of Mislevy's marginalized *maximum a posteriori* procedure (MMAP) procedure. Analyses include a simulation assessing accuracy of GGUM estimates where number of items, number of response categories, and sample size are varied, plus a comparison of MMAP and MML estimates. The rationale for this research is to reduce the data demands associated with parameter estimation in the GGUM--marginal Bayesian estimates of GGUM item parameters may require far fewer subjects--and to improve estimation accuracy. When there are few respondents near the outer ends of the attitude continuum, over- and under-estimation of item parameters can occur in these regions. Setting informative prior distributions on item parameters and using a Bayesian estimation approach can limit errors in estimation.

Poster Session 1

#125

IRT

Chanho Park
Daniel M. Bolt

University of Wisconsin-Madison
University of Wisconsin-Madison

Multilevel IRT and Cross-National Skill Profile Differences on TIMSS 2003

The current study considers three-level IRT models for representation of cross-national skill profile differences on the TIMSS 2003 mathematics assessment. Two multilevel models are considered using different coding approaches. An item feature model characterizes item difficulty with respect to item features such as item content, cognitive process, and item format. An alternative skill attribute model uses the attributes identified by Tatsuoka, Corter, & Tatsuoka as predictors of item difficulty. In both models, random weights are attached to the difficulty predictors. Both models are fit using a Markov chain Monte Carlo (MCMC) procedure implemented in WinBUGS. Cross-national differences are accounted for through variability in the difficulty predictor weights across country, and are illustrated using a proficiency scaling framework. This research also provides a useful illustration of some of the successes and limitations of multilevel IRT modeling.

#47

IRT

Javier Revuelta
Carmen Ximénez

Universidad Autonoma de Madrid
Universidad Autonoma de Madrid

Psychometric models for structured responses with application to estimation of knowledge states

This presentation introduces a class of item response models for structured responses. Items are described by a vector of components and a design matrix. The vector of components indicates the subtasks that constitute the item response process. The design matrix is a binary matrix that relates the response alternatives to the components. Each row of the design matrix indicates whether or not each component is successfully applied for each response alternative. Then, the item response is a pattern of binary indicators. As a result, the model is a logistic function whose scale and location parameters are linear functions of the parameters of the components. These models are referred to as generalized log-linear item response models (GLLIRM). The presentation includes an empirical application with real data. The test measures mathematics literacy. Item responses can be incorrect, partially correct and incorrect. Moreover, correct responses are classified into several groups, depending on the response process, and the same occurs for partially correct and incorrect response. We applied the GLLIRM and show how the component vector and design matrix can be used to model these responses. The model provides an estimate of the latent trait and classifies examinees into latent classes or knowledge states.

#34

LDA

Leah H. Rubin
Katie Witkiewitz

University of Illinois at Chicago
University of Illinois at Chicago

Addicted to defaults: Maximum likelihood vs. restricted maximum likelihood estimation in the analysis of alcohol treatment outcomes

Recent advances in quantitative methodology and statistical computing power has provided applied researchers with the tools for analyzing large, multi-site, longitudinal datasets. Several examples of advanced longitudinal data analysis can be found in studies on alcohol initiation, consumption, and post-treatment outcomes. Despite the widely available computing options one of the remaining struggles for researchers in this area is dealing with missing data, which is very common in studies of alcohol use. Common approaches for handling correlated unbalanced data and the change in variability over time in mixed regression models are maximum likelihood (ML) and restricted maximum likelihood estimation (REML). Both methods are based on a maximum likelihood estimation approach and each are used as default estimators in commonly used statistical programs. There are tradeoffs with selecting ML over REML in situations with unbalanced data. The benefits associated with one method tend to be the costs associated with the alternative method. In this poster we draw comparisons between ML and REML using an example analysis of alcohol treatment outcomes.

Poster Session 1

#102

MDS

Naoko Kuga
Shin-ichi Mayekawa

Tokyo Institute of Technology
Tokyo Institute of Technology

Metric Unfolding via PREFMAP

Given the exact squared distances between the sets of row-objects and the column-objects, Shönemann's Classical Metric Multi-dimensional Unfolding (CMMDU) recovers the true configuration of the row/column objects. However, when the real dataset is analyzed, CMMDU often fails to produce the solution of the predetermined dimensionality. In this research, we present a new method of CMMDU which can recover the exact configuration from the linearly transformed squared distances such as the column standardized squared distance matrix. In our method, CMMDU is reformulated as the external analysis (PREFMAP) in which the row eigen vector matrix from CMMDU plays the role of the external configuration. It can also be used with the error perturbed dataset in a similar way as the original CMMDU is used. By using this method as the initial configuration, we expect a faster convergence for the general unfolding programs.

#171

MVA

Richard A. Harshman
Margaret E. Lundy

University of Western Ontario
University of Western Ontario

A randomization method of obtaining valid p-values for model changes selected "post hoc"

When model changes are guided by *post hoc* assessment of the observed improvements in prediction or fit (e. g. , in stepwise regression), it is usually impossible to obtain p-values for these improvements by conventional analytical methods. We describe a computer-intensive alternative that accurately estimates these p-values by using a modified randomization / permutation test procedure that empirically determines the appropriate null distributions. To demonstrate the method, we use it to get valid p-values for step improvements found during standard stepwise multiple regression. Our method corrects for bias caused by the increased 'capitalization on chance' intrinsic to post hoc variable selection; it does this by introducing an equivalent post hoc selection step into the process generating the null-hypothesis values. The method also corrects for an "inconsistency" bias by eliminating or "pruning out" permuted cases that are inconsistent with prior step results; without such pruning, the method would underestimate significance except on the first step. In a Monte Carlo sample of one million cases, the p-values estimated for fit improvements during a three-step stepwise multiple regression did not show a statistically detectable bias at any step. Potential applications include significance tests for more complex sequential methods, stepwise canonical correlation / MANOVA, and discriminant analysis.

#106

OTR

Ben Babcock
Rick Guyer

University of Minnesota
University of Minnesota

Randomized Methods for Partial Paired Comparison Scaling

This research examined how well six different randomized partial paired comparison (PPC) methods reproduced the full paired comparison scale values on the Minnesota Importance Questionnaire (MIQ). The MIQ is a vocational paired comparison battery consisting of 20 stimuli and a zero scale. Randomized PPC methods differ from past PPC methods, because previous PPC methods used some sort of algorithm to decide which subset of pairs were to be taken. The randomized methods randomly assign pairs as to eliminate potential subset-unique scale results. Scale values were found using logistic regression based on the Bradley-Terry-Luce model. Results showed that, while all methods were in high agreement concerning the ranks of the stimuli, the randomized multiple forms designs, which consisted of breaking the full scale into several forms by randomly assigning pairs to said forms, performed better overall than the conditions where pairs were randomly omitted. The estimated scale values of the 20 stimuli and the zero scale using the randomized 140 and 70 pairs form designs correlated significantly higher with the full paired comparison scale values than did the other methods.

Poster Session 1

#79

SEM

Victor L. Willson
Ronghua Sun

Texas A & M University
Texas A & M University

Application of Circular Statistics to Psychological Functioning

Circular statistics is a branch of statistical methodology that focuses on direction as a statistic of interest in two, three-, or multidimensional space. Applications to psychology or education related to psychological processes reported in published literature have previously focused on multidimensional scaling to characterize location in M-space. This paper summarizes current statistical theory of circular functions and proposes a two-step procedure employing structural equation modeling to estimate model parameters and fit for multigroup location with moderator processes. This is shown to be represented as a multilevel modeling problem in which the first level includes individual covariation of scores, and the second level focuses on the spatial structure of groups. Spatial statistics are estimated using nonlinear transformations utilized in spatial models common to biology and oceanography. Of particular interest is the potential for cross-level interactions between spatial characteristics of the second level and SEM parameters of the first level. The model is applied to a common problem in psychology, representation of individuals placed in groups specified into spatial quadrants based on multivariate interval scale scores. The Myers-Briggs Type Indicator Test (MBTI) is presented as an example of the type of data structure that might be encountered, and a hypothetical example is given.

Poster Session 2

#103

AAP, OTR

Maki Miyake

Tokyo Institute of Technology

Application of Graph Clustering Method for Lexical-semantic Network of the Synoptic Gospels

The purpose of this study is for a computational explanation of the mutual relationships among the Gospels, as one of a series of researches searching for application and systematization of linguistic resources. This research represents a graph clustering method to automatically construct a lexical-semantic network of the Synoptic Gospels in the New Testament. The lexical-semantic network is illustrated by a simple graph, whose vertices correspond to the words/concepts and whose edges to the lexical-semantic paths. As to generate the network of the Synoptic Gospels, we apply our original clustering algorithm called Recurrent Markov Cluster Algorithm (RMCL), whose key idea was originated from Markov Cluster Algorithm by van Dongen. In this study, we produce a concise lexical-semantic panorama of the Synoptic Gospels, which is consisted of lexical or semantic clusters that can be taken as concepts each. Furthermore, some typical subject categories of the Gospels are observed as concept clusters linked one another. Consequently, we confirm that RMCL is a powerful technique to classify and reconnect the words in texts and at the same time we find a possibility to employ it as a sort of ontology generator for the text analysis.

#24

AAP

Dong Gi Seo
David J. Weiss

University of Minnesota
University of Minnesota

Person-fit with the real data

Person-fit indexes evaluate the fit of an individual's test performance in item response theory (IRT) models. Detection of a non-fitting examinee is important because the non-fitting examinee's trait estimate, $\hat{\theta}$, might not provide an adequate description of his/her true trait level (θ). Various statistics have been proposed to detect non-fitting examinees. The person-fit statistic that has been the focus of most research and application is the l_z index. Previous monte carlo simulation research has investigated the normality of the l_z distribution with each true θ value and estimated θ value, and the accuracy of using different estimation methods was investigated when different types of non-fitting responses were present. However, a limitation of most studies of the performance of l_z is the use of short test lengths and known item parameter to estimate θ . In a real testing application, the true item parameter would be unknown. Therefore, empirical research using estimates of item parameters is necessary. This study provides a more practical basis for the application of l_z by using real test data from the Psychology 1001 course and item parameters estimated from that data.

Poster Session 2

#135

AAP

Darrin Grelle
Rod Dishman
Robert Vandenberg

The University of Georgia
The University of Georgia
The University of Georgia

The Measurement of Exercise Attitude Change after a Short-Term
Intervention Using Combinations of IRT and LGM

Accurately measuring change is essential to determining the success of an intervention. Latent growth modeling (LGM) assumes that the measure used to measure change is invariant across administrations. This study explored using IRT to better assess an individual's latent trait level on a five item exercise attitude scale after an intervention aimed at increasing physical activity. Participants were divided between control and intervention. All were measured before, in the middle of, and after the intervention. IRTLRDIF was used to detect DIF between control and intervention and across the three time periods using Time One intervention data as the referent. The program identified one item with significantly different difficulty and discrimination across all five comparisons. The IRTLRDIF estimates were used to compute Thetas for individuals at all three time periods. MPLUS was used to measure differences between a standard multiple indicator LGM assuming measurement invariance and a model with the computed thetas as the manifest indicators. The second model yielded better fit indicating that combining IRT and LGM might provide a more accurate measure of change than either approach alone. The theoretical implications of these findings will be evaluated using additional analyses aimed at validating the current results.

#126

CCC

Tatsuo OTSU
Takamitsu HASHIMOTO
Kojiro SHOJIMA
Tomoichi ISHIZUKA

The National Center for University Entrance Examinations
The National Center for University Entrance Examinations
The National Center for University Entrance Examinations
The National Center for University Entrance Examinations

Sex differences in Subject Selection Behavior of The National Center Test

We present here a practical combination of non-hierarchical clustering method and Jacques Bertin's matrix graphics. The National Center Test (NCT) is a nationwide examination conducted by the NCUEE for Japanese university admission. More than 500,000 applicants take the test every year. We analyzed the relationship between an applicant's test subject selection on NCT and his or her subsequent university application for the first period in 2004. For detailed analysis, we classified university departments into ten clusters by k-means method. The classification was based on the mean scores of successful and unsuccessful applicants in the four subjects of languages and mathematics. Although there was no significant difference in the pass ratio between male and female applicants in the clusters, we found a significant gender difference in the subject choice ratio for Physics and Biology. Female applicants preferred Biology than male applicants in every department cluster. The structure of the cross classified frequencies are lucidly recognized by Jacques Bertin's weighted matrix presentation.

#4

CTT

Kyung T. Han

University of Massachusetts at Amherst

No Such Thing Content Validity: Is Construct Validity a Safe Zone?

Since the emergence of three terms of the validity, i. e. , content validity, criterion-related validity, and construct validity, endless arguments between content validity and construct validity continued and it seems not to be stopped unless one type of validity reveals itself as superior than the other. The confusion of test validity due to misunderstanding and types of validity is explained in this chapter. Without losing utility of having previous concepts of validity, new terms and suggested use of a new two-way framework for both the process of test validation and the process of test construction are introduced. The prism-like two-way framework introduced in this chapter enables test construction and test validation in a framework at the same time and as a result the effectiveness of the spectrum of evidence of validity can be maximized.

Poster Session 2

#127

CTT

Daniel Serrano

*University of North Carolina at Chapel Hill*Estimation of Cronbach's α Under General Covariance Structure and Missing Data Via the Linear Mixed Model

In a recently published paper, Kistner and Muller provided an exact Gaussian distribution for α under general covariance structure and complete data. The robustness of their distribution to the presence of missing data is unknown. However, the approximation may work well asymptotically, even in the presence of missing data. In this study, expressions for α are derived under general covariance structure as a function of variance components. This permits estimation of α via ratios of components obtained from the linear mixed model. The linear mixed model provides a convenient means for handling missing data and complex error structures. This approach has advantages over standard methods for estimating α , which require case deletion in the presence of missing data. A numerical example is used to demonstrate the estimation of α under a variety of error covariance structures (e. g. parallel, τ equivalent, AR1, and blocked dependence) in the presence of missing data via the linear mixed model. In addition, the approximation of the distribution proposed by Kistner and Muller is evaluated for each covariance structure in the presence of missing data.

#46

FAC

Kentaro Hayashi
Peter M. Bentler
Ke-Hai Yuan*University of Hawaii at Manoa
University of California, Los Angeles
University of Notre Dame*On the Likelihood Ratio Test for the Number of Factors in Exploratory Factor Analysis

In the exploratory factor analysis, when the number of factors exceeds the true number of factors, the likelihood-ratio test (LRT) statistic no longer follows the chi-square distribution due to a problem of rank deficiency and of non-identifiability of model parameters. As a result, decisions regarding the number of factors may be incorrect. Typically, we obtain too many factors than the true number of factors. We support our argument on the overfactoring by a simulation study.

#122

IRT

Kentaro NAKAMURA
Hideki TOYODA*Waseda University
Waseda University*Investigating students' ratings for evaluation of university teaching

Students' ratings are frequently used for evaluation of university teaching. It is quite possible that the ratings may be affected by each student's motivation, diligence or involvement in the course. Although the hierarchical rater model (HRM) was proposed for rated responses to examine leniency and consistency of raters, it assumes a situation where some experts rate the responses of many examinees to open-ended items. In students' evaluation of teaching, number of raters are usually large. So we investigate the model which constrains parameters among raters, and discuss the problems analyzing real data.

#167

IRT

Ying Liu
Jay Verkuilen*University of Illinois at Urbana-Champaign
University of Illinois at Urbana-Champaign*A Multidimensional Unfolding Item Response Model for Forced Choice Data

The forced choice response format is popular to handle nuisance effects such as social desirability and halo effects in personality scales because subjects have to make tradeoffs between two concrete items, not choosing whether to endorse an item compared to a vague alternative of not endorsing. However, such scales have been historically difficult to score and most rules used are heuristic, not model-based. We propose a multidimensional unfolding IRT model for this response format. Estimation is done using maximum likelihood and MCMC.

Poster Session 2

#177

IRT, LDA

Weiwei Cui
James S. Roberts

*University of Maryland
Georgia Institute of Technology*

A Multidimensional Unfolding Item Response Model for Forced Choice Data

This paper illustrates the implementation and application of the generalized graded unfolding model for repeated measures (GGUM-RM) to analyze polytomous item responses to tests/questionnaires that are repeatedly administered to a single sample of respondents. The model is appropriate for measuring changes across repeated measurements of attitudes, preferences, and individual differences in certain developmental processes that occur in distinct stages. The GGUM-RM directly parameterizes latent change over time in an explicit fashion and also accounts for the correlations of latent trait scores across repeated measurements of the same individuals. The GGUM-RM is suitable for responses derived from repeated administration of either the same test/questionnaire form or alternate test/questionnaire forms with a subset of common items across forms.

The parameters of the GGUM-RM can be estimated using a fully Bayesian procedure via WinBUGS. The mean and the variance-covariance for the latent variables can be directly estimated by modeling the hyperparameters of the distribution of the latent variable. A recovery simulation is conducted to test the influence of sample size, number of items, and the proportion of common items across alternative forms. A real data analysis example is also given to illustrate how this model works in practice.

#49

LDA

You Dong Jun
Kim Cheongtag

*Seoul National University
Seoul National University*

On specifying the null model for incremental fit indices in latent growth curve model

LGM (Latent growth curve model) is the analytic method using the frame of structural equation modeling to estimate latent curve. One of the most prominent merits using SEM framework is that it provides fit indices that can be used for judge the fit of the researcher's model. We suspect that the fit indices that provide most SEM packages provide may not reflect the goodness of model fit. In the present study we propose two different types of null models: one not assuming the latent curve only and the other not assuming any structural relations. We claim that the widely-used the latter type of null model may not be applied to LGM. Instead the null model not assuming latent curve is a reasonable one for the LCM. In study 1 and 2, Monte Carlo experiments were conducted to ascertain whether the presented reasonable null model is valid. In study 3, real data is used to ascertain whether the presented reasonable null model is valid. These results provide some insights into what fit indices should be used in LCM.

#67

MVA

Hye Won Suk
Cheongtag Kim

*Seoul National University
Seoul National University*

Comparison of the multivariate analysis techniques for fMRI data analysis:
focusing on principal component analysis, independent component analysis and factor analysis

Sophisticated data analysis techniques are needed for fMRI data analysis due to the characteristics of fMRI data such as small signal intensity and noisiness. Univariate techniques have been widely used to localize the regions of activation till now, where researchers are interested in whether there is activation or not in a specific voxel. However, recently, multivariate techniques attract researchers' attention with some results showing that meaningful information could be extracted not from each voxel but from activation patterns across the whole voxels. In this context, this study was proposed to apply multivariate analysis techniques based on linear transformation to fMRI data and compare the patterns which each technique extracts. In study 1, multivariate techniques such as principal component analysis, rotated principal component analysis, independent component analysis and factor analysis were applied to simulated fMRI data. With the systematic variation in temporal and spatial correlation in fMRI data, rotated principal component analysis and factor analysis performed better in recovering original activation patterns. Study 2 illustrates the results when the multivariate techniques were applied to real fMRI data for vision areas. Like in study 1, only rotated principal component analysis and factor analysis extracted compatible patterns with the knowledge about retinotopy.

Poster Session 2

#100

OTR

Juan Pascual-Leone
 E. Manolo Romero Escobar
 Janice Johnson
 Sergio Morra

York University
York University
York University
Universita di Genova

Mathematical modeling of free recall in the Mental Attention Memory task

The Mental Attention Memory (MAM) task is a verbal span task that yields a good measure of mental-attentional (M-) capacity. M-capacity is a content-free resource for boosting activation of task-relevant schemes. MAM includes three subtasks that vary in degree of interference with the basic recall task and result in progressive increase in mental demand. We contrast two theory-driven models of developmental performance in the MAM. Both models are based on the Theory of Constructive Operators. Model-1 is based on structural meta-subjective task analysis; it assumes a developmentally increasing M-capacity, but involves no parameter estimation. Model-2 is a mathematical model of online processing, which assumes a developmentally increasing M plus decay of active memory units; one decay-rate parameter is estimated. The models are tested for goodness-of-fit to data collected from participants from 7 years old to adults. Goodness-of-fit is tested by mean and variance contrast, linear regression coefficients and root mean squared scaled deviation estimation and comparison, squared deviation of proportion recalled across the different developmental groups sampled, and analyses of serial position data. Based on results, a refined alternative Model-3 is proposed and tested. Implications for measurement of mental attentional capacity and quantitative modeling of span tasks are discussed.

#120

OTR

Rick Guyer
 Ben Babcock

University of Minnesota
University of Minnesota

Effect of Sample Size on the Stability of Paired Comparisons Scale Values

There has been very little research on sample size requirements for paired comparisons (PC) scaling. Consequently, this study examined the stability of PC scale values estimates as a function of sample size. Scale values for a 20-stimulus PC instrument were obtained with the Bradley-Terry model using logistic regression. Three sample sizes were examined: 25, 50, and 100 individuals. There were 21 groups in the conditions with 25 and 50 individuals per group, and 10 groups for the 100 individual condition. Scale values for each group were compared to those estimated from the reduced full sample. Reduced full samples were computed by removal of the cases common to the group from the full sample of 1,050. Based on the inter-group correlations, the stability of the scale value estimates was optimal with 100 individuals. The observed standard errors of the scale value estimates were found to negatively correlate with the average scale values. This correlation was moderated by group size, as the relationship weakened as group size increased. There was evidence that scale values became sensitive to sampling variation as the sample size decreased. Implications for the use of PC scaling and for future research in the field are discussed.

Poster Session 2

#80

SEM

Qi Chen
Victor Willson

Texas A & M University
Texas A & M University

Effect of Covariate Magnitude on Latent Class Analysis Determination

The current study was motivated by problems the researchers encountered in data analysis using latent class analysis (LCA), in which the inclusion of a covariate led to a completely different classification of participants compared to classification without the covariate in an SEM framework. Previous research in the area of LCA indicates that most researchers use real data to demonstrate the effects of covariates but no simulation study has been done to investigate systematically the effect of the magnitude of the covariance of an exogenous variable with latent class estimation. Specifically, how are conditional and unconditional LCA affected by the inclusion of an exogenous covariate in structural equation modeling of the relationship?

A Monte Carlo study of LCA with covariates of high and low (including zero) covariance with class membership was conducted. MPLUS was used to generate the data and then model the effects of covariates on LCA under various conditions. LCA groups of 2, 3, 4, and 5 are examined for sample sizes of 50, 100, and 200 per group. The quality of classification as a function of covariate magnitude was compared to classification without covariate and the influence of covariate magnitude on LCA determination is discussed.

#175

SEM

Vitor Manuel Pereira Bertoquini
José Luís Pais Ribeiro

Universidade do Porto
Universidade do Porto

Measurement Invariance of a Subjective Well-Being Model

Aims: The objective of this study is to analyse the measurement invariance of the Subjective Well-Being across two samples.

Methods: The samples consists of 1334 college students aged between 17 and 52 (M=22.83 DP=4.85) and of 227 college students aged between 17 and 42 (M=23 DP=3.77). Life Satisfaction was measured with the SWLS and PANAS was used to assess the frequency of Positive and Negative Affect. To assess Subjective Well-Being a composite index was created from these three components [NA - (PA + SWLS)]. Confirmatory factor analysis (AMOS 5.0) was used to analyse the SWB model. The fit of each model was compared to the previous, less constrained model to assess invariance. To establish configural invariance, a freely specified model is tested for each dataset. To establish metric invariance, a model with equal factor loadings is tested.

Results: The results of this study indicate that the configural (equivalent structure) and metric invariance (equivalent factor loadings) were achieved. Strict invariance was rejected.

Conclusions: The findings support the view of higher-order construct of SWB with three factors. The three factors of SWB seem to be validly measured and the partial factorial invariance across groups lends support for some measurement equivalence.

Symposiums

Functional Data Analyses of Music Performance

Organizer: Jim Ramsay

McGill University

Theodoro Koulis
Dan Levitin
Jim Ramsay

McGill University
McGill University
McGill University

Input-Output Systems in Psychoacoustics

Continuous response digital interfaces are becoming popular in psychoacoustics for measuring in real time psychological responses to episodes such as a musical performance. Many psychoacoustics experiments of this nature may be viewed as collections of input-output systems. The statistical problem is linking time-varying covariates to the continuous response variate. Using online data obtained from an experiment in psychoacoustics, we showcase new statistical tools that incorporate dynamical elements of the response. We outline the issues involved in analyzing input-output systems when the exact form of the underlying mathematical model is not known, and present a calibration method to facilitate inter-subject and intra-subject comparisons.

Janeen Loehr
Caroline Palmer

McGill University
McGill University

Analyzing finger taps with FDA techniques

Music performance, typing, and other fast motor sequences display important interactions among finger movements, arising from factors such as coupling (non-independence among fingers) and anticipatory motion (preparing for upcoming events). We describe applications of functional data analysis to musicians' finger movements in tapping sequences, that help clarify the roles of co-articulation and anticipatory motion in finger movement trajectories.

David Campbell
Sean Hutchins

McGill University
McGill University

Frequency trace functional ANOVA models for repetition priming

In two experiments, participants listened to a short sequence of musical tones and were instructed to sing back the pitch of the final tone as soon as possible. We manipulated whether the final tone was a repetition of a previous tone, the number of intervening tones between the pitch repetitions, and whether the final tone was globally expected or unexpected. We use wavelets to determine the instantaneous frequency of the participants' sung response, and track the path of the sung pitches. We measure how quickly the subject is able to reach the intended frequency through a functional ANOVA model. We then compare this functional ANOVA model with a simple ANOVA model, which only regards the onset of the sung pitch, and their abilities to determine the latency of the sung pitches.

Caroline Palmer
Simone Dalla Bella

McGill University
University of Finance and Management, Warsaw

Individual differences in pianists' finger dynamics

Studies of the recognition of personal identity rely mainly on static images such as photographs and fingerprints. In contrast, dynamic information in rapid movements, as in music performance, is highly individualized. To examine whether musical movement properties characterize personal identity, we studied finger movements of skilled pianists performing melodies on a musical keyboard, and used functional data analytic techniques to describe changes in velocity and acceleration trajectories that form dynamic musical signatures.

Data problems in tests and questionnaires

Organizers: Klaas Sijtsma & L. Andries van der Ark

Tilburg University

Joost R. van Ginkel

*Tilburg University*Missing data: Multiple imputation of item scores in test and questionnaire data under a two-way ANOVA model

Method two-way with normally distributed errors is a simple multiple-imputation method for missing data in test and questionnaire data. This method imputes scores that are based on a two-way ANOVA model of persons by items. Although this method is simple, it may produce biased results in a two-way ANOVA framework. An attempt was made to develop a version of two-way imputation that produces unbiased results. To this end, method two-way with normally distributed errors was combined with the data augmentation algorithm. Simulation results show that this method produces unbiased results in the intra-class correlation coefficient, the mean of squares in ANOVA, and the item means.

Klaas Sijtsma

Tilburg University

L. Andries van der Ark

Tilburg University

Wobbe P. Zijlstra

*Tilburg University*Outlier detection in test and questionnaire data

Outlier detection in item scores from tests and questionnaires for the measurement of psychological concepts has to deal with highly discrete data, for example, incorrect/correct scoring (0/1) and ordered rating scale scores (e.g., 0, ..., 4). This is different from 'classical' approaches to detecting outliers that deal with continuous variables. As a result of this difference, classical methods for identifying and accommodating outliers are not readily applicable to highly discrete data. This study proposes two new types of outlier indices for highly discrete item scores, one of which is closely related to person-fit analysis. The identified outliers were investigated for their influence with respect to the statistical analysis at hand. Ten real data sets were analyzed to this end, and recommendations were formulated for outlier identification and accommodation in test and questionnaire data.

Wilco H. M. Emons

*Tilburg University*On the use of person-fit analysis in test and questionnaire data

The purpose of person-fit analysis is to detect and diagnose item-response vectors that are unlikely under the test theory model that describes the data. For these vectors, the validity of the resulting test score may be questionable. In recent years, person-fit methods were successfully applied in the context of educational testing and psychological assessment. Person-fit methods, however, may also be useful in many other research areas, in which multiple-item scales are used to measure some latent constructs as well. Examples include scales that are used in clinical research, sociology, marketing, and quality of life research. Potential threats to the validity of individual response protocols in these kinds of applications include a lack of motivation to answer the questionnaire seriously (e.g., students completing the questionnaire for credits), a poor understanding of the questions due to language problems, and response tendencies (e.g., the tendency to avoid extreme response options). In this presentation, I will discuss the usefulness and limitations of person-fit approaches in test and questionnaire data from a broad perspective. Important features of different applications are discussed, such as the distinction between maximum performance (e.g., intelligence testing, often measured by a set of dichotomously scored items) and typical performance (e.g., attitude, preferences; often measured by a set of polytomously scored items). I also discuss how aggregated person-fit results may be useful for further understanding of the psychometric properties of a test or questionnaire.

Data problems in tests and questionnaires

Organizers: Klaas Sijtsma & L. Andries van der Ark

Tilburg University

L. Andries van der Ark

Tilburg University

Wilco H. M. Emons

Tilburg University

Klaas Sijtsma

Tilburg University

Identifying answer copier and source in real-life examinations

A typical problem in high-stakes examinations is that examinees may be tempted to copy answers from their neighbors. The most compelling evidence is obtained when a copier is caught red-handed. Alternative means of obtaining evidence have been suggested; in particular, the statistical analysis of patterns of item scores in an effort to determine the likelihood that this pattern or parts of it have been copied from a 'source', most likely a neighbor. These analyses result in answer-copying statistics that are usually compared to a reference distribution. Obtaining an accurate empirical reference distribution requires a large sample, whereas many examinations do not have that many examinees.

We propose two answer-copying statistics that allow for the presence of different test versions and for the location of t though the sample was relatively small (N = 231), whereas comparing the statistics with a reference distribution resulted in the detection of fewer copiers and sources.

Model Selection in Latent Variable Models

Organizer: Kristopher J. Preacher

University of North Carolina at Chapel Hill

Discussant: James H. Steiger

Vanderbilt University

Kristopher J. Preacher

University of North Carolina at Chapel Hill

Why We Should Adopt a Model Selection Strategy

The philosophy of science can offer guidance to scientists interested in selecting one model from a set of rival models by clarifying the goals of the model-fitting process. Specifically, the metatheory of science advocated by Imre Lakatos will be described and offered as a justification for replacing the prevailing practice of evaluating the fit of isolated models against arbitrary benchmarks with an alternative strategy, in which rival models are compared in terms of their relative abilities to generalize to new data. It is suggested that the currently prevalent practice of model evaluation does not adequately or efficiently achieve the aims of science, and that model evaluation is important only insofar as it reflects a model's ability to cross-validate better than its rivals. Areas to be emphasized in future research on model selection, with specific reference to latent variable models, will be outlined.

Roy Levy

University of Maryland

Gregory R. Hancock

University of Maryland

A Generalized Model Comparison Framework in Structural Equation Modeling

The current work will generalize an existing model comparison framework for structural equation modeling (SEM) so that it can be made applicable to a larger class of modeling situations. This will be accomplished in two stages. The first is the development of statistical tests in a bootstrapping paradigm that are not subject to the limitations of traditional tests based on asymptotic distribution theory of maximum likelihood estimation. More specifically, Levy and Hancock (2005) developed a model comparison framework that draws from results in determining model relation and statistical tests based on asymptotic distribution theory to address model comparison questions of distinguishability and difference in fit. In the current project new tests will be developed that are applicable to a larger class of models, resulting in a generalized model comparison framework. The second stage is an assessment of these new tests in terms of their use in the larger model comparison framework. The performance of these tests over the existing tests that employ asymptotic distributions will be evaluated by simulating data from known conditions and applying the framework to arrive at conclusions regarding model distinguishability and difference in fit.

Model Selection in Latent Variable ModelsOrganizer: Kristopher J. Preacher
Discussant: James H. SteigerUniversity of North Carolina at Chapel Hill
Vanderbilt University

Kristian Markon

*University of Minnesota*Minimum Description Length Goodness-of-Fit Tests

Information-theoretic criteria provide a compelling basis for latent variable model selection. So far, however, attention has focused on the use of information-theoretic criteria to assess the relative fit of two or more models, rather than the absolute fit of any single model. Here, a test of absolute fit is introduced for information-theoretic criteria in the Minimum Description Length (MDL) family, which includes the Bayesian Information Criterion (BIC) among other criteria. The test is based on comparison of the fit of a model of interest to that of a null nonparametric model. Simulation results suggest that the test performs well in assessing the fit of latent variable models. Use of the test is illustrated in an example from recently published comparisons of latent class and latent trait models of psychological disorder.

Michael C. Neale

*Virginia Commonwealth University*Comparative Fit Statistics in Raw Data Analysis

Direct analysis of raw data by full information maximum likelihood (FIML) instead of summary statistics (SS) such as means and covariances is becoming increasingly popular, due in part to its elegant handling of missing data. Under FIML, likelihood-ratio tests (LRTs) between models with different numbers of parameters operates equivalently to the analysis of SS. However, many comparative fit statistics such as the Bayesian Information Criterion involve the sample size (N) and the degrees of freedom of the model. Calculation of the sample size under FIML with missing data is not straightforward. Neither is the computation of the degrees of freedom. While software such as Mx uses the number of raw data observations minus the number of parameters, this is not equivalent to the degrees of freedom when fitting models to SS, where the number of observed covariances and means is used. Some divergence between the FIML and SS results is therefore to be expected. Akaike's Information Criterion is based solely on the degrees of freedom, and differs only by a constant from the SS case. Caution in the use of fit statistics other than AIC and the LRT seems warranted.

Modeling Response Times

Organizer: Eric-Jan Wagenmakers

University of Amsterdam

Pablo Gomez

*DePaul University*Variability in the Decision Boundaries and Starting Point in Accumulator Models

Accumulator models of RT have been very successful at accounting data from a variety of two-choice tasks. A crucial component of these models is the variability in their parameters. Most implementations of these type of models (e.g., Ratcliff's diffusion model) routinely incorporate variability in starting point of the accumulation process; this allows the model to account for conditions in which the mean error RTs are faster, than the mean correct RTs. Variability in the position of the decision boundaries, however, has traditionally not been implemented in this models. There is some overlap in the predictions of these two implementations (variability in starting point and variability in decision boundaries); however, there are also some differential predictions. This research focuses on the parameter spaces in which these differential prediction emerge, and in its empirical implications.

Han van der Maas

*University of Amsterdam*Phase Transition in Speed-Accuracy Trade-off

Sequential sampling models of choice reaction time predict a continuous trade-off between speed and accuracy. In this talk, a phase-transition model is proposed, in which sudden changes in reaction time and accuracy occur as function of small changes in payoffs for speed and accuracy. In an experiment, payoffs were slowly varied between values favoring guessing and values favoring accurate responding. As predicted by the PhTM, jumps between the states differed in position depending on the direction of change in the payoffs. New data-analytical techniques are used in order to analyze hysteresis. Also, the possibility of new sequential sampling models for a discontinuous trade-off will be discussed.

Modeling Response Times

Organizer: Eric-Jan Wagenmakers

University of Amsterdam

Adele Diederich

*International University Bremen*Modeling the Effects of Payoff on Response Bias in a Perceptual Discrimination Task: A Further Test on how Three Different Hypotheses Incorporate Response Biases into a Sequential Sampling Decision Process

Three hypothesis, *bound--change hypothesis*, *drift-rate-change hypothesis* and *two-stage-processing hypothesis* were proposed to account for data from a perceptual discrimination task in which three different response deadlines were involved and three different payoffs were presented prior to each individual trial. It was shown how the three different hypotheses incorporate response biases into a sequential sampling decision process; how payoffs and deadlines affect choice probabilities; and the hypotheses' predictions of choice times and choice probabilities. The two-stage-processing hypothesis gave the best account, especially for the choice probabilities whereas the drift-rate-change hypothesis had problems predicting choice probabilities as a function of deadlines (Diederich & Busemeyer, 2005). Here, a further test is provided including an addition hypothesis, *the mixture--of--processes hypothesis*.

Eric-Jan Wagenmakers

University of Amsterdam

Scott Brown

University of California at Irvine

Raoul P. P. P. Grasman

University of Amsterdam

Peter C. M. Molenaar

*Penn State*On the Relation Between the Mean and the Standard Deviation of a Response Time Distribution

Although it is generally accepted that the standard deviation of a response time (RT) distribution increases with the mean, surprisingly little work has investigated the precise nature of this relation. Here we show that one of the most popular models for RT and accuracy (i.e., Ratcliff's diffusion model) predicts an approximately linear relation between RT mean and RT standard deviation. We tested this prediction in several large data sets. Across a wide range of different tasks we consistently found an approximately linear relation between RT mean and RT standard deviation. This strong empirical regularity supports the use of the coefficient of variation to compare variabilities while controlling for differences in baseline speed of processing.

A. Voss

University of Freiburg

K.C. Klauer

*University of Freiburg*A Diffusion-Model Analysis of the IAT

The Implicit Association Test (IAT) is a well known RT-based measure of implicit attitude. In spite of its omnipresence in social psychological research, little is known about the cognitive mechanisms behind the IAT effect. A diffusion-model analysis may help to clarify this issue. In Study 1, the IAT effect is decomposed into three dissociable components: Relative to the compatible phase, (1) ease and speed of information accumulation are lowered in the incompatible phase, (2) more cautious speed-accuracy settings are adopted, and (3) non-decisional components of processing require more time. Studies 2 and 3 analyze interindividual differences in these components: While correlations between different IATs can be traced back in large part to differential speed-accuracy settings, variance related to the attitude to be measured was concentrated in the compatibility effect on information accumulation.

Scott Brown

University of California at Irvine

A.A.J. Marley

University of Victoria

Andrew Heathcote

*University of Newcastle, Australia*How does stimulus and response history affect decision-making?

Models of choice response time (RT) have become quite sophisticated, and accommodate many details of observed RT distributions and response probability patterns. However, the most modern and sophisticated models of choice RT are mostly silent on the topic of sequential effects, even though many empirical regularities have been observed in that area. We describe a modified choice RT model tailored to suit absolute identification data. This model includes two additions that allow it accommodate several of the most important sequential effects. The model successfully predicts that stimulus repetitions yield accuracy and RT bonuses, and that the similarity of the current stimulus to the previous stimulus affects both RT and accuracy in complex ways. We discuss some general points related to how *any* choice RT model can incorporate sequential effects, and why this might be important in the context of sources of variability (i.e., errors).

IRT Models for Continuous Responses and Response Times

Organizers: Cees A. W. Glas & Wim J. van der Linden
Moderator: Cees A. W. Glas

University of Twente
University of Twente

Rinke Klein Entink
Jean-Paul Fox
Wim J. van der Linden

University of Twente
University of Twente
University of Twente

Simultaneous Modeling of Response and Response Time Data with Covariates for the Person Parameters

With the introduction of computerized testing the collection of response times on test items has become straightforward. This additional information can be used to improve the accuracy of person and item parameter estimation. Further, it helps us to explore empirical relationships between response time/speed and response accuracy.

A hierarchical modeling framework was developed which provides a tool for the simultaneous analysis of responses and response times on a test. At the first level, as a response model, the two-parameter normal-ogive model with the usual parameter for the person's ability is used. In addition, a lognormal model with a parameter for the person's speed is used to model the response times. At the second level, a multivariate regression model for two person parameters is specified.

The joint modeling of ability and speed allows us to evaluate the role of alternative explanatory variables as well as different within-subject covariance structures. Since the item parameters are allowed to correlate, the model produces more accurate estimates of all parameters. A Bayesian MCMC method for simultaneous estimation of all model parameters was developed. Results from a simulation study and a real data example are presented.

Cees A. W. Glas
Oxana Korobko

University of Twente
University of Twente

Multivariate IRT Models for Mixed Responses

A multidimensional IRT model for mixed responses, an MML estimation procedure and a procedure for the evaluation of model fit using Lagrange multiplier tests are presented. The approach is exemplified using data from the central examinations in secondary education in the Netherlands. These data fitted an IRT model with a three-dimensional latent ability space. These three factors can be labeled as a "Language-dimension", a "Science-dimension" and an "Economy-dimension". Lagrange multiplier tests are used to test the validity of the pattern of factor scores with this interpretation.

Wim J. van der Linden

University of Twente

Using Response Times for Item Selection in Adaptive Testing

Response times on items can be used to improve item selection in adaptive testing provided a probabilistic model for their distribution is available. In this research, we used a hierarchical modeling framework with separate first-level models for the responses and response times and a second-level model for the distribution of the ability and speed parameters in the population of test takers. The framework allows us to retrofit an empirical prior distribution for the ability parameter each time a new response time is recorded.

In an example with an adaptive version of the Law School Admission Test, we show how this additional update of the posterior distribution of the ability leads to a substantial improvement of the ability estimator.

Wim J. van der Linden
Krista Breithaupt
David Chuah
Oliver Zhang

University of Twente
American Institute of Certified Public Accountants
American Institute of Certified Public Accountants
American Institute of Certified Public Accountants

Detecting Differential Speededness of Tests

A potential undesirable effect of giving different test forms to different examinees is differential speededness, which happens if the time limit is tight and some of the forms have items that are more time intensive than others. This paper shows how a lognormal response-time model can be used for estimating differences in speed and time intensities between test takers and subtests and detecting differential speededness. The results are then used in posterior predictive model fit analyses and simultaneous residual analyses of responses and response times to detect differential speededness.

An empirical data set for a multistage test in the computerized CPA Exam will be used to demonstrate the procedures. The results indicate tendencies to different time use across examinees on different subtests. However, these tendencies can be shown to be due to a warming up of the examinees at the beginning of the test rather than their running out of time near the end of it. We therefore conclude that the exam was not differentially speeded.

Cees A. W. Glas
Wim J. van der Linden

University of Twente
University of Twente

Testing Local Independence Assumptions for Responses and Response Times in an MML Framework

A marginal maximum likelihood (MML) framework for hierarchical models for speed and accuracy on test items is presented. The models are hierarchical arrangements of item response theory models for response times and dichotomously-scored items and multivariate normal models for their item and person parameters.

First, procedures for MML estimation of the model parameters are presented. Second, we discuss how to test the validity of the models using two different classes of statistical tests. The first class is likelihood-ratio tests that evaluate the hierarchically ordered versions of the models. The second class consists of Lagrange multiplier tests on three different assumptions of conditional independence in the model. A number of simulated examples will be used to show the feasibility of the estimation and testing procedure.

Both the computation of standard errors for the estimates and Lagrange multiplier tests require the inversion of a matrix of second-order derivatives. For application to the large item banks used for computerized adaptive testing, this inversion becomes unfeasible. Several approximation methods to obtain the needed matrices are evaluated. Finally, these approximations are used to evaluate the fit of the models in an application with data from computerized adaptive testing.

Intensive Longitudinal Data: Applications in Functional and Point Process Modeling	
Organizers: Theodore A. Walls & James O. Ramsay Discussant: Joseph L. Schafer	University of Rhode Island & McGill University Pennsylvania State University

Theodore A. Walls *University of Rhode Island*
Intensive Longitudinal Data: Current Progress and Future Prospects

This presentation will review the scientific goals of psychological studies that produce ILD and characterize the families of statistical models with high suitability for these data. In addition, applications in multilevel modeling and some extensions will be considered briefly.

Stephen L. Rathbun *University of Georgia*
Point Process Models for the Effects of Time-Varying Covariates on Rates of Repeated Behavioral Events

Examples of repeated events that occur at discrete points in time can be found in the behavioral studies, including the lighting of cigarettes by smokers, and acts of aggressive behavior in children. The rates at which such events occur may depend on time-varying covariates (e.g., mood), time of day, and characteristics of individual subjects (e.g., education level). This paper introduces a point process model that takes into account all of these potential sources of variation in event rates. Event rates are measured through an intensity function, expressed in units of the number of cigarettes smoked by hour, for example. A proportional intensity model is proposed, similar to the Cox proportional hazards model in survival analysis, under which the intensity is written as the product of the baseline intensity and a function of the covariates. The baseline intensity is written as a function of time of day, and is modeled non-parametrically. Both time-dependent and time-independent covariates may be included in the model. Subject-dependent random effects may also be included to take into consideration variation among subjects in baseline event rates. This model will be illustrated using data from an Ecological Momentary Assessment of smoking.

Carlotta Ching Ting Fok *McGill University*
 James O. Ramsay *McGill University*
Techniques for the analysis of Event Timings and Strengths

A functional data analysis model is developed for longitudinal and functional data from events that occur at random times. The model examines the event intensity, or the instantaneous rate of occurrence, and estimates the variable values of the data.

Event data are data that occur at random times, collected usually either using hand-held computer or manually by recording the event occurrence. For these events, each observation is the time t_i of an event, and the state x_i of a variable at the time of occurrence. If only event times are recorded, the data are called a point process. If both event times and variable values are recorded, the data are a marked point process. The variable values are known as the mark. The model, derived by the sum of the two log-likelihood functions, estimates the intensity function and the smoothed function for the mark variable simultaneously.

An example is the application of the model to a set of lupus data that involve the medical histories of 300 lupus sufferers over 20 years to examine the flare intensity and severity of lupus symptoms of each patient.

James O. Ramsay *McGill University*
Modeling and Controlling Input/Output Systems

Where a system produces a time-varying output $y(t)$ in response to a time-varying input $x(t)$, there are many modeling options. Among the most important are dynamic models, where, in the simplest case, the model fits the rate of change $Dy(t)$. A few examples are offered. If the system can be modified, then system designers will look for control loops that will reduce or eliminate undesirable aspects of how the system responds to changes in inputs. Control loops tend to use three types of information: (P) the immediate value of the input $x(t)$, (I) the past history of the input, and (D) the rate of change of the input. Each of these has its advantages and liabilities.